

Modelos Lineales Generalizados

M^{te} Asunción Martínez Mayoral
Javier Morales Socuéllamos

Dpto. Estadística y Matemática Aplicada
Centro de Investigación Operativa
Universidad Miguel Hernández

No esta permitida la reproducción total o parcial de este libro, ni de su tratamiento informático ni la transmisión de ninguna forma por cualquier medio, ya sea electrónico, por fotocopia, por registro u otros métodos, sin el permiso previo y por escrito de los titulares del copyright.

Impreso en España. Printed in Spain

Autor:

*M^aAsunción Martínez Mayoral
Javier Morales Socuëllamos*

ISBN: 84-95893-01-0

Depósito Legal: A-1544-2001

Fotocomposición, impresión y encuadernación:
Gráficas Limencop S.L. ELCHE

Editor: Universidad Miguel Hernández

Prólogo

Este manual de Modelos Lineales Generalizados surge del trabajo continuado de tres años de docencia impartiendo la asignatura Métodos de Regresión en la Licenciatura en Ciencias y Técnicas Estadísticas de la Universidad Miguel Hernández de Elche.

Hemos tratado de hacer una revisión a nivel básico, de los modelos lineales generalizados (GLM, abreviatura con la que nos referiremos frecuentemente a estos modelos, y que proviene del inglés *Generalized Linear Models*). La pretensión al confeccionar la asignatura fue la de conseguir que el estudiante, a la vista de unos datos, aprendiera a proponer cuestiones a resolver y a la vez modelos estadísticos útiles para contestarlas. Se le ofrecían como herramientas una serie de modelos lineales generalizados comunes en la práctica. Este manual sigue esa filosofía; asociado a cada concepto teórico tratado, se propone un ejemplo para ilustrar cómo se llevaría a cabo la modelización e interpretación de los resultados del análisis.

La resolución de los ejemplos está hecha con el paquete estadístico de libre disposición $\mathcal{R}$, desarrollado por *AT&T Bell Laboratories* por Rick Becker, John Chambers y Allan Wilks, iniciado como una réplica del lenguaje-software *S-Plus*. $\mathcal{R}$ está disponible en la dirección web <http://www.R-project.org/>. La flexibilidad y sencillez de este lenguaje para la programación lo hace una herramienta eficaz y versátil para el análisis estadístico. Además, $\mathcal{R}$ contiene implementados una buena cantidad de modelos lineales generalizados, con lo que la resolución de la mayoría de los modelos que proponemos es inmediata. Para aquellos poco familiarizados con este software, existe un manual del lenguaje en la página web citada.

B

Índice general

1. Introducción a los GLM	1
1.1. Limitaciones en el Modelo Lineal Normal	1
1.2. La Generalización del Modelo Lineal Normal	5
1.2.1. Elementos de un Modelo Lineal Generalizado (GLM)	6
1.3. Ejemplos de datos modelizables con GLM's	7
1.4. Modelos de cuasi-verosimilitud	10
2. Conceptos Básicos de los GLM	13
2.1. La familia Exponencial	13
2.2. Modelos Lineales Generalizados	16
2.3. Relación media–varianza en la familia exponencial	16
2.4. Ajuste del modelo	18
2.4.1. Necesidad de métodos numéricos para estimar por máxima verosimilitud	18

2.4.2.	El ajuste por MÁXIMA-VEROSIMILITUD	20
2.4.3.	Métodos Numéricos de Obtención de los EMV's	23
2.4.4.	Mínimos Cuadrados Ponderados Iterados	24
2.4.5.	Estimación del parámetro de escala ϕ	26
2.5.	Inferencia	27
2.5.1.	Distribución en el muestreo de U	28
2.5.2.	Distribución en el muestreo del $EMV(\beta)$	29
2.5.3.	Estadístico DEVIANCE: Bondad del ajuste	34
2.6.	Comparación de modelos	38
2.6.1.	Comparaciones con la deviance	39
2.6.2.	Comparaciones con el estadístico de Wald	40
2.6.3.	Comparaciones con los estadísticos C_p y AIC	45
2.7.	Chequeo del Modelo Ajustado: Residuos	50
2.7.1.	Tipos de residuos	52
2.7.2.	Residuos en $\mathcal{R}$	54
2.8.	Análisis de la Covarianza (ANCOVA). Ejemplos	58
2.9.	Ejercicios	63
3.	GLM para Respuesta Binaria	73
3.1.	Introducción. Ejemplos	73
3.2.	Hipótesis y Links para variables binarias	77
3.3.	Modelos de dosis-respuesta	79

3.3.1. Modelo Lineal	80
3.3.2. Modelo Probit	80
3.3.3. Modelo Logit	81
3.3.4. Modelo Complementario Log-Log	81
3.4. Estimación y Bondad de Ajuste	81
3.5. Residuos	84
3.6. Modelización y Análisis. Ejemplos	86
3.7. Ejercicios	105
4. Tablas de Contingencia y Modelos Log-Lineales	117
4.1. Introducción y Ejemplos	117
4.2. Tipos de muestreo y distribución de frecuencias	120
4.3. Modelos Log-Lineales	124
4.4. Estimación máximo-verosímil. Bondad de Ajuste	126
4.5. Consideraciones en el análisis de Tablas de Contingencia	128
4.6. Modelos Logit Multinomial	135
4.6.1. Variable a explicar Binaria	137
4.6.2. Variable a explicar Multinomial	141
4.6.3. Modelos Logit Generalizados	143
4.7. Modelos logit-multinomial y variables ordinales	147
4.7.1. Variable a explicar nominal y predictor ordinal	150
4.7.2. Variable a explicar ordinal y predictor nominal	150

4.7.3. Variable a explicar ordinal y predictor ordinal	152
4.8. Modelos de cambio	154
4.8.1. Modelos Mover-Stayer	155
4.8.2. Simetría	158
4.9. Ejercicios	162
5. Otros Modelos Lineales Generalizados	181
5.1. Introducción	181
5.2. Regresión de Poisson	181
5.3. GLM's para Datos Longitudinales	187
5.3.1. Introducción y Ejemplos	187
5.3.2. Curvas Exponenciales de Crecimiento	189
5.3.2.1. Respuesta Continua	190
5.3.2.2. Respuesta Discreta: conteos	192
5.3.3. Curva Logística de Crecimiento	193
5.3.4. Curva de Crecimiento de Gompertz	195
5.4. GLM's para Datos de Supervivencia Censurados	197
5.4.1. Conceptos Generales: Distribuciones asimétricas y Censura	197
5.4.2. Funciones de probabilidad para Tiempos de Supervivencia .	199
5.4.3. Modelización de la Censura	200
5.4.3.1. Modelos Hazard Proporcionales	201
5.4.3.2. Representación de Poisson	202

5.4.3.3. Distribución Exponencial	203
5.4.3.4. Distribución Weibull	204
5.4.3.5. Distribución del valor extremo	205
5.5. Ejercicios	206
6. Modelos de Cuasi-Verosimilitud	211
6.1. Introducción y Ejemplos	211
6.2. Cuasi-verosimilitud y estimación	213
6.3. Sobredispersión: Modelización y Análisis	218
6.4. Ejercicios	225
Índice de Tablas	227
Bibliografía	233

1

Introducción a los GLM

1.1. Limitaciones en el Modelo Lineal Normal

Muchos de los modelos de toda la vida son modelos lineales generalizados (GLM). Por mencionar algunos, tenemos:

- regresión lineal múltiple
- análisis de la varianza (ANOVA)
- análisis probit
- análisis logit para proporciones
- análisis de ítem
- modelos log-lineales
- modelos de regresión para datos de supervivencia
- polinomios inversos

- modelos basados en la familia exponencial
- etc

Sea cual sea el tipo de datos que pretendemos analizar, para una modelización adecuada siempre es básico preguntar y contestar una serie de cuestiones:

1. ¿Cuál es la variable respuesta, esto es, la variable a explicar?
2. ¿Son siempre homogéneas las condiciones de observación?
3. ¿Cómo cambia la respuesta *media* cuando se modifican las condiciones de experimentación?
4. ¿Qué variables pueden explicar algo sobre la respuesta?

Las respuestas a estas cuestiones constituyen la guía esencial de por dónde abordar el análisis de los datos para obtener el tipo de conclusiones deseadas.

Cuando además corroboramos que la respuesta verifica

- normalidad
- homogeneidad de varianzas
- linealidad (aditividad) de los efectos sistemáticos

proponemos como modelización el **Modelo lineal Normal**:

$$Y = X\beta + \epsilon \quad (1.1)$$

donde $\epsilon \sim N(0, \sigma^2 I)$, de forma que

$$E(Y) = \mu = X\beta$$

$$Var(Y_i) = \sigma^2$$

Sin embargo, el mundo de los datos no es perfectamente NORMAL. Muchas veces, los datos provienen de otras distribuciones que no pueden satisfacer nunca

	$E(Y)$	$Var(Y)$
$Y \sim N(\mu, \sigma^2)$	μ	σ^2
$Y \sim Br(\mu)$	$\mu \in [0, 1]$	$\mu(1 - \mu)$
$Y \sim Po(\mu)$	$\mu \in (0, +\infty)$	μ
$Y \sim Ga(a\mu, a)$	$\mu \in (0, +\infty)$	μ/a

Tabla 1.1: Distribuciones usuales. Relación entre la media y la varianza.

los requisitos del modelo normal respecto a la relación media-varianza. Algunas se ilustran en la Tabla 1.1, junto con su media y su varianza.

Cuando no se verifican las hipótesis del modelo lineal normal, la dinámica usual es intentar alguna **transformación** de las variables, de forma que se corrijan las deficiencias y así poder aplicar el modelo normal. Sin embargo, surgen algunos inconvenientes:

- Algunas transformaciones no están definidas en las fronteras del espacio muestral (como la logit).
- No es siempre directa la interpretación de la variable transformada.
- Generalmente no existe una escala única que corrija a la vez “todos los defectos” (violaciones de las hipótesis) que encontramos en el ajuste de un modelo lineal (como por ejemplo, linealidad y heterocedasticidad conjuntamente).

Ejemplo 1.1 (Regresión de Poisson) *Los datos son relativos al número de quejas a médicos de un servicio de urgencias en un hospital (Le, 1998, pag.113). El interés básico del análisis es encontrar relaciones entre el número de quejas que recibe un médico y el hecho de que sea hombre o mujer, haya sido residente o no en un servicio de urgencias, sus ingresos y el número de horas y consultas que ha realizado en allí. Algunos de los datos se presentan en la Tabla 1.2.*

La respuesta es “ Y_{ijk} = número de quejas que recibe el k -ésimo médico de sexo i ($= 1$ hombre, $= 2$ mujer) en la categoría j de residente en urgencias ($= 1$ no, $= 2$ sí)”. Dado que el número de quejas está limitado del número de consultas realizadas,

No.visitas	Quejas	Residente	Sexo	Ingresos	Horas
2014	2	Y	F	263.03	1287.25
3091	3	N	M	334.94	1588.0
879	1	Y	M	206.42	705.25
1780	1	N	M	226.32	1005.50
3646	11	N	M	288.91	1667.25

Tabla 1.2: Datos de un servicio de urgencias hospitalario.

podríamos asumir que Y_{ijk} se distribuye según una Poisson con tasa de ocurrencias λ_{ijk} y media proporcional al número de consultas realizadas, Z_{ijk} , esto es, $Y_{ijk} \sim \text{Po}(\lambda_{ijk}Z_{ijk})$; la tasa de ocurrencias estaría relacionada con el resto de variables explicativas.

En la Figura 1.1 tenemos representada la variable respuesta, cuya distribución es claramente asimétrica. Es imposible utilizar el logaritmo para corregir la asimetría dado que hay varios ceros. Ni siquiera la transformación $Y^{2/3}$ consigue algo de simetría. La modelización está resuelta en el Ejemplo 5.1.

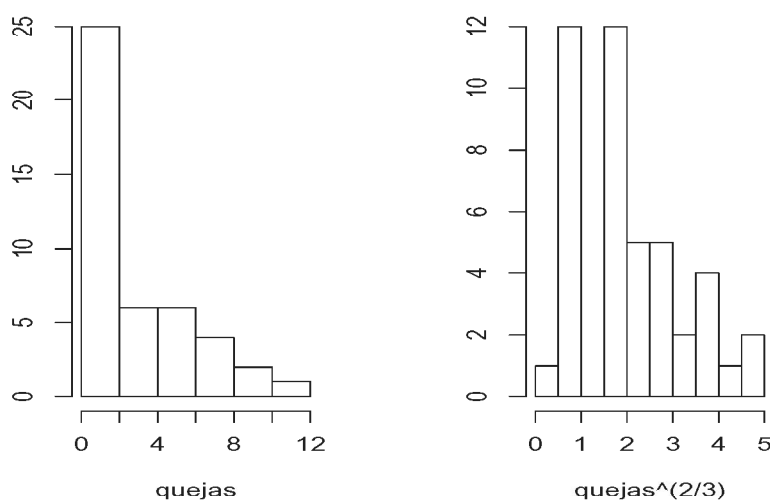


Figura 1.1: Número de quejas a médicos en un Servicio de Urgencias.

Por ejemplo, en problemas con datos basados en conteos procedentes de una distribución de Poisson, suele ocurrir que:

- los efectos sistemáticos a menudo son multiplicativos: $E(Y) = a \exp(bX)$.
- $\sqrt{Y}$ aproxima homogeneidad de varianza.
- $Y^{2/3}$ aproxima simetría o normalidad.
- $\log(Y)$ produce aditividad en los efectos sistemáticos ($E(\log(Y)) = \alpha + \beta Z$),

de modo que ninguna escala única corrige simultáneamente todas las propiedades deseadas para el ajuste de un modelo lineal.

1.2. La Generalización del Modelo Lineal Normal

Los **modelos lineales generalizados (GLM)** son una alternativa a transformaciones de la respuesta, justificadas por falta de linealidad y homogeneidad de la varianza.

En todos los modelos lineales generalizados, las propiedades básicas de los estimadores (como por ejemplo, las varianzas) son insensibles a la distribución asumida; dependen básicamente de:

- la relación asumida entre media-varianza,
- el grado de independencia/incorrelación entre las observaciones.

Las hipótesis básicas de un modelo lineal generalizado (GLM) son:

- **independencia** entre las respuestas;
- la “respuesta media” cambia con las condiciones, pero no la “forma funcional” de la distribución;
- la respuesta media, o alguna transformación de ella, cambia de modo lineal cuando las condiciones cambian.

Así, se trabaja con aquella distribución de probabilidad y aquel funcional para el cambio en la respuesta media de la variable observada, que mejor acomoda los datos. Esto es, los GLM permiten especificar, separadamente:

- la distribución de los datos, o lo que es básico, que es la relación media-varianza
 $\rightsquigarrow$ **función de varianza**;
- las relaciones de linealidad entre la respuesta media y los predictores
 $\rightsquigarrow$ **función link**.

Además,

✓ unifican todos los modelos en los que la respuesta (continua o discreta, en la familia exponencial) guarda algún tipo de relación lineal con los predictores;

✓ no fuerzan transformaciones de las variables a escalas no naturales, difíciles de interpretar;

✓ utilizan un mismo algoritmo de estimación: *mínimos cuadrados ponderados iterados*.

1.2.1. Elementos de un Modelo Lineal Generalizado (GLM)

- Las variables respuesta Y_i , $i = 1, \dots, n$, comparten la misma distribución en la familia exponencial.
- Un conjunto de variables explicativas X y de parámetros β .
- (Reparametrización) Una función monótona llamada *link*, $g(\cdot)$, proporciona el *predictor lineal* η ,

$$g(\mu_i) = \eta_i = X_i^T \beta,$$

con $E(y_i) = \mu_i$.

En la Tabla 1.3 aparecen algunos modelos y links usuales en la modelización de GLM's.

Modelos	Links
Normal	identidad μ
Binomial	inverso $1/\mu$
Poisson	inverso cuadrático $1/\mu^2$
Gamma	raíz cuadrada $\sqrt{\mu}$
Gausiano inverso	exponencial $(\mu + c_1)^{c_2}$
	log $\log(\mu)$
	logit $\log(\frac{\mu}{1-\mu})$
	cloglog $\log(-\log(\mu))$
	probit $\Phi^{-1}(\mu)$

Tabla 1.3: Algunos modelos y links usuales en ajustes de GLM's.

1.3. Ejemplos de datos modelizables com GLM's

Presentamos a continuación algunos ejemplos con datos en los que es viable la modelización a través de modelos lineales generalizados. La mayoría de estos ejemplos son tratados y analizados en temas sucesivos.

Ejemplo 1.2 (Respuesta Binaria) *En un experimento se sometió a cierto número de cucarachas a cinco horas de exposición a disulfato de carbono gaseoso a varias concentraciones (Bliss, 1935, también en Dobson, 1995, pag.109). Se pretendía investigar la relación existente entre la dosis de disulfato administrada y la resistencia de los insectos. Los datos aparecen en la Tabla 1.4.*

La variable respuesta es “ Y_i = número de escarabajos muertos en un total de n_i sometidos a una misma dosis de pesticida x_i ”. La distribución habitual es binomial, $Y_i \sim Bi(n_i, \pi_i)$ para $i = 1, \dots, n$, con π_i = probabilidad de muerte a dosis x_i . Ver Ejemplo

Ejemplo 1.3 (Respuesta multinomial) *Los datos de la Tabla 4.1 (Roberts et al, 1981, también en Dobson, 1995, pag.121) provienen de un estudio de pacientes con una forma de cáncer de piel llamado melanoma maligno. En una muestra de $N = 400$ pacientes se recogió información sobre la localización del tumor y su tipo histológico. Los datos son el número de pacientes en cada combinación de tipo de tumor y localización. El interés básico del análisis es investigar la relación entre el tipo de tumor y su localización.*

log.dosis (x_i)	No. insectos (n_i)	No. muertos (y_i)
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Tabla 1.4: Datos de mortalidad de cucarachas.

Dado el tipo de muestreo realizado (se seleccionó a un conjunto de pacientes con melanoma y se les fue clasificando en función de los dos criterios mencionados), la distribución de la variable respuesta “ Y_{ij} = número de tumores de tipo i en la localización j ” sería multinomial. Ver Ejemplo 4.1.

tipo tumor	Localización			
	cabeza/cuello	tronco	extremidades	total
Hutchinson	22	2	10	34
melanoma superficial	16	54	115	185
nodular	19	33	73	125
indeterminado	11	17	28	56
total	68	106	226	400

Tabla 1.5: Melanoma maligno: frecuencias por tipo de tumor y localización.

Ejemplo 1.4 (Respuesta ordinal) Los datos que se presentan en la Tabla 1.6 (Agresti, 1990, pag. 262) corresponden a un estudio en el que se pretendía concluir sobre la relación entre el grado de satisfacción en el trabajo y los ingresos percibidos, ambas variables categorizadas y de tipo ordinal. Si existe asociación, un segundo objetivo es predecir el grado de satisfacción en el trabajo (variable respuesta ordinal) en función de los ingresos percibidos (predictor ordinal). El interés del análisis es resolver estas cuestiones incorporando la información ordinal de las variables. Ver Ejemplo 4.12.

Ejemplo 1.5 (Respuesta normal) Los datos en la Tabla 1.7 (Dobson, 1990, pag.11) corresponden a un estudio en el que semillas genéticamente iguales son asignadas

Ingresos (\$)	Muy Insatisfecho	Poco Insatisfecho	Moderad Satisfecho	Muy Satisfecho
< 6000	20	24	80	82
6000 – 15,000	22	38	104	125
15,000 – 25,000	13	28	81	113
> 25000	7	18	54	92

Tabla 1.6: Satisfacción en el trabajo e ingresos.

(aleatoriamente), bien a un entorno enriquecido nutricionalmente (tratamiento), bien a condiciones estándar (control). Una vez han crecido todas las plantas, se recolectan, secan y pesan. El interés del análisis es investigar el efecto del tratamiento utilizado sobre el peso seco (en gramos) de las plantas en cuestión. Ver Ejemplo 2.17.

En este caso podemos asumir normalidad para la variable respuesta “ Y_{ij} = peso seco (gr) de la planta j en el grupo i (= 1 tratamiento, = 2 control)”, y tratar de captar diferencias entre el tratamiento y el control.

control	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
tratamiento	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69

Tabla 1.7: Datos sobre el crecimiento de plantas.

Ejemplo 1.6 (Respuesta Continua Asimétrica) Los datos en la Tabla 1.8 (Example U de Cox y Snell, 1981, también en Dobson, 1990, pag.47), corresponden a tiempos de supervivencia (en semanas) para pacientes de leucemia y su correspondiente conteo inicial de células blancas en la sangre (en escala $\log_{10}$). El interés del análisis es intentar predecir el tiempo de supervivencia Y en función del número inicial de células blancas x_i . Ver Ejercicio 3.

Una distribución usual en la modelización de tiempos de supervivencia es la exponencial, que es un caso particular de la gamma.

Ejemplo 1.7 (Respuesta Longitudinal) En economía se presta mucha atención a comparar las razones de crecimiento entre diferentes países. La formación del capital doméstico bruto es un índice importante para medir el crecimiento económico de un país. En la Tabla 1.9 (Oliver, 1970, también en Lindsey, 1997, pag.71) tenemos los

tiempos y_i	65	156	100	134	16	108	121	4	39
células x_i	3.36	2.88	3.63	3.41	3.78	4.02	4.00	4.23	3.73
tiempos y_i	143	56	26	22	1	1	5	65	
células x_i	3.85	3.97	4.51	4.54	5.00	5.00	4.72	5.00	

Tabla 1.8: Tiempos de supervivencia (en semanas) para pacientes de leucemia y conteo inicial de células blancas en la sangre (en escala $\log_{10}$).

datos de este índice en el Reino Unido para el período comprendido entre 1948 y 1967. El objetivo del análisis es modelizar la evolución de dicho índice de cara a una previsión del futuro.

La variable respuesta 'índice económico' tiene un crecimiento exponencial los primeros años, y acaba estabilizándose; podría asumirse una distribución exponencial. Ver Ejemplo 5.3.

año	1948	1949	1950	1951	1952	1952	1953	1954	1955	1956
gdpcf	1422	1577	1700	1889	2106	2359	2552	2829	3103	3381
año	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966
gdpcf	3492	3736	4120	4619	4731	4906	5860	6331	6686	7145

Tabla 1.9: Formación de capital doméstico bruto en Gran Bretaña durante 1948–1966.

1.4. Modelos de cuasi-verosimilitud

Los modelos de **cuasi-verosimilitud** son modelos lineales generalizados aún más generales, en los que en lugar de especificarse la distribución de la variable respuesta, se concreta únicamente:

- la función link: $g(\mu) = \eta = X\beta$;
- y la función de varianza $V(\mu)$ tal que $\text{Var}(y) = \phi V(\mu)$.

Son una buena alternativa a modelos en los que la respuesta presenta sobredispersión respecto de la distribución asumida, esto es, que presentan mayor variabilidad de la que postula dicha distribución.

Ejemplo 1.8 (Respuesta Poisson con sobredispersión) Los datos de la Tabla 1.10 (Bissell, 1972, también en Garthwaite et al, 1995, pag. 271) corresponden al número de defectos encontrados en rollos de fábrica de varias longitudes. El interés del análisis es investigar la relación entre el número de fallos (respuesta) y la longitud del rollo. Ver Ejemplo 6.5.

La respuesta ' Y_i = número de fallos' podría responder a una distribución de Poisson,; sin embargo hay indicios para sospechar que su varianza es mayor de la que le corresponde en caso de ser Poisson, quizás $\text{Var}(Y_i) = \phi E(Y_i)$, con $\phi > 1$. La modelización ese plantea a partir de la relación media varianza, sin considerar ninguna verosimilitud particular.

observación	No.Fallos	longitud del rollo
1	6	551
2	4	651
3	17	832
4	9	375
5	14	715
6	8	868
7	5	271

Tabla 1.10: Número de taras versus la longitud de los rollos.

2

Conceptos Básicos de los GLM

2.1. La familia Exponencial

La función de densidad de una variable aleatoria Y_i cuya distribución pertenece a la familia exponencial viene dada por:

$$f(y_i; \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}, \quad (2.1)$$

donde $a_i()$, $b()$ y $c()$ son funciones específicas de la distribución en cuestión, y θ_i es el denominado *parámetro natural o canónico*. Si ϕ es conocido, entonces tenemos un modelo de la *familia exponencial lineal*. Si ϕ es desconocido, el modelo pertenece a la clase de los *modelos de dispersión exponenciales*.

Muy a menudo, resulta que $a_i(\phi) = a_i \phi$, donde a_i es una constante de peso conocida y ϕ es llamado el *parámetro de dispersión o escala*.

Una formulación más general para expresar la función de densidad de una distribución de la familia exponencial es la siguiente,

$$f(y_i; \theta_i) = s(y_i)t(\theta_i) e^{r(y_i)w(\theta_i)} = \exp\{r(y_i)w(\theta_i) + c(\theta_i) + d(y_i)\}, \quad (2.2)$$

con $c(\theta_i) = \log(t(\theta_i))$ y $d(y_i) = \log(s(y_i))$.

Si $r(y_i) = y_i$, decimos que la distribución en (2.2) viene dada en *forma canónica*. A veces, a $w(\theta_i)$ se le llama *parámetro canónico* de la distribución. Si hay otros parámetros además del parámetro de interés θ , que forman parte de las funciones r, w, c y d , se consideraran como *parámetros de ruido*, y serán tratados como si fueran conocidos.

Ejemplo 2.1 (Distribución Normal)

$$Y_i \sim N(\mu_i, \sigma^2) \rightsquigarrow E(Y_i) = \mu_i$$

$$\begin{aligned} f(y_i; \mu_i, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}. \end{aligned}$$

$$\begin{aligned} \theta_i &= \mu_i \\ b(\theta_i) &= \theta_i^2 \\ a_i(\phi) &= \sigma^2 \\ c(y_i, \phi) &= -\frac{y_i^2}{2\phi} - \frac{\log(2\pi\phi)}{2} \end{aligned}$$

Ejemplo 2.2 (Distribución Poisson)

$$Y_i \sim Po(\mu_i) \rightsquigarrow E(Y_i) = \mu_i$$

$$\begin{aligned} f(y_i; \mu_i) &= \frac{\mu_i^{y_i}}{y_i!} \exp(-\mu_i) \\ &= \exp\{y_i \log \mu_i - \mu_i - \log y_i!\}. \end{aligned}$$

$$\begin{aligned}
\theta_i &= \log \mu_i \\
b(\theta_i) &= \exp(\theta_i) \\
a_i(\phi) &= 1 \\
c(y_i, \phi) &= -\log(y_i!).
\end{aligned}$$

Ejemplo 2.3 (Distribución Binomial)

$$Y_i \sim Bi(n_i, \pi_i) \rightsquigarrow E(Y_i) = n_i \pi_i = \mu_i$$

$$\begin{aligned}
f(y_i; \pi_i) &= \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \\
&= \exp \left\{ y_i \underbrace{(\log \pi_i - \log(1 - \pi_i))}_{\text{logit}(\pi_i)} + n_i \log(1 - \pi_i) + \log \binom{n_i}{y_i} \right\}.
\end{aligned}$$

$$\begin{aligned}
\theta_i &= \log \left(\frac{\mu_i}{n_i - \mu_i} \right) = \log \left(\frac{\pi_i}{1 - \pi_i} \right) \\
b(\theta_i) &= n_i \log(1 + e^{\theta_i}) \\
a_i(\phi) &= 1 \\
c(y_i, \phi) &= \log \binom{n_i}{y_i}.
\end{aligned}$$

Ejemplo 2.4 (Distribución Gamma)

$$Y \sim Ga(\nu, \nu/\mu_i) \rightsquigarrow E(Y_i) = \mu_i$$

$$\begin{aligned}
f(y_i; \mu_i, \nu) &= \left(\frac{\nu}{\mu_i} \right)^\nu \frac{y_i^{\nu-1}}{\Gamma(\nu)} \exp(-\nu y_i / \mu_i) \\
&= \exp \left\{ \frac{y_i(-1/\mu_i) - \log(\mu_i)}{1/\nu} + \nu \log(\nu) + \log(\Gamma(\nu)) + (\nu - 1) \log y_i \right\}
\end{aligned}$$

$$\begin{aligned}
\theta_i &= -1/\mu_i \\
b(\theta_i) &= -\log(-\theta_i) \\
a_i(\phi) &= 1/\nu = \phi \\
c(y_i, \phi) &= \frac{1}{\phi} \log \frac{1}{\phi} - \log(\Gamma\left(\frac{1}{\phi}\right) + \left(\frac{1}{\phi} - 1\right) \log(y_i)).
\end{aligned}$$

2.2. Modelos Lineales Generalizados

Un **Modelo Lineal Generalizado** (*Generalized Linear Model (GLM)*) consiste en:

1. Variables respuesta independientes $Y_1, \dots, Y_n$ que tienen la misma forma distribucional (paramétrica) dentro de la familia exponencial.
2. Un vector de parámetros $\beta' = (\beta_1, \dots, \beta_p)$ y una matriz de diseño X .
3. Una función monótona y diferenciable llamada **función link** $g(\cdot)$, que define la relación entre la media $\mu_i = E(Y_i)$ y el **predictor lineal** $\mathbf{x}_i' \beta = \eta_i$, es decir,

$$g(\mu_i) = \eta_i = \mathbf{x}_i' \beta.$$

Nota 2.2.1 En los GLM asumiremos que el parámetro ϕ se mantiene constante, sin interferir en la linealidad del predictor lineal respecto de β .

Nota 2.2.2 Llamamos **link canónico** a la función $g(\cdot)$ que define el parámetro canónico θ_i de un modelo en términos de su media μ_i ,

$$g(\mu_i) = \theta_i.$$

2.3. Relación media–varianza en la familia exponencial

En los miembros de la familia exponencial existe una relación especial entre la media y la varianza, que aprovecharemos posteriormente para obtener el ajuste y

estimar los parámetros de interés.

Denotemos por $l(\theta, \phi; y)$ a la función de log-verosimilitud obtenida con una única observación y ,

$$l(\theta; y) = \log f(y; \theta, \phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi),$$

y por U a la función *score* asociada,

$$U = \frac{\partial \log f(y; \theta, \phi)}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}. \quad (2.3)$$

El valor esperado de la función *score* es cero, y en consecuencia, la respuesta media es igual a la derivada de la función $b(\theta)$:

$$E(U) = 0 \Rightarrow E(Y) = b'(\theta). \quad (2.4)$$

Demostración:

$$\begin{aligned} E(U) &= \int \underbrace{\frac{\partial \log f(y; \theta, \phi)}{\partial \theta}}_{\frac{1}{f} \cdot \frac{\partial f}{\partial \theta}} f(y; \theta, \phi) dy = \int \frac{\partial}{\partial \theta} f(y; \theta, \phi) dy \\ &= \frac{\partial}{\partial \theta} \int f(y; \theta, \phi) dy = 0. \end{aligned}$$

Luego

$$0 = E(U) = \frac{E(y) - b'(\theta)}{a(\phi)} \Rightarrow E(y) = b'(\theta), \text{ cqd.}$$

La varianza de la función *score* es igual al valor esperado de su derivada cambiado de signo. En consecuencia, la varianza de la respuesta Y se puede escribir como el producto de la función de escala $a(\phi)$ y la derivada segunda de la función $b(\theta)$:

$$\text{Var}(U) = -E(U') \Rightarrow \text{Var}(Y) = a(\phi)V(\mu), \quad (2.5)$$

donde $V(\mu) = b''(\theta)$ es la llamada **función de varianza**.

Demostración:

Puesto que $E(U) = 0$ y $Var(U) = E(U^2) - E(U)^2 \rightarrow Var(U) = E(U^2)$. Con lo que basta probar que $E(U^2) = -E(U')$, siendo $U' = \partial U / \partial \theta$.

De (2.4) se tiene que $E(U) = 0 \Leftrightarrow \int \frac{\partial l}{\partial \theta} f dy = 0$. Tomando derivadas de nuevo,

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \int \frac{\partial l}{\partial \theta} f dy = \int \frac{\partial}{\partial \theta} \left(\frac{\partial l}{\partial \theta} f \right) dy \\ &= \int \frac{\partial^2 l}{\partial \theta^2} f + \frac{\partial l}{\partial \theta} \frac{\partial f}{\partial \theta} dy \\ &\quad \text{como } \partial l / \partial \theta = f' / f \rightsquigarrow f' = f \cdot \partial l / \partial \theta \\ &= E \left(\frac{\partial^2 l}{\partial \theta^2} \right) + \underbrace{\int \left(\frac{\partial l}{\partial \theta} \right)^2 f dy}_{E(U^2)}, \end{aligned}$$

de donde se concluye con que $E(U^2) = -E(U')$.

Y dado que $E(U') = -b''(\theta)/a(\phi)$, se tiene que

$$Var(U) = \frac{Var(Y)}{a(\phi)} = \frac{b''(\theta)}{a(\phi)},$$

y por lo tanto $Var(Y) = a(\phi)b''(\theta)$, cqd.

2.4. Ajuste del modelo

2.4.1. Necesidad de métodos numéricos para estimar por máxima verosimilitud

Nada mejor que un par de ejemplos para ilustrar la necesidad de métodos numéricos en la estimación de los parámetros de un modelo lineal generalizado. De hecho, en el modelo lineal normal, las ecuaciones normales para la estimación son analíticas, pero no es lo usual en cualquier GLM, como vemos a continuación en un modelo Poisson sencillo.

Ejemplo 2.5 (Ajuste del modelo Normal) Planteado el modelo lineal usual, $E(Y) = X\beta$, se obtiene la expresión explícita (analítica) de los estimadores máximo-verosímiles $\hat{\beta}$, simplemente derivando la log-verosimilitud $l(\theta; \mathbf{y})$ respecto de β , e igualando a cero. Dada la forma de la log-verosimilitud,

$$\begin{aligned} l(\theta; \mathbf{y}) &= \sum_{i=1}^n \left[\frac{y_i \theta_i - \theta_i^2 / 2}{2\sigma^2} - \frac{y_i^2 / \sigma^2 + \log(2\pi\sigma^2)}{2} \right] \\ &= \sum_{i=1}^n \left[\frac{y_i \mathbf{x}_i' \beta - (\mathbf{x}_i' \beta)^2 / 2}{2\sigma^2} - \frac{y_i^2 / \sigma^2 + \log(2\pi\sigma^2)}{2} \right] \end{aligned}$$

derivada respecto de β , e igualada a cero da lugar a:

$$\begin{aligned} \frac{\partial l(\theta; \mathbf{y})}{\partial \beta_j} = 0 &\rightsquigarrow \sum_{i=1}^n [y_i x_{ij} - x_i' \beta x_{ij}] = 0, \quad j = 1, \dots, p \\ &\rightsquigarrow X'(Y - X\beta) = 0 \\ &\rightsquigarrow \hat{\beta} = (X'X)^{-1} X'Y. \end{aligned}$$

Ejemplo 2.6 (Ajuste de un modelo Poisson) Planteado el modelo de predicción $\log[E(Y)] = X\beta$, y en particular $\log[E(Y_i)] = \beta_0 + \beta_1 x_i$, obtendríamos el EMV de β a partir de la log-verosimilitud

$$l(\mu; \mathbf{y}) = \sum_{i=1}^n y_i \log \mu_i - \mu_i,$$

con $\mu_i = E(Y_i)$, que expresada en términos de $\beta = (\beta_0, \beta_1)$ resulta

$$l(\beta; \mathbf{y}) = \sum_{i=1}^n y_i (\beta_0 + \beta_1 x_i) - \exp(\beta_0 + \beta_1 x_i).$$

Las ecuaciones que dan lugar a las estimaciones máximo-verosímiles son:

$$\begin{aligned} 0 = \frac{\partial l(\beta; \mathbf{y})}{\partial \beta_0} &= \sum_{i=1}^n y_i - \exp(\beta_0 + \beta_1 x_i) \rightsquigarrow \\ &\rightsquigarrow n\bar{y} = e^{\beta_0} \sum_{i=1}^n \exp(\beta_1 x_i), \end{aligned}$$

y

$$0 = \frac{\partial l(\boldsymbol{\beta}; \mathbf{y})}{\partial \beta_1} = \sum_{i=1}^n x_i y_i - x_i \exp(\beta_0 + \beta_1 x_i) \rightsquigarrow$$

$$\rightsquigarrow \sum_{i=1}^n x_i y_i = e^{\beta_0} \sum_{i=1}^n x_i \exp(\beta_1 x_i).$$

Ecuaciones que no son lineales en $\boldsymbol{\beta}$ y por lo tanto fuerzan la estimación numérica de su EMV.

2.4.2. El ajuste por MÁXIMA-VEROSIMILITUD

Notación

Consideramos un conjunto de datos que provienen de variables observadas $Y_1, \dots, Y_n$, cuya distribución es de la familia exponencial,

$$Y_i \sim F(\theta_i, \phi).$$

La log-verosimilitud conjunta viene dada por

$$l(\theta; \mathbf{y}) = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^n c(y_i; \phi), \quad (2.6)$$

con θ_i el parámetro canónico, relacionado con $\mu_i = E(Y_i)$, y ϕ un parámetro de escala. Consideramos el ajuste de un modelo lineal generalizado de la forma:

$$g(\mu_i) = \eta_i = \mathbf{x}_i' \boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathbb{R}^p.$$

Ecuaciones a resolver

Para obtener el EMV de $\boldsymbol{\beta}$ hemos de maximizar (2.6) respecto de $\boldsymbol{\beta}$. Dicha maximización es expresable a modo de una regresión por mínimos cuadrados ponderados, inmersa en un procedimiento iterativo.

El **ajuste por máxima-verosimilitud** se obtiene resolviendo las ecuaciones

$$U_j = \frac{\partial l(\theta; y)}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial l_i(\theta_i; y_i)}{\partial \beta_j} = 0, \quad j = 1, \dots, p, \quad (2.7)$$

es decir,

$$\sum_{i=1}^n \frac{(y_i - \mu_i) x_{ij}}{\text{Var}(Y_i) g'(\mu_i)} = 0, \quad \text{para } j = 1, \dots, p, \quad (2.8)$$

que escritas en forma matricial resultan:

$$U = X' M^{-1} (Y - \mu) = 0, \quad (2.9)$$

donde $U' = (U_1, \dots, U_p)$, $X = (x_{ij})_{n \times p}$, $Y' = (Y_1, \dots, Y_n)$, $\mu' = (\mu_1, \dots, \mu_n)$ y $M = \text{diag}\{m_1, \dots, m_n\}$, con $m_i = \text{Var}(Y_i) g'(\mu_i)$.

Notar que tanto μ como M son función de los parámetros a estimar, β .

Demostración:

1. Las observaciones son independientes, luego

$$\frac{\partial l(\theta; \mathbf{y})}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial l_i(\theta_i; y_i)}{\partial \beta_j}.$$

2. Dada la conexión directa $y_i \leftarrow \theta_i \leftarrow \mu_i \leftarrow \eta_i \leftarrow \beta$, la derivada del *score* respecto del parámetro de interés β se obtendrá aplicando la Regla de la Cadena:

$$\frac{\partial l_i(\theta_i; y_i)}{\partial \beta_j} = \frac{\partial l_i(\theta_i; y_i)}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j}, \quad (2.10)$$

donde

$$\begin{aligned} \frac{\partial l_i(\theta_i; y_i)}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)} \\ \mu_i = b'(\theta_i) &\rightarrow \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} \end{aligned}$$

$$g(\mu_i) = \eta_i \rightarrow g'(\mu_i) = \frac{\partial \eta_i}{\partial \mu_i} \rightarrow \frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}$$

$$\eta_i = \mathbf{x}_i' \beta \rightarrow \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Así, sustituyendo en (2.10) tenemos:

$$\sum_{i=1}^n \frac{\partial l_i(\theta_i; y_i)}{\partial \beta_j} = \sum_{i=1}^n \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)g'(\mu_i)} = 0. \quad \nabla$$

Nota 2.4.1 Si $g(\mu_i)$ es el link canónico, entonces $V(\mu_i) = 1/g'(\mu_i)$ y (2.8) queda simplificada a

$$\sum_{i=1}^n \frac{y_i x_{ij}}{a_i(\phi)} = \sum_{i=1}^n \frac{\mu_i x_{ij}}{a_i(\phi)}, \quad j = 1, \dots, p. \quad (2.11)$$

Demostración:

Por ser g el link canónico, $g(\mu_i) = \theta_i$, luego $\frac{\partial g(\mu_i)}{\partial \mu_i} = \frac{\partial \theta_i}{\partial \mu_i}$.

Como $\mu_i = E(Y_i) = b'(\theta_i)$, se tiene que

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) = V(\mu_i).$$

Así,

$$\frac{1}{\partial g'(\mu_i)} = \frac{\partial \mu_i}{\partial \theta_i} = V(\mu_i).$$

La obtención de (2.11) es directa.

Nota 2.4.2 Si $a_i(\phi) = a_i \phi$, con los a_i 's conocidos, entonces (2.8) queda

$$\sum_{i=1}^n \frac{(y_i - \mu_i)x_{ij}}{a_i V(\mu_i)g'(\mu_i)} = 0, \quad j = 1, \dots, p, \quad (2.12)$$

que no depende del parámetro de escala ϕ .

2.4.3. Métodos Numéricos de Obtención de los EMV's

Las ecuaciones que se obtienen para resolver el ajuste por máxima verosimilitud, (2.8), generalmente no son lineales en β_j , $j = 1, \dots, p$. Surge entonces la necesidad de utilizar métodos numéricos para obtener las estimaciones (máximo-verosímiles) $\hat{\beta}_j$, $j = 1, \dots, p$, y en consecuencia obtener el ajuste $\hat{\mu}_i = g^{-1}(\mathbf{x}'_i \hat{\beta})$. Como posibilidades tenemos:

1. Newton-Raphson:

$$\hat{\beta}^{(r)} = \hat{\beta}^{(r-1)} - H_{\hat{\beta}^{(r-1)}}^{-1} U_{\hat{\beta}^{(r-1)}}. \quad (2.13)$$

2. Scoring de Fisher:

$$\hat{\beta}^{(r)} = \hat{\beta}^{(r-1)} + \mathcal{I}_{\hat{\beta}^{(r-1)}}^{-1} U_{\hat{\beta}^{(r-1)}}, \quad (2.14)$$

donde

$$\begin{aligned} U &= \left(\frac{\partial l}{\partial \beta_j} \right)_j \\ H &= \left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right)_{jk} \\ \mathcal{I} &= E(H) = \sum_{i=1}^n E \left(\begin{pmatrix} \frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} \end{pmatrix}_{jk} \right), \end{aligned}$$

y U_{β} , H_{β}^{-1} e $\mathcal{I}_{\beta}^{-1}$ son las matrices U , H^{-1} e $\mathcal{I}^{-1}$ evaluadas en β .

Demostración:

1. Newton-Raphson se apoya en el desarrollo en serie de Taylor:
Si β^* es solución de $U(\beta) = U_{\beta} = 0$, y $\beta^{(0)}$ es un punto cualquiera, entonces podemos escribir aproximadamente

$$0 = U_{\beta^*} = U_{\beta^{(0)}} + H_{\beta^{(0)}} (\beta^* - \beta^{(0)}),$$

donde $H = \partial U / \partial \beta = (\partial^2 l / \partial \beta_j \partial \beta_k)$ es la matriz hessiana.

Podemos así construir un procedimiento iterativo que finalmente llegará al cero de U a través de,

$$\hat{\beta}^{(r)} = \hat{\beta}^{(r-1)} - H_{\hat{\beta}^{(r-1)}}^{-1} U_{\hat{\beta}^{(r-1)}}.$$

2. El método Scoring de Fisher aprovecha el mismo algoritmo iterativo de Newton-Raphson sustituyendo la matriz hessiana por la matriz de información de Fisher, $\mathcal{I} = -E(H)$, dando lugar a (2.14).

Cualquiera de los procedimientos iterativos de búsqueda de los EMV's es expresable en términos de un ajuste iterativo de mínimos cuadrados ponderados, lo que da lugar a la denominación del método de estimación como **Ajuste por Mínimos Cuadrados Ponderados Iterados**. Repasemos en primer lugar la justificación y procedimiento de estimación por mínimos cuadrados ponderados, para pasar a continuación a expresar el algoritmo iterativo de Fisher en términos de un ajuste similar.

2.4.4. Mínimos Cuadrados Ponderados Iterados

Propuesto el ajuste del modelo lineal $Y = \mathbf{X}\beta + \epsilon$, donde $\epsilon \sim N_n(0, \sigma^2 V)$, V es una matriz conocida, definida positiva y simétrica tal que existe una matriz K con $V = KK'$, para conseguir los estimadores máximo-verosímiles de β dentro de los supuestos del modelo lineal normal, basta con transformar

$$\left. \begin{array}{l} Z = K^{-1}Y \\ M = K^{-1}X \\ \delta = K^{-1}\epsilon \end{array} \right\} \text{ y ajustar } Z = M\beta + \delta, \text{ donde } \delta \sim N_n(0, \sigma^2 I)$$

La estimación por **mínimos cuadrados ponderados** consiste en resolver

$$\min_{\beta} (Z - M\beta)'(Z - M\beta),$$

que da lugar al EMV

$$\begin{aligned} \beta^* &= (M'M)^{-1}M'Z = (X' \underbrace{(K^{-1})'K^{-1}}_{(KK')^{-1}=V^{-1}} X)^{-1} X' \underbrace{(K^{-1})'K^{-1}}_{(KK')^{-1}=V^{-1}} Y \\ &= (X'V^{-1}X)^{-1}X'V^{-1}Y. \end{aligned}$$

La expresión del algoritmo iterativo Scoring de Fisher como un ajuste de mínimos cuadrados ponderados iterados viene dada por:

$$\hat{\beta}^{(r)} = (X'W_{\hat{\beta}^{(r-1)}}^{-1}X)^{-1} X'W_{\hat{\beta}^{(r-1)}}^{-1} Z_{\hat{\beta}^{(r-1)}}, \quad (2.15)$$

donde

$$\begin{aligned} \mathcal{I} &= -E \left[\left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right)_{jk} \right] = - \sum_{i=1}^n E \left(\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} \right)_{jk} \\ W &= \text{diag}\{ \text{Var}(Y_i) g'(\mu_i)^2; \ i = 1, \dots, n \} \\ Z &= X\beta + (\mathbf{y} - \mu) \text{diag}\{ g'(\mu_1), \dots, g'(\mu_n) \}. \end{aligned}$$

Demostración:

Para llegar a la expresión (2.15) procedemos, en primer lugar, multiplicando ambos lados de (2.14) por la matriz $\mathcal{I}_{\hat{\beta}^{(r-1)}}$, de donde

$$\mathcal{I}_{\hat{\beta}^{(r-1)}} \hat{\beta}^{(r)} = \mathcal{I}_{\hat{\beta}^{(r-1)}} \hat{\beta}^{(r-1)} + U_{\hat{\beta}^{(r-1)}}. \quad (2.16)$$

Si calculamos la expresión de los términos de $\mathcal{I}_{\beta} \equiv \mathcal{I}(\beta)$,

$$\begin{aligned} \mathcal{I}(\beta)_{jk} &= -E \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right] = - \sum_{i=1}^n E \left[\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} \right] \\ &= \sum_{i=1}^n E \left[\left(\frac{\partial l_i}{\partial \beta_j} \right) \left(\frac{\partial l_i}{\partial \beta_k} \right) \right] = \sum_{i=1}^n E \left[\frac{(y_i - \mu_i)^2 x_{ij} x_{ik}}{\text{Var}(Y_i)^2 g'(\mu_i)^2} \right] \\ &= \sum_{i=1}^n \frac{x_{ij} x_{ik}}{\text{Var}(Y_i) g'(\mu_i)^2}. \end{aligned}$$

Es decir,

$$\mathcal{I}_{\beta} = X'W_{\beta}^{-1}X, \quad (2.17)$$

donde $W = \text{diag}\{ \text{Var}(Y_i) g'(\mu_i)^2, \ i = 1, \dots, n \}$, y puesto que

$$U_{\beta} = X'W_{\beta}^{-1}Z_{\beta}^*,$$

con $Z_i^* = (y_i - \mu_i) g'(\mu_i)$, $i = 1, \dots, n$, la ecuación (2.16) queda:

$$X'W_{\hat{\beta}^{(r-1)}}^{-1} X \hat{\beta}^{(r)} = X'W_{\hat{\beta}^{(r-1)}}^{-1} X \hat{\beta}^{(r-1)} + X'W_{\hat{\beta}^{(r-1)}}^{-1} Z_{\hat{\beta}^{(r-1)}}^*,$$

de donde se obtiene la expresión

$$\begin{aligned} X'W_{\hat{\beta}^{(r-1)}}^{-1}X\hat{\beta}^{(r)} &= X'W_{\hat{\beta}^{(r-1)}}^{-1}Z_{\hat{\beta}^{(r-1)}}, \\ \rightsquigarrow \hat{\beta}^{(r)} &= (X'W_{\hat{\beta}^{(r-1)}}^{-1}X)^{-1}X'W_{\hat{\beta}^{(r-1)}}^{-1}Z_{\hat{\beta}^{(r-1)}}, \end{aligned}$$

similar a la ecuación de ajuste de regresión por *mínimos cuadrados ponderados* (MCP), con W como matriz de pesos. La diferencia con el procedimiento habitual de MCP estriba en la necesidad de *iterar*; de ahí la coletilla de *iterados*.

Notar que tanto los pesos W como las respuestas Z dependen del vector de parámetros a estimar β . En concreto, en el algoritmo dependerán del valor obtenido para $\hat{\beta}$ en la última iteración.

Nota 2.4.3 Especificando una estimación inicial $\hat{\beta}^{(0)}$ de β , las ecuaciones en (2.15) se pueden resolver iterativamente hasta obtener una secuencia de estimaciones $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}, \dots$ que convergen al estimador máximo-verosímil $\hat{\beta}$.

$$\hat{\beta}^{(r)} = (X'W_{\hat{\beta}^{(r-1)}}^{-1}X)^{-1}X'W_{\hat{\beta}^{(r-1)}}^{-1}Z_{\hat{\beta}^{(r-1)}}.$$

Nota 2.4.4 La estimación inicial de β se puede obtener tomando $\hat{\mu}_i = y_i$, y ajustando si fuera necesario.

Nota 2.4.5 Si $a_i(\phi) = a_i\phi$, entonces la matriz de información de Fisher (2.17) se puede expresar como

$$\mathcal{I} = \frac{1}{\phi} X'W^{-1}X, \quad (2.18)$$

donde en este caso, $W = \text{diag}\{a_i V(\mu_i) g'(\mu_i)^2, i = 1, \dots, n\}$.

2.4.5. Estimación del parámetro de escala ϕ

Desafortunadamente, aunque ϕ sea conocido, como ocurre en los modelos Poisson y binomial, necesitaremos estimarlo.

Cuando $a_i(\phi) = a_i\phi$, un *estimador consistente* de $\phi (= \text{Var}(Y_i)/a_iV(\mu_i))$ que permite también la estimación (consistente) de β es el **estimador corregido de los grados de libertad**, definido como:

$$\tilde{\phi} = \frac{1}{n-p} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)}. \quad (2.19)$$

La relación entre este estimador corregido de los grados de libertad y el *estadístico de Pearson generalizado* es directa:

$$\chi^2(y; \hat{\mu}) = (n-p)\tilde{\phi}. \quad (2.20)$$

2.5. Inferencia

Una vez seleccionado y ajustado un modelo concreto, se plantea el problema de inferir sobre la calidad del ajuste: dar intervalos de confianza de los parámetros, resolver contrastes de hipótesis y calcular la bondad del ajuste. Para ello es imprescindible conocer la distribución en el muestreo de los estimadores y de otros estadísticos que serán útiles para medir la bondad del ajuste.

La idea básica es que consiguiendo estimadores consistentes, estos serán insesgados y asintóticamente normales por el Teorema Central del Límite.

En el Teorema Central del Límite se basa la obtención de las distribuciones en el muestreo de los estadísticos:

- Scores U
- $\text{EMV}(\beta_j) = \hat{\beta}_j$
- Deviance (estadístico de bondad de ajuste),

que se presentan a continuación.

Teorema 1 (Teorema Central del Límite) Si $\hat{\theta}$ es un estimador consistente de un parámetro θ (es decir, $\hat{\theta}_n \xrightarrow{P} \theta \Leftrightarrow \lim_{n \rightarrow \infty} \Pr(|\hat{\theta}_n - \theta| > \epsilon) = 0$), y $\text{Var}(\hat{\theta})$ es su varianza, entonces para muestras grandes tenemos, aproximadamente al menos, que

1. $\hat{\theta}$ es un estimador insesgado de θ
2. $\hat{\theta} \sim N(\theta, \text{Var}(\hat{\theta}))$, o lo que es igual, $(\hat{\theta} - \theta)^2 / \text{Var}(\hat{\theta}) \sim \chi_1^2$.

Generalizando el teorema anterior a p dimensiones,

Teorema 2 (Teorema Central del Límite p -dimensional) Sea $\theta' = (\theta_1, \dots, \theta_p)$. Si $\hat{\theta}$ es un estimador consistente de θ y V su matriz de varianzas-covarianzas, entonces

1. $\hat{\theta}$ es, asintóticamente, un estimador insesgado de θ .
2. Si V es no singular, entonces

$$(\hat{\theta} - \theta)' V^{-1} (\hat{\theta} - \theta) \sim \chi_p^2.$$

Si V es singular, entonces tenemos varias alternativas:

- a) Obtener una inversa generalizada de V , V^- , y entonces

$$(\hat{\theta} - \theta)' V^- (\hat{\theta} - \theta) \sim \chi_p^2.$$

- b) Expresar el modelo en términos de un vector paramétrico $\varphi \in \mathbb{R}^p$ tal que la matriz de varianzas-covarianzas W de su estimador $\hat{\varphi}$ sea no singular, y así

$$(\hat{\varphi} - \varphi)' W^{-1} (\hat{\varphi} - \varphi) \sim \chi_p^2.$$

2.5.1. Distribución en el muestreo de U

Siendo $U_j = \frac{\partial l}{\partial \beta_j}$, $j = 1, \dots, p$, la derivada de la log-verosimilitud, y utilizando que

$$E(U_j) = 0, \quad \text{Var}(U) = E(UU^t) = -E(U') = \mathcal{I},$$

donde $\mathcal{I}$ es la matriz de información de Fisher.

Por el TCL se tiene que, al menos asintóticamente,

$$U \sim N(0, \mathcal{I}) \Rightarrow U' \mathcal{I}^{-1} U \sim \chi_p^2. \quad (2.21)$$

2.5.2. Distribución en el muestreo del $EMV(\beta)$

Cuando se encuentra un vector $\hat{\beta}$ solución del ajuste por MCPI, éste es consistente, asintóticamente normal y asintóticamente eficiente. La media de su distribución asintótica normal es β , y su matriz de varianzas-covarianzas la inversa de la matriz de información de Fisher, $\mathcal{I}^{-1}$, con

$$\mathcal{I}_{jk} = \sum_{i=1}^n \frac{x_{ij}x_{ik}}{a_i(\phi)V(\mu_i)[g'(\mu_i)]^2}$$

En concreto, para muestras grandes se tiene

$$WD = (\hat{\beta} - \beta)' \mathcal{I} (\hat{\beta} - \beta) \sim \chi_p^2, \quad (2.22)$$

donde WD recibe el nombre de *estadístico de Wald*, y equivalentemente

$$\hat{\beta} \sim N(\beta, \mathcal{I}^{-1}). \quad (2.23)$$

Demostración:

Supongamos que la función de log-verosimilitud tiene un único máximo en $\beta = \beta$ y que este estimador quede cerca del verdadero valor de β . Consideremos también la aproximación de primer orden del vector score,

$$U(\beta) = \underbrace{U(\hat{\beta})}_0 + H(\hat{\beta})(\beta - \hat{\beta}).$$

Asintóticamente $H(\hat{\beta}) \doteq E[H(\hat{\beta})] = -\mathcal{I}(\beta) \equiv -\mathcal{I}$, de donde para muestras grandes

$$U(\beta) \cong -\mathcal{I}(\beta - \hat{\beta}) \rightarrow \hat{\beta} - \beta \cong \mathcal{I}^{-1}U.$$

Tratando a $\mathcal{I}$ como constante, tenemos

$$E[\hat{\beta} - \beta] \cong \mathcal{I}^{-1}E[U] = 0 \Rightarrow E[\hat{\beta}] = \beta,$$

es decir, que $\hat{\beta}$ es un estimador insesgado de β .

La matriz de varianzas-covarianzas será

$$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^t] \cong \mathcal{I}^{-1}E[UU^t]\mathcal{I}^{-1} = \mathcal{I}^{-1}.$$

Así, por el TCL para muestras grandes,

$$WD = (\hat{\beta} - \beta)' \mathcal{I}(\hat{\beta} - \beta) \sim \chi_p^2,$$

o equivalentemente,

$$\hat{\beta} \sim N(\beta, \mathcal{I}^{-1}).$$

Nota 2.5.1 Excepto para los modelos lineales con variables respuesta con una distribución Normal, los resultados anteriores están apoyados en aproximaciones para muestras grandes. También la matriz de información de Fisher $\mathcal{I}$ depende a menudo de los parámetros β , de tal forma que para utilizarla necesitaremos evaluarla en $\beta = \hat{\beta}$. Ocasionalmente, $-H(\hat{\beta})$ se utilizará en lugar de $\mathcal{I}(\hat{\beta})$ como una estimación de $\mathcal{I}(\beta)$. Cuando n es grande, la diferencia entre $\mathcal{I}(\hat{\beta})$ y $-H(\hat{\beta})$ no será muy relevante; algunos prefieren $-H(\hat{\beta})$ porque no está basada en una medida promedio.

Resultado 2.5.1 Los elementos de la diagonal de la inversa de la matriz de información de Fisher (evaluada en el EMV) proporcionan una medida de la precisión obtenida en las estimaciones,

$$\text{Var}(\hat{\beta}_j) = v_{jj}.$$

donde v_{jj} es el elemento correspondiente en la matriz $\mathcal{I}^{-1}$.

Resultado 2.5.2 Los intervalos de confianza para las estimaciones $\hat{\beta}_j$ cuando el tamaño muestral es suficientemente grande, vienen dados por

$$\hat{\beta}_j \pm z_{\alpha/2} \sqrt{v_{jj}}, \quad (2.24)$$

siendo α el nivel de confianza deseado y $z_{\alpha/2}$ el cuantil $\alpha/2$ de la normal estándar.

Resultado 2.5.3 Podemos examinar la correlación entre las estimaciones utilizando la matriz $\mathcal{I}$,

$$\text{corr}(\hat{\beta}_j, \hat{\beta}_k) = \frac{v_{jk}}{\sqrt{v_{jj}v_{kk}}}.$$

Resultado 2.5.4 Cuando hacemos un contraste de hipótesis para eliminar (o no) una variable del modelo, por ejemplo $H_0 : \beta_j = 0$, podemos utilizar el estadístico de Wald

$$WD_j = (\hat{\beta}_j - \beta_j)^2 / v_{jj},$$

que tiene una distribución asintótica χ^2_1 , o lo que es igual, el estadístico de contraste

$$Z = \frac{\hat{\beta}_j}{\sqrt{v_{jj}}} \sim N(0, 1). \quad (2.25)$$

Notar que las distribuciones especificadas para WD y Z son aproximaciones asintóticas y por tanto sólo han de utilizarse como guía. En el caso de la respuesta normal, dichas distribuciones muestrales son exactas.

Si ϕ es desconocido, ha de ser estimado; en lugar del estadístico Z o del estadístico de Wald habremos de utilizar las versiones t de Student y F habituales. Esto se verá en la sección de Comparación de modelos.

Ejemplo 2.7 (Estimación en un modelo Poisson) Consideramos los datos en la Tabla 2.1, donde Y representa el número de casos observados para diversos valores de una covariable $X \in \{-1, 0, 1\}$. Podemos asumir entonces una distribución de Poisson para la respuesta y (Ver Ejercicio 6).

Vamos ajustar un modelo de la forma

$$g(\mu_i) = \mu_i = \eta_i = \beta_1 + \beta_2 x_i,$$

Y	2	3	6	7	8	9	10	12	15
X	-1	-1	0	0	0	0	1	1	1

Tabla 2.1: Datos Poisson.

es decir, explicamos $\mu = E(Y)$ con la variable x , y los relacionamos con el link identidad. Queremos calcular las estimaciones de los parámetros, sus errores estándar y correlaciones. Para ello necesitaremos la matriz de diseño X y calcular con ella la matriz de información de Fisher $\mathcal{I}$.

Para ajustar el modelo en $\mathcal{R}$, introducimos primeramente los datos con:

```
y <- c(2, 3, 6, 7, 8, 9, 10, 12, 15)
x <- c(-1, -1, 0, 0, 0, 0, 1, 1, 1)
```

y ajustamos el modelo propuesto solicitando también la matriz de diseño X con el argumento ' $x = T$ ':

```
fit <- glm(y ~ x, family = poisson(link = "identity"), x = T)
```

Los resultados del ajuste provienen de `summary(fit)` y son:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	7.4516	0.8841	8.428	< 2e - 16
x	4.9353	1.0892	4.531	5,86e - 06

Todos los coeficientes del modelo resultan significativos al 95% en función del estadístico de Wald (columna 'z-value' en la tabla anterior), como puede comprobarse a través de sus p -valores asociados (columna $Pr(> |z|)$). La recta ajustada tiene la forma:

$$E(Y_i) = 7,4516 + 4,9353 x_i.$$

La estimación del error estándar de las estimaciones viene dada en la columna 'Std. Error' de la tabla anterior, y son 0.8841 y 1.0892 respectivamente para β_1 y β_2 .

A partir de esto, es fácil construir intervalos de confianza para los parámetros del modelo:

$$\begin{aligned} IC(\beta_1, 95\%) &= 7,45 \pm 1,96(0,88) = (5,72, 9,18) \\ IC(\beta_2, 95\%) &= 4,93 \pm 1,96(1,09) = (2,79, 7,066) \end{aligned}$$

Si pretendemos sin embargo, calcular la matriz de varianzas-covarianzas completa, con el fin de calcular también la correlación entre las estimaciones, hemos de utilizar la matriz de información de Fisher;

$$\mathcal{I}_{jk} = \sum_i \frac{x_{ij}x_{ik}}{\text{Var}(Y_i)g'(\mu_i)^2} \Rightarrow \mathcal{I} = \begin{pmatrix} \sum_i \frac{1}{\beta_1 + \beta_2 x_i} & \sum_i \frac{x_i}{\beta_1 + \beta_2 x_i} \\ \sum_i \frac{x_i}{\beta_1 + \beta_2 x_i} & \sum_i \frac{x_i^2}{\beta_1 + \beta_2 x_i} \end{pmatrix}$$

Utilizando el Resultado 2.4.5 y la expresión en (2.18), tan sólo necesitamos la matriz de diseño X y la matriz W . La matriz de diseño X la recuperamos con `X <- fit$x`. Dado que $V(\mu_i) = \mu_i$ y $g()$ es el link identidad ($g'(\mu_i) = 1$), la matriz W tiene la forma: $W = \text{diag}\{a_i V(\mu_i) g'(\mu_i)^2\} = \text{diag}\{\mu_i\}$. La calculamos en $\mathcal{R}$ con:

$$W <- \text{diag}(\text{fitted}(fit))$$

Así, la matriz de información $\mathcal{I}$ y la matriz de varianzas-covarianzas de β , $\mathcal{I}^{-1}$, se obtienen con:

$$\begin{aligned} Ifisher &<- t(X) \%* \%solve(W) \%* \%X \\ Var.Cov &<- solve(Ifisher) \end{aligned}$$

y resultan, respectivamente,

$$\mathcal{I} = \begin{pmatrix} 1,57 & -0,55 \\ -0,55 & 1,04 \end{pmatrix}, \quad \text{Var.Cov} = \begin{pmatrix} 0,78 & 0,42 \\ 0,42 & 1,19 \end{pmatrix}$$

Así, el error estándar de las estimaciones y sus correlaciones resultan:

$$\begin{aligned} se(\hat{\beta}_1) &= \sqrt{0,7817} = 0,88, & se(\hat{\beta}_2) &= \sqrt{1,1863} = 1,09. \\ Corr(\hat{\beta}_1, \hat{\beta}_2) &= \frac{0,4166}{\sqrt{0,78}\sqrt{1,18}} = 0,43, \end{aligned}$$

2.5.3. Estadístico DEVIANCE: Bondad del ajuste

Un aspecto importante en el ajuste de un modelo es determinar si describe adecuadamente o no los datos observados. Cuando ajustamos un modelo lineal generalizado, juzgamos la adecuación del modelo comparando la verosimilitud del modelo ajustado con la verosimilitud del modelo saturado.

El modelo saturado es un modelo de forma similar al modelo propuesto que describe de modo perfecto los datos. Por tanto, tiene poca utilidad desde el punto de vista de ajuste de un modelo. Sin embargo, es útil para medir cómo un ajuste concreto se parece a un ajuste “perfecto”. El **modelo saturado** asociado a un modelo propuesto viene caracterizado por:

- utiliza la misma distribución para la respuesta (no necesariamente con los mismos parámetros)
- utiliza el mismo link
- el número de parámetros es igual al número de datos ($p = n$), y por lo tanto no quedan grados de libertad para los residuos.

Ejemplo 2.8 (Normal 1) Considerando $Y_1, \dots, Y_n$ v.a. normales, con $E(Y_i) = \mu_i$ y varianza común σ^2 , se propone un modelo en el que todas las medias μ_i son iguales, $\mu_1 = \dots = \mu_n = \mu$. En el modelo saturado se usa la misma distribución (normal), el mismo link (identidad), y hay un parámetro a estimar por cada dato ($\mu_i \rightsquigarrow y_i$):

$$\begin{aligned} \text{Modelo Propuesto } E(Y_i) &= \mu, \quad i = 1, \dots, n \rightsquigarrow \hat{\mu}_i = \bar{y} \\ \text{Modelo Saturado } E(Y_i) &= \mu_i, \quad i = 1, \dots, n \rightsquigarrow \tilde{\mu}_i = y_i. \end{aligned}$$

Ejemplo 2.9 (Normal 2) Considerando $Y_1, \dots, Y_n$ v.a. normales, con $E(Y_i) = \mu_i$ y varianza común σ^2 , se propone un modelo en el que las medias μ_i vienen determinadas por una variable explicativa x_i como $\mu_i = \beta_1 + \beta_2 x_i$. El modelo saturado es

el mismo del Ejemplo 2.8.

$$\begin{aligned}
 \text{Modelo Propuesto } E(Y_i) &= \beta_1 + \beta_2 x_i, \quad i = 1, \dots, n & \rightsquigarrow \hat{\mu}_i &= \hat{\beta}_1 + \hat{\beta}_2 x_i, \\
 & & \text{con } \hat{\beta} &= EMV(\beta), \\
 \text{Modelo Saturado } E(Y_i) &= \mu_i, \quad i = 1, \dots, n & \rightsquigarrow \tilde{\mu}_i &= y_i.
 \end{aligned}$$

Ejemplo 2.10 (Poisson) Considerando $Y_1, \dots, Y_n$ v.a. independientes que se distribuyen Poisson, con $E(Y_i) = \mu_i$. Se propone un modelo en el que todas las medias μ_i son iguales, $\mu_1 = \dots = \mu_n = \mu$. En el modelo saturado se usa la misma distribución (poisson), el mismo link (identidad), y hay un parámetro a estimar por cada dato ($\mu_i \rightsquigarrow y_i$):

$$\begin{aligned}
 \text{Modelo Propuesto } E(Y_i) &= \mu_i, \quad i = 1, \dots, n & \rightsquigarrow \hat{\mu}_i &= \bar{y} \\
 \text{Modelo Saturado } E(Y_i) &= \mu_i, \quad i = 1, \dots, n & \rightsquigarrow \tilde{\mu}_i &= y_i.
 \end{aligned}$$

Ejemplo 2.11 (Binomial) Considerando $Y_1, \dots, Y_n$ v.a. independientes que se distribuyen $Bi(n_i, \pi_i)$. Se propone un modelo en el que la probabilidad de éxito es similar en todas las pruebas, $\pi_1 = \dots = \pi_n = \pi$; en el modelo saturado se usa la misma distribución (binomial), el mismo link (identidad), y hay un parámetro a estimar por cada dato ($\pi_i \rightsquigarrow y_i$):

$$\begin{aligned}
 \text{Modelo Propuesto } E(Y_i) &= n_i \pi, \quad i = 1, \dots, n & \rightsquigarrow \hat{\mu}_i &= n_i \hat{\pi} \\
 & & \text{con } \hat{\pi} &= EMV(\pi), \\
 \text{Modelo Saturado } E(Y_i) &= n_i \pi_i, \quad i = 1, \dots, n & \rightsquigarrow \tilde{\mu}_i &= n_i \tilde{\pi}_i \\
 & & \text{con } \tilde{\pi}_i &= y_i / n_i.
 \end{aligned}$$

Supongamos que hemos ajustado un modelo $g(\mu) = X\beta$ que proporciona como estimaciones máximo-verosímiles $\hat{\beta}$. Consideremos el modelo saturado correspondiente, en cuyo ajuste se obtiene $EMV(\beta) = \hat{\beta}_{sat}$.

Para medir la distancia entre el modelo ajustado y el saturado (ajuste perfecto), podemos considerar el estadístico del cociente de verosimilitudes

$$\lambda = \frac{f(y; \hat{\beta}, \phi)}{f(y; \hat{\beta}_{sat}, \phi)}, \quad (2.26)$$

supuesto ϕ conocido.

Si el modelo ajustado es “pobre”, $f(y; \hat{\beta}, \phi)$ será mucho menor que $f(y; \hat{\beta}_{sat}, \phi)$ ($\rightsquigarrow \lambda < 1$). Si el modelo ajustado es “bueno”, describirá bien los datos y por lo tanto $f(y; \hat{\beta}, \phi) \cong f(y; \hat{\beta}_{sat}, \phi)$ ($\rightsquigarrow \lambda \approx 1$).

Equivalentemente a λ , podemos utilizar

$$S = -2\log\lambda = -2[l(\hat{\beta}; y, \phi) - l(\hat{\beta}_{sat}; y, \phi)], \quad (2.27)$$

donde l denota la log-verosimilitud para β con datos observados y , y parámetro de escala ϕ conocido. S se denomina **deviance escalada**.

Una característica de los modelos lineales generalizados es que el criterio de máxima-verosimilitud utilizado para obtener estimadores $\hat{\beta}$ coincide con el criterio de bondad de ajuste basado en la deviance: maximizar la log-verosimilitud es equivalente a minimizar la deviance (dado que la log-verosimilitud evaluada en los EMV del modelo saturado es una constante).

Si expresamos el modelo en forma canónica y $\hat{\theta}_{i,sat}$ y $\hat{\theta}_i$ denotan, respectivamente, las EMV para el modelo saturado y el modelo propuesto, podemos escribir S como:

$$\begin{aligned} S &= -2 \left[\sum_{i=1}^n \frac{y_i \hat{\theta}_i - b(\hat{\theta}_i)}{a_i(\phi)} + c(y_i; \phi) - \left(\frac{y_i \tilde{\theta}_i - b(\tilde{\theta}_i)}{a_i(\phi)} + c(y_i; \phi) \right) \right] \\ &= 2 \sum_{i=1}^n \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))}{a_i(\phi)}. \end{aligned} \quad (2.28)$$

Si además, $a_i(\phi) = a_i\phi$, entonces se puede escribir

$$S(\mathbf{y}; \hat{\mu}, \phi) = \frac{D(\mathbf{y}; \hat{\mu})}{\phi}, \quad (2.29)$$

con

$$D = 2 \sum_{i=1}^n \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))}{a_i}, \quad (2.30)$$

que no depende de ϕ y es el output *deviance* que proporcionan usualmente los paquetes estadísticos. D se denomina **deviance no escalada**.

Ejemplo 2.12 (Continuación del Ejemplo 2.8. Normal) Dada la log-verosimilitud

$$l(\beta; \mathbf{y}, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - \mu_i)^2,$$

se obtiene:

$$S = \frac{1}{\sigma^2} \sum_i (y_i - \bar{y})^2, \quad D = \sum_i (y_i - \bar{y})^2,$$

con $\bar{y} = \sum_i y_i / n$. El estadístico deviance está relacionado con la varianza muestral, $s^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2$, mediante

$$S = (n-1)s^2/\sigma^2 \sim \chi_{n-p}^2,$$

que es la distribución exacta de S .

Ejemplo 2.13 (Continuación del Ejemplo 2.10. Poisson) La log-verosimilitud para datos Poisson es

$$l(\mu; \mathbf{y}) = \sum_i y_i \log(\mu_i) - \sum_i \mu_i - \sum_i \log(y_i!).$$

Se obtiene pues:

$$S = 2 \sum_{i=1}^n y_i \log \left(\frac{y_i}{\bar{y}} \right),$$

con $\bar{y} = \sum_i y_i / n$.

Como en el modelo Poisson $\phi = 1$, entonces $S = D \sim \chi_{n-1}^2$.

Ejemplo 2.14 (Continuación del Ejemplo 2.11. Binomial) Dada la forma de la log-verosimilitud para datos binomiales,

$$l(\phi, \mathbf{y}) \propto \sum_i (y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i)),$$

se tiene:

$$S = D = 2 \sum_i y_i \log \left(\frac{y_i}{n_i \hat{\pi}_i} \right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i (1 - \hat{\pi}_i)} \right).$$

Cuando ϕ es conocida, la deviance S mide cómo de grande es la desviación del modelo ajustado respecto de los datos (modelo saturado). Si S es grande, el modelo ajustado proporciona un ajuste pobre. Un valor pequeño de S es indicador de un buen ajuste. Para decidir qué se considera “ S grande” y qué “ S pequeño”, es preciso utilizar la distribución en el muestreo del estadístico.

Cuando el modelo propuesto proporciona un buen ajuste de los datos, entonces

$$S \sim \chi_{n-p}^2, \text{ asintóticamente,} \quad (2.31)$$

donde n es el número de datos a ajustar y p el número de parámetros del modelo.

Ejemplo 2.15 (Continuación del Ejemplo 2.7) *En aquel ejemplo, con datos Poisson ($\phi = 1$) y un ajuste con el link identidad,*

$$E(Y_i) = \beta_1 + \beta_2 x_i,$$

se había obtenido $D = 1,8947$ con 7 grados de libertad. Comparando con una χ_7^2 , se tiene que $D < \chi_7^2(0,95)$, indicando que el modelo ajusta bien los datos.

A pesar de que la deviance tiene un valor limitado para evaluar la bondad del ajuste, es útil para comparar los ajustes de dos modelos anidados.

2.6. Comparación de modelos

Aunque un modelo con más parámetros proporcione un ajuste mejor que otro con menos parámetros, cabe plantearse si en realidad todos los parámetros estimados son necesarios. Esta cuestión se puede resolver con tests basados en la deviance y en el estadístico de Wald, siempre y cuando se trate de modelos anidados. Otros estadísticos como el C_p de Mallows o el AIC permiten la comparación más general de modelos no necesariamente anidados.

Dos modelos en competencia pueden ser comparados mediante la deviance cuando tienen la misma distribución y función link y sólo difieren en el número de parámetros, es decir, cuando se trata de modelos anidados.

Dos modelos M_1 y M son **anidados** ($M_1 \subset M$) si, siendo X_1 y X sus respectivas matrices de diseño, $M_1 : g(\mu) = X_1 \beta_1$ y $M : g(\mu) = X \beta$, se tiene que las p_1 columnas de X_1 están contenidas en las p columnas de X , con $p_1 < p$.

Sea pues $\beta' = (\beta'_1, \beta'_2)$, con $\dim(\beta_1) = p_1$ y $\dim(\beta) = p$. Se plantea pues, si es necesario estimar el vector paramétrico completo β , o si por el contrario basta con

estimar β_1 . El contraste a resolver es pues:

$$\begin{aligned} H_0 : \beta_2 &= 0 \\ H_1 : \beta_2 &\neq 0. \end{aligned}$$

Nos limitaremos al caso en que $a_i(\phi) = a_i\phi$ y consideramos dos situaciones: ϕ conocido y ϕ desconocido

2.6.1. Comparaciones con la deviance

Sean D_1 y D las deviances (no-escaladas) asociadas a los modelos ajustados M_1 y M ,

$$\begin{aligned} D_1 &= -2\phi [l(\hat{\beta}_1; \mathbf{y}) - l(\hat{\beta}_{sat})] \\ D &= 2\phi [l(\hat{\beta}; \mathbf{y}) - l(\hat{\beta}_{sat})]. \end{aligned}$$

Si ϕ es conocido. Puesto que tanto D_1/ϕ como D/ϕ tienen distribuciones χ^2 (cada uno bajo la hipótesis de que es un modelo adecuado), y además D es independiente de $D_1 - D$, se tiene que:

$$\frac{D_1 - D}{\phi} \sim \chi_{p-p_1}^2. \quad (2.32)$$

Para algunas distribuciones en que ϕ es conocido, como son la binomial o la Poisson ($\phi = 1$), el estadístico χ^2 de (2.32) es suficiente para llevar a cabo la comparación de modelos. Bastará con calcular el p-valor asociado al estadístico de (2.32) y concluir a favor del rechazo del modelo más sencillo si el resultado es significativo; si no lo es, la conclusión es la habitual: no hay evidencias para rechazar el modelo más simple propuesto en H_0 .

Ejemplo 2.16 (Continuación del Ejemplo 2.7) *En aquel ejemplo, con datos Poisson ($\phi = 1$) y un ajuste con el link identidad,*

$$E(Y_i) = \beta_1 + \beta_2 x_i,$$

se había obtenido $D = 1,8947$ con 7 grados de libertad. Comparando con el modelo nulo, $E(Y_i) = \beta_1$, para el que se obtiene una deviance de $D_1 = 18,42$ con 8 grados de libertad, se tiene que $D_1 - D = 16,5259$, que ha de ser comparado con una χ_1^2 . Puesto que $D_1 - D > \chi_1^2(0,95) = 3,841$, se rechaza el modelo nulo a favor del modelo ajustado $E(Y_i) = \beta_1 + \beta_2 x_i$.

Si ϕ es desconocido. Cuando ϕ es desconocido, ha de ser estimado.

A pesar de que ya fue presentado un estimador consistente de ϕ , (2.19), cuando se trabaja con deviances es conveniente utilizar un estimador alternativo, no consistente). Si el modelo propuesto proporciona un buen ajuste de los datos, se puede estimar ϕ con la **deviance media**,

$$\hat{\phi} = \frac{D}{n - p}, \quad (2.33)$$

donde D es la deviance no escalada ($D = \phi S$), y $E(S) \doteq n - p$ asintóticamente, cuando el modelo propuesto es adecuado.

$\hat{\phi}$ es práctico y adecuado como estimador de ϕ cuando $a_i(\phi) = a_i\phi$ y además se sabe que $\frac{1}{\phi} \times \min\{\frac{1}{a_1}, \dots, \frac{1}{a_n}\}$ es razonablemente grande.

En estas situaciones en que ϕ es desconocido y estimado, el test χ^2 es reemplazado por el test F construido a partir del cociente de dos distribuciones χ^2 independientes:

$$\frac{(D_1 - D)/\phi(p - p_1)}{D/\phi(n - p)} = \frac{(D_1 - D)/(p - p_1)}{D/(n - p)} \sim F_{p-p_1, n-p}. \quad (2.34)$$

Cuando los datos son normales, las distribuciones asintóticas de los estadísticos presentados son distribuciones exactas.

2.6.2. Comparaciones con el estadístico de Wald

Otra forma de comparar modelos anidados es a través de la distribución en el muestreo de los estadísticos máximo-verosímiles, en concreto basándose en el estadístico de Wald:

Si ϕ es conocido, entonces

$$WD = (\hat{\beta}_2 - \beta_2)' \mathcal{I}_{22}(\hat{\beta}_2 - \beta_2) \sim \chi_{p_2}^2,$$

o equivalentemente,

$$\hat{\beta}_2 \sim N(\beta_2, \mathcal{I}_{22}^{-1}).$$

En particular, cuando se trata de un contraste unidimensional, $\dim(\beta_2) = 1$, utilizar el estadístico de Wald es equivalente a utilizar el estadístico normal estándar:

$$Z = \frac{\hat{\beta}_2}{\sqrt{\text{Var}(\hat{\beta}_2)}} \sim N(0, 1). \quad (2.35)$$

Si ϕ es desconocido, ha de ser estimado, y la matriz de covarianzas $\mathcal{I}$ sustituida por una estimación $\hat{\mathcal{I}}$ (ver la Nota 2.5.1 y el resultado en (2.18)). Así, la distribución del estadístico de Wald se convierte en una F :

$$WD = (\hat{\beta}_2 - \beta_2)' \hat{\mathcal{I}}_{22}^{-1} (\hat{\beta}_2 - \beta_2) \sim F_{p_2, \nu},$$

y la del estadístico normal (2.35), en una t de Student con ν grados de libertad, de donde $\hat{\beta}_2 \sim St(\beta_2, \hat{\mathcal{I}}_{22}^{-1}, \nu)$. Los grados de libertad ν de las distribuciones t y F son los correspondientes a la deviance con la que se ha estimado ϕ (en general, $n - p$).

En particular, para contrastes en una dimensión, el estadístico normal (2.35) se convierte en

$$Z_t = \frac{\hat{\beta}_2}{\sqrt{\text{Var}(\hat{\beta}_2)}} \sim t_\nu. \quad (2.36)$$

Una desventaja de utilizar el test de Wald es que asume que la distribución del estadístico de contraste es normal, y por lo tanto simétrica. Sin embargo, esto puede no ser cierto, especialmente en muestras pequeñas, y por lo tanto el test puede resultar poco fiable. Una aproximación alternativa cuando ϕ es conocido, es basar este tipo de contrastes en intervalos de confianza para la deviance escalada, y utilizar (2.32).

Ejemplo 2.17 (Modelo Lineal Normal tratado como GLM) *Los datos en la Tabla 2.2 (Dobson, 1990, pag.11) corresponden a un estudio en el que semillas genéticamente iguales son asignadas (aleatoriamente), bien a un entorno enriquecido nutricionalmente (tratamiento), bien a condiciones estándar (control). Una vez han crecido todas las plantas, se recolectan, secan y pesan. El interés del análisis es investigar el efecto del tratamiento utilizado sobre el peso seco (en gramos) de las plantas en cuestión, esto es, si existen diferencias entre el peso de las plantas tratadas y el de las no tratadas.*

En este caso podemos asumir normalidad para la variable respuesta “ Y_{ij} = peso seco (gr) de la planta j en el grupo i ($= 1$ tratamiento, $= 2$ control)”

control	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
tratamiento	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69

Tabla 2.2: Datos sobre el crecimiento de plantas.

Lo resolvemos utilizando $\mathcal{R}$. En primer lugar, introducimos los datos observados para los individuos del grupo control (*ctl*) y para los del grupo tratamiento (*trt*). Creamos asimismo, una variable factor que etiqueta a estos individuos con *ctl* o *trt* según corresponde:

```
ctl <- c(4,17, 5,58, 5,18, 6,11, 4,50, 4,61, 5,17, 4,53, 5,33, 5,14)
trt <- c(4,81, 4,17, 4,41, 3,59, 5,87, 3,83, 6,03, 4,89, 4,32, 4,69)
peso <- c(ctl, trt)
n <- length(ctl)
tratamiento <- factor(rep(c("ctl", "trt"), c(n, n)))
```

El modelo en el que consideramos un efecto del tratamiento es:

$$E(Y_{ij}) = \beta_1 + \beta_2 I_i^T, \quad i = 1, 2; \quad j = 1, \dots, 10$$

con $I_1^T = 0$ en el grupo control e $I_2^T = 1$ en el grupo tratamiento, equivalente a:

$$E(Y_{ij}) = \mu_i, \quad \forall \quad j = 1, \dots, 10,$$

con $\mu_1 = \beta_1$ y $\mu_2 = \beta_1 + \beta_2$.

La sintaxis en $\mathbf{R}$ para conseguir el ajuste y una descripción del mismo es:

```
summary(fit1 <- glm(peso ~ tratamiento, family = "gaussian").
```

Puesto que el comando '*glm*' por defecto utiliza el modelo normal, bastaría con haber escrito

```
glm(respuesta ~ tratamiento)
```


El resumen del modelo consiste, primero en una descriptiva de los residuos deviance, y después en la estimación de los parámetros del modelo, junto con el error estándar, el estadístico de Wald (t) y el p -valor asociado a éste:

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0710	-0.4937	0.0685	0.2463	1.3690

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	5.0320	0.2202	22.850	9.55e-15
tratamientotrt	-0.3710	0.3114	-1.191	0.249

Es decir, la respuesta media en el grupo control es estimada con $\hat{\mu}_{1j} = 5,0320$, y en el grupo tratamiento con $\hat{\mu}_{2j} = 5,0320 - 0,3710 = 4,661$.

El estadístico de Wald para contrastar $\beta_i = 0$, $i = 1, 2$, es el t de Student:

$$\begin{aligned}
 t_i &= \hat{\beta}_i / \text{se}(\hat{\beta}_i) \sim t_{n-p}, \text{ con } n = 20, p = 2 \\
 &= \begin{cases} 5,0320/0,2202 = 22,850 & \text{para } \beta_1 \\ -0,3710/0,3114 = -1,191 & \text{para } \beta_2, \end{cases}
 \end{aligned}$$

cuyos p -valores asociados aparecen en la última columna de la Tabla Coefficients. El p -valor asociado a la comparación de los grupos, $H_0 : \beta_2 = 0$ (coeficiente 'tratamientotrt'), resulta $p = 0,249$. Así, los datos no evidencian diferencias significativas entre los individuos que recibieron el tratamiento y los que quedaron como controles.

Por otro lado, es preciso estimar el parámetro de dispersión ϕ , que en la familia normal es $\phi = \sigma^2$; la estimación se obtiene con la deviance media $D/n - p$, aproximadamente:

(Dispersion parameter for gaussian family taken to be 0.4849583)

Por último, aparecen la deviance del modelo nulo (Null Deviance), que es el modelo que ajusta todos los datos con la misma media, $E(Y_{ij}) = \mu$, $\forall i, j$, y la del

modelo ajustado (Residual deviance), junto con los grados de libertad asociados al estadístico chi-cuadrado, que son:

Null deviance: 9.4175 on 19 degrees of freedom
Residual deviance: 8.7293 on 18 degrees of freedom

Junto a estas deviances se presentan el AIC (AIC : 46,176), que es un indicador de la bondad del ajuste (lo veremos en la Sección 2.6.3), y el número de iteraciones necesarias para obtener la solución utilizando el método Scoring de Fisher para estimar (en este caso, sólo 2 iteraciones).

Para resolver el contraste $H_0 : \beta_2 = 0$ con la deviance, a través de una comparación de modelos, hay que ajustar un modelo más simple SIN efecto tratamiento, que en este caso coincide con el modelo nulo, es decir,

```
summary(fit2 <- update(fit1, . ~ . - tratamiento))
```

equivalente a

```
summary(fit2 <- glm(respuesta ~ 1))
```

El estadístico F basado en las deviances, adecuado para hacer la comparación de modelos ya que el parámetro de escala $\phi = \sigma^2$ es desconocido,

$$\frac{(D_2 - D_1)/(2 - 1)}{D_1/(20 - 2)} \sim F_{1,18}$$

se obtiene con el comando

```
anova(fit2, fit1, test = "F")
```

que proporciona la salida:

Analysis of Deviance Table:

Response: respuesta

	<i>Resid. Df</i>	<i>Resid. Dev</i>	<i>Df</i>	<i>Deviance</i>	<i>F</i>	<i>Pr(> F)</i>
<i>1</i>	19	9.4175				
<i>tratamiento</i>	18	8.7293	1	0.6882	1.4191	0.249

El p -valor asociado a la comparación de modelos coincide con el del test basado en el estadístico de Wald. Los dos tests son equivalentes siempre para la comparación de dos grupos en el modelo lineal normal.

2.6.3. Comparaciones con los estadísticos C_p y AIC .

Existen varias herramientas más para la comparación de modelos en situaciones no tan restrictivas como las de los modelos anidados. Los estadísticos C_p de Mallows y AIC (Akaike Information Criteria) permiten comparaciones de modelos más generales, y por supuesto, resultan útiles en la búsqueda de modelos parsimoniosos cuando el número de variables explicativas disponibles es grande. Estos estadísticos consideran, además de la influencia en la deviance que tienen las diversas variables explicativas, el número de parámetros a estimar. El objetivo es conseguir modelos parsimoniosos (sencillos): que expliquen mucho (poca deviance) a poco coste (pocos parámetros).

En un modelo lineal normal, el **estadístico C_p de Mallows** toma la forma:

$$C_p = RSS_p/s^2 - n + 2p,$$

donde n es el número de datos, RSS_p es la suma de cuadrados residual para un modelo con p parámetros estimados, y s^2 un estimador insesgado de la varianza, obtenido como la suma de cuadrados residual media cuando se consideran en el modelo todas las variables explicativas disponibles.

Cuanto mayor sea el número de variables explicativas en el modelo, más próximo a 1 resultará el cociente RSS_p/s^2 , pero el valor del estadístico C_p recibirá mayor penalización ($2p - n$) por incorporar más variables. Cuando el modelo ajustado con p parámetros es adecuado, entonces $E(RSS_p) = (n-p)\sigma^2$; por otro lado, $E(s^2) = \sigma^2$, con lo que se tendrá que $E(C_p) = p$. Modelos adecuados son diagnosticados pues, cuando el valor del C_p se aproxima a p .

El **estadístico AIC** (Akaike information criterion) es una versión del estadístico C_p en términos de la verosimilitud. En los modelos lineales generalizados con un parámetro de dispersión ϕ conocido, se calcula como:

$$AIC = D + 2 p \hat{\phi}, \quad (2.37)$$

siendo D la deviance del modelo ajustado y p el número de parámetros estimados, o lo que es lo mismo, el número de grados de libertad consumidos por el modelo. Los cambios en el AIC debidos añadir o quitar un término en un modelo, provienen tanto del cambio ocasionado en la deviance, como de la dimensión del término de diferencia (a menudo un término involucra más de un grado de libertad). Cuando se comparan modelos en los que ϕ es estimado, es importante mantener $\hat{\phi}$ constante.

La relación, aproximada, entre los estadísticos C_p y AIC en el modelo normal viene dada por:

$$AIC = \hat{\sigma}^2(C_p + n). \quad (2.38)$$

Si pretendemos iniciar un procedimiento de tipo STEPWISE para conseguir ajustar un modelo parsimonioso, es claro que el criterio ha de estar basado en añadir términos siempre que el ajuste mejore sustancialmente y eliminar términos que no contribuyan a una mejora importante, considerando por supuesto, el equilibrio respecto al coste del número de parámetros a estimar.

Si se procede según un criterio de búsqueda BACKWARD, en cada paso se elimina aquella variable que proporciona una reducción más importante en el C_p (AIC). Si se procede hacia adelante (FORWARD), se introduce el término que reduce más el valor del estadístico C_p (AIC). Todos los procedimientos se detienen cuando se llega a un modelo “frontera”, o cuando ningún paso más hace decrecer el C_p (AIC).

En $\mathcal{R}$ hay varios comando útiles para realizar una búsqueda automática o manual de las variables relevantes en la explicación del modelo según el criterio AIC. Estos comandos son *drop1*, *add1* y *step*.

El comando, *drop1* (respectivamente *add1*), realiza la comparación del modelo actual con todos los que se obtienen eliminando (respectivamente añadiendo) una sola de las variables candidatas a desaparecer (respectivamente ser incluida) del modelo. Estos comandos permiten además la comparación de modelos mediante tests Chi-cuadrado y F. Proporcionan por defecto para cada contraste, los valores de la deviance (*Deviance*), el estadístico del cociente de verosimilitudes (*LRT*), los grados

de libertad (Df) asociados y el AIC del modelo que surge. El analista es el ha de ajustar manualmente en función de los resultados que observa.

El procedimiento de $\mathcal{R}$ denominado *step* resuelve automáticamente la búsqueda de un modelo parsimonioso bajo el criterio del AIC, procediendo de manera secuencial hacia adelante y hacia atrás a la vez, o en un único sentido, elegible con el argumento *direction=c("both", "backward", "forward")*; por defecto funciona hacia atrás (ver ayuda en $\mathcal{R}$). El procedimiento se detiene cuando la eliminación o inclusión de cualquier término no proporciona una reducción en el AIC.

Hay que tener en cuenta que un efecto principal no es candidato a desaparecer del modelo hasta que no han desaparecido todos los términos de interacción con otras variables.

Es importante anotar que en la selección de variables en un modelo, se puede optar por seguir el criterio AIC (o C_p), o bien el de significatividad de las variables (tests basados en la deviance de modelos anidados). Cada uno conducirá a resultados distintos la mayoría de las veces. Eso sí, el AIC será siempre lícito para la comparación de modelos no-anidados, mientras que la deviance no permitirá ese tipo de comparaciones.

Ejemplo 2.18 (Torus Mandibularis: Selección de variables) Consideremos el ajuste de los datos que se presentan en el Ejercicio 6 del Tema 4, (Muller y Mayhall, 1970, también en Krzanowski, 1998, pag.230). Estos datos versan sobre la incidencia de un rasgo morfológico denominado Torus Mandibularis, que consiste en una pequeña protuberancia en la barbilla, y que afecta a tres tribus de esquimales. Se han considerado diversas variables que pudieran tener más o menos importancia sobre la incidencia de esta malformación. Pretendemos obtener el modelo más parsimonioso (simple) posible.

Las cuatro variables observadas son:

inc: incidencia de la malformación: la variable a explicar, con dos opciones de respuesta, que son Sí $\mapsto$ presencia de Torus mandibularis y No $\mapsto$ ausencia.

pop: tribu esquimal, explícitamente controlada para comparar la incidencia de la malformación en las diversas tribus consideradas. De hecho, se tomó un tamaño muestral fijo para cada grupo poblacional, lo que obliga a incluir esta variable en cualquier modelo que planteemos (como veremos en el Tema 4).

sex: el sexo del individuo: $H \mapsto$ hombre y $M \mapsto$ mujer.

edad: categorizada en los grupos $[1, 10]$, $[11, 20]$, $[21, 30]$, $[31, 40]$, $[41, 50]$ y $[51, +]$.

La modelización de este tipo de datos se hace a través de modelos con verosimilitud Poisson para los conteos de las celdas, en la variable 'conteos', y link logaritmo, como veremos en el Tema 4. Todas las variables de clasificación y las interacciones entre ellas pueden servir en principio para predecir los conteos. Como hay muchos posibles modelos (si consideramos todas las posibilidades a la hora de incluir las variables de clasificación y sus interacciones), podemos proceder, bien hacia adelante partiendo del modelo más simple posible, bien hacia atrás desde el modelo más "grande" sin ser saturado. Seguiremos este último procedimiento.

El ajuste se lleva a cabo según una modelización de Poisson para predecir las frecuencias esperadas en las celdas a través de los factores de clasificación y sus interacciones. Excluyendo la interacción de orden 4 (modelo saturado) tenemos:

```
fit <- glm(conteos ~ (pop + sex + edad + inc)^3, family = poisson)
```

La Deviance de este modelo resulta $D = 21,607$, con 10 grados de libertad, y un $AIC = 387,4$.

Ahora bien, ¿podemos eliminar algún término poco relevante del modelo? Para contestar esta pregunta podemos seguir dos criterios: 1) basarnos en la significatividad asociada a cada término (comparación de deviances) 2) basarnos en la reducción del AIC (o Cp). Procederemos de las dos formas para ilustrar el proceso.

Si optamos por el criterio basado en la significatividad de los términos del modelo, con el comando `drop1(fit, test = "Chisq")` obtenemos la tabla:

	Df	Deviance	AIC	LRT	Pr(Chi)
none		21.61	387.40		
pop:sex:edad	10	32.17	377.96	10.56	0.39254
pop:sex:inc	2	23.60	385.40	2.00	0.36873
pop:edad:inc	10	42.27	388.06	20.66	0.02358
sex:edad:inc	5	24.93	380.73	3.33	0.64958

con lo que el primer candidato a salir sería el término '*sex : edad : inc*', ya que es el menos significativo. Si utilizáramos como criterio el AIC, el primer candidato a salir del modelo sería '*pop : sex : edad*'. Ya vemos que los resultados según uno y otro criterio empiezan a diferir:

Para llegar al modelo óptimo con sólo las variables significativas habríamos de continuar con el proceso, excluyendo términos en función del resultado de '*drop1*', y añadiendo (si se hace una búsqueda adelante-atrás) en función del de '*add1*'.

Procediendo siempre hacia atrás con '*drop1*' y utilizando el criterio de significatividad en las comparaciones con la deviance, obtenemos finalmente el modelo:

$$\text{conteos} \sim \text{pop} + \text{edad} + \text{inc} + \text{pop} : \text{edad} + \text{pop} : \text{inc} + \\ + \text{edad} : \text{inc} + \text{pop} : \text{edad} : \text{inc},$$

según el cual, son significativas para explicar los conteos en las celdas, las variables '*pop*', '*edad*' e '*inc*', y las interacciones entre ellas (el sexo del individuo no influye). El AIC del ajuste es $AIC = 357,7$ y la deviance de $D = 43,875$ con 36 grados de libertad. Según esto, la malformación está asociada con la edad y la tribu; además, la edad no afecta de la misma forma en todas las tribus (como se deduce de los términos de interacción con la variable '*inc*').

• Si optamos por el criterio AIC para obtener el modelo más parsimonioso, podemos utilizar el procedimiento automático de $\mathcal{R}$ denominado *step*,

$\text{fit.aic} <- \text{step}(\text{fit})$

que proporciona el modelo:

$$\text{conteos} \sim \text{pop} + \text{sex} + \text{edad} + \text{inc} + \text{edad} : \text{inc},$$

con un $AIC = 347,09$ y una deviance de $D = 75,30$ con 57 grados de libertad. Eliminar cualquier otro término elevaría el valor del AIC, como se aprecia en la siguiente tabla:

	<i>Df</i>	<i>Deviance</i>	<i>AIC</i>
<i>none</i>		75.30	347.09
- <i>sex</i>	1	78.11	347.91
- <i>pop</i>	2	215.86	483.66
- <i>edad:inc</i>	5	225.74	487.53

Es decir, según esto, la malformación sólo está asociada con la edad, a diferencia de la conclusión que obteníamos utilizando la significatividad asociada a la comparación de deviances.

2.7. Chequeo del Modelo Ajustado: Residuos

Los residuos se pueden utilizar para explorar la adecuación del ajuste de un modelo respecto a la elección de la función de varianza, la función link y los términos a incluir en el predictor lineal. Los residuos pueden también ser útiles para detectar observaciones influyentes o valores anómalos que requieran una investigación más intensa. Para modelos lineales generalizados, requerimos una definición algo más amplia de *residuos*, que sea aplicable a todas las distribuciones que pueden reemplazar a la Normal.

Los gráficos de residuos pueden ser muy útiles a la hora de detectar fallos sistemáticos en un modelo para ajustar unos datos. Detectados los fallos, la labor siguiente consiste en solventar las deficiencias y reajustar el modelo. Sin embargo, cuando la respuesta toma sólo unos pocos valores, como ocurre en pruebas Bernoulli, la utilidad de los residuos es limitada.

La especificación incorrecta de un modelo se puede deber a una serie de factores:

- elección incorrecta de la distribución de probabilidad,
- especificación incorrecta de la forma en que la respuesta media cambia con las variables explicativas, ya sea porque
 - la componente sistemática $\eta()$ está mal especificada,
 - o la función link $g()$ no es apropiada,
- variables explicativas no incorporadas en el modelo,
- funciones incorrectas de las variables explicativas en el modelo, incluyendo interacciones desconocidas entre ellas, o no disponibilidad de suficientes funciones diferentes,
- dependencia entre las observaciones, por ejemplo, a lo largo del tiempo (autocorrelación).

Veamos, sobre el modelo normal, la forma de calcular los residuos, para luego generalizarla a los GLM.

En el modelo lineal normal, podemos expresar la variable dependiente de la forma:

$$y = \hat{y} + (y - \hat{y}) = Hy + \underbrace{(I - H)y}_{\text{residuos}},$$

es decir, *datos* = *valores ajustados* + *residuos*, con $H = X(X'X)^{-1}X'$ la **matriz gorro**. Esta matriz es idempotente y simétrica; su traza es igual a p , que es la dimensión del vector de parámetros, y $0 \leq h_{ii} \leq 1$.

Además, en el modelo normal, los residuos R verifican

$$E(R) = 0 \quad \text{y} \quad \text{Var}(R) = (I - H)\sigma^2.$$

Podríamos pensar que podemos obtener residuos con varianza unidad simplemente dividiéndolos por s , donde s^2 es la varianza residual que se obtiene al ajustar el modelo en cuestión. Sin embargo, de la expresión anterior, es claro que la varianza de R no es fija sino que varía cuando varía X . Por lo tanto, para obtener residuos con varianza unidad, estos han de ser corregidos también por las X . Así, surge el *residuo estudentizado* como:

$$r_i = \frac{y_i - \hat{y}_i}{s\sqrt{1 - h_{ii}}},$$

donde h_{ii} es el i -ésimo elemento de la diagonal de H . Así, cuando el residuo estudentizado de un dato es grande (por ejemplo, mayor que 2), se dice que dicho dato puede ser un outlier.

El **valor leverage** h_{ii} proporciona una medida de lo remota que resulta, en el espacio de las variables X , la i -ésima observación de las $n - 1$ restantes. Así, los valores leverage resultan diagnósticos útiles cuando buscamos observaciones que podrían tener una influencia indebida en el ajuste del modelo. Como la media de los valores h_{ii} es p/n , una regla de salto fácil de aplicar es considerar como “potencialmente influyente” cualquier observación que tiene $h_{ii} > 2p/n$, y candidata por tanto, a ser revisada con atención.

El recíproco de un elemento h_{ii} de la diagonal de H , $1/h_{ii}$, llamado la **replicación efectiva**, se puede interpretar, a grosso modo, como el número de observaciones que proporcionan información sobre $\hat{y}_i$.

En un GLM, la matriz gorro del ajuste viene dada en el procedimiento de MCPI (2.15) por:

$$H = W^{-1/2} X (X' W^{-1} X)^{-1} X' W^{-1/2}, \quad (2.39)$$

que depende del valor de los parámetros a través de $W = \text{diag}\{Var(Y_i)g'(\mu_i)^2\}_{i=1}^n$. Así pues, habrá de ser evaluada sobre las estimaciones máximo-verosímiles obtenidas en el ajuste. Esta matriz gorro la utilizaremos para estandarizar residuos.

2.7.1. Tipos de residuos

Los análogos a los *residuos estudentizados de Pearson* en el caso de la normal son los que llamamos **residuos de Pearson**:

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{Var}(Y_i)}}. \quad (2.40)$$

La suma de sus cuadrados es el *estadístico chi-cuadrado de Pearson*,

$$\chi^2 = \sum_i (r_i^P)^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\widehat{Var}(Y_i)}.$$

El estadístico de Pearson se utiliza, más que como un estadístico de bondad de ajuste, como una medida de variación residual.

Puesto que la variabilidad de los residuos depende de las variables explicativas X , al igual que en el modelo normal, corregimos su efecto dividiendo por $\sqrt{1 - h_{ii}}$. Obtenemos así los **residuos de Pearson estandarizados**:

$$r_i^{PS} = \frac{r_i^P}{\sqrt{1 - h_{ii}}}. \quad (2.41)$$

Ejemplo 2.19 En datos binomiales, el residuo de Pearson (no estandarizado) es:

$$r_i^P = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}}.$$

Ejemplo 2.20 En datos Poisson, el residuo de Pearson es:

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

Una desventaja de los residuos de Pearson es que su distribución para respuestas no-normales es a menudo, marcadamente asimétrica, y por lo tanto pueden diferir sus propiedades de las de los residuos del modelo lineal normal.

Anscombe propuso definir otro tipo de residuos cuya distribución fuera lo más Normal posible (cuando el modelo era adecuado). Wedderburn (ver Barndorff-Nielsen, 1978) definió esos residuos a través de una transformación de la forma

$$A(\cdot) = \int \frac{1}{V^{1/3}(\mu)} d\mu.$$

Se definen los **residuos de Anscombe** como:

$$r_i^A = \frac{A(y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i) \sqrt{V(\hat{\mu}_i)}}.$$

Ejemplo 2.21 En datos binomiales,

$$V(\mu) = n\pi(1 - \pi) = \mu(n - \mu)/n,$$

y el residuo de Anscombe resulta:

$$r_i^A = \frac{3(y_i^{1/3} - (n_i \hat{\pi}_i)^{1/3})}{(n_i \hat{\pi}_i)^{1/3}}.$$

Ejemplo 2.22 En datos Poisson, $V(\mu) = \mu$ y $A(\mu) = (3/2)\mu^{2/3}$, y el residuo de Anscombe es:

$$r_i^A = \frac{3/2(y_i^{2/3} - \hat{\mu}_i^{2/3})}{\hat{\mu}_i^{1/6}}.$$

Si la deviance se utiliza como una medida de discrepancia para un modelo lineal generalizado, entonces cada unidad contribuye una cantidad D_i a dicha medida, de forma que $\sum_i D_i = D$,

$$D_i = 2 \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{a_i}.$$

Definimos pues, el **residuo deviance** como:

$$r_i^D = \text{sgn}(y_i - \hat{\mu}_i) \sqrt{D_i}, \quad (2.42)$$

con ϕ estimada y $\text{sgn}(y_i - \hat{\mu}_i)$ el signo de dicha diferencia. Tenemos pues que

$$\sum_{i=1}^n (r_i^D)^2 = D,$$

donde D es la *deviance no escalada* para el modelo ajustado.

El **residuo deviance estandarizado** se obtiene con:

$$r_i^{DS} = \frac{r_i^D}{\sqrt{1 - h_{ii}}}. \quad (2.43)$$

Los residuos Deviance estandarizados hacen resaltar cualquier observación que “contribuye demasiado” a la deviance del modelo.

El **residuo verosimilitud** se define como:

$$r_i^L = \text{sgn}(y_i - \hat{\mu}_i) \sqrt{h_{ii}(r_i^{PS})^2 + (1 - h_{ii})(r_i^{DS})^2}. \quad (2.44)$$

La utilidad de este residuo reside en el hecho de que $(r_i^L)^2$ es aproximadamente igual al cambio en la deviance escalada que resulta cuando eliminamos la i -ésima observación en el ajuste. Así, la sensibilidad del ajuste de un modelo a la eliminación de la i -ésima observación se puede calcular considerando el tamaño de r_i^L .

Con el residuo r_i^L podemos definir un **estadístico de Cook modificado**:

$$D_i = |r_i^L| \sqrt{\frac{(n - p)h_{ii}}{p(1 - h_{ii})}}, \quad (2.45)$$

con n el número de datos y p el número de parámetros estimados en el modelo.

2.7.2. Residuos en $\mathcal{R}$

En *SPlus/R* tenemos varias opciones para conseguir residuos. Una vez hemos ajustado un modelo lineal generalizado con el comando *glm*,

$$\boxed{\text{fit} <- \text{glm}(\text{respuesta} \sim \text{covariables}, \text{family} = ())}$$

podemos calcular los residuos:

pearson: Se obtienen con

```
residuals(fit, type = "pearson")
```

deviance: Se obtienen con

```
residuals(fit, type = "deviance")
```

o simplemente

```
residuals(fit)
```

working: Son los residuos ponderados por la derivada de la función link,

$$r_i^W = (y_i - \hat{\mu}_i) \frac{\partial \hat{\eta}_i}{\partial \mu_i}.$$

Se obtienen con

```
residuals(fit, type = "working")
```

o simplemente

```
fit$residuals
```

response: Son simplemente las diferencias entre los valores observados y sus predicciones con el modelo ajustado, $y_i - \hat{\mu}_i$. Se obtienen con:

```
residuals(fit, type = "response")
```

A partir de ellos podemos calcular cualesquiera de los residuos antes especificados y así llevar a cabo un chequeo completo del modelo ajustado.

Ejemplo 2.23 (Continuación del Ejemplo 2.7: Residuos) Pretendemos a continuación, construir los residuos de Pearson, los estandarizados de Pearson, Los deviance, los deviance estandarizados, los residuos verosimilitud y las distancias de Cook modificadas. Comentaremos el ajuste del modelo en función de los residuos y determinaremos qué observaciones son las más influyentes en el cálculo de la deviance. Utilizamos todos los resultados obtenidos en el Ejemplo 2.7.

Para estandarizar los residuos necesitamos la matriz H , que se calcula según (2.39), utilizando la matriz de diseño X y la matriz $W^{-1/2}$. Tanto X como W habían sido calculadas en el Ejemplo mencionado:

```
iW.sqrt <- -sqrt(solve(W))
H <- -(iW.sqrt %* %X) %* %Var.Cov %* %(t(X) %* %iW.sqrt)
```

de donde obtenemos una matriz 9×9 , cuya diagonal son los valores leverage:

```
h.leverage <- -diag(H)
```

es decir, $h = (0,451, 0,451, 0,105, 0,105, 0,105, 0,105, 0,226, 0,226, 0,226)$. Utilizando como regla de salto $h > 2p/n - 0,44$, la primera y segunda observación son las que resultan más remotas respecto de las restantes observaciones, en el espacio de las variables X (el valor de su x es -1, frente a 0 y 1 que es el valor de las restantes observaciones).

A continuación, calculamos los residuos estandarizados y las distancias de Cook modificadas. Las diferencias entre los distintos tipos de residuos son muy pequeñas, como puede apreciarse en la Figura 2.1. Se puede hablar de aleatoriedad de los residuos. Sus magnitudes son además, considerablemente pequeñas. Las observaciones más influyentes desde la perspectiva de los residuos, son la sexta y la novena (más la última): $(Y, x) = (9, 0)$ y $(15, 1)$, con estimaciones respectivas de $\hat{\mu} = 7,45$ y $12,39$. Desde el punto de vista de las distancias de Cook, las más influyentes son la séptima $(Y, x, \hat{\mu}) = (10, 1, 12,38)$ y la novena. La observación tal que al prescindir de ella en el ajuste modifica más la deviance del modelo es, como puede apreciarse en el gráfico de residuos Deviance, la novena (la que tiene un mayor residuo deviance).

La sintaxis utilizada para calcular y representar los residuos es:

```
par(mfrow=c(2,2))
plot(pear.std<-resid(fit1,type="pearson")/sqrt(1-h.leverage),
     xlab="",ylab="",main="Res.Pearson std")
plot(dev.std<-resid(fit1,type="deviance")/sqrt(1-h.leverage),
     xlab="",ylab="",main="Res.Deviance std")
plot(vero<-sign(res-fitted(fit1)) * sqrt(h.leverage*pear.std^2+
     (1-h.leverage)*dev.std^2), xlab="",ylab="",main="Res.Verosim")
plot(abs(vero)*sqrt((length(res)-2)*h.leverage/(2*(1-h.leverage))),
     xlab="",ylab="",main="Cook modif")
```

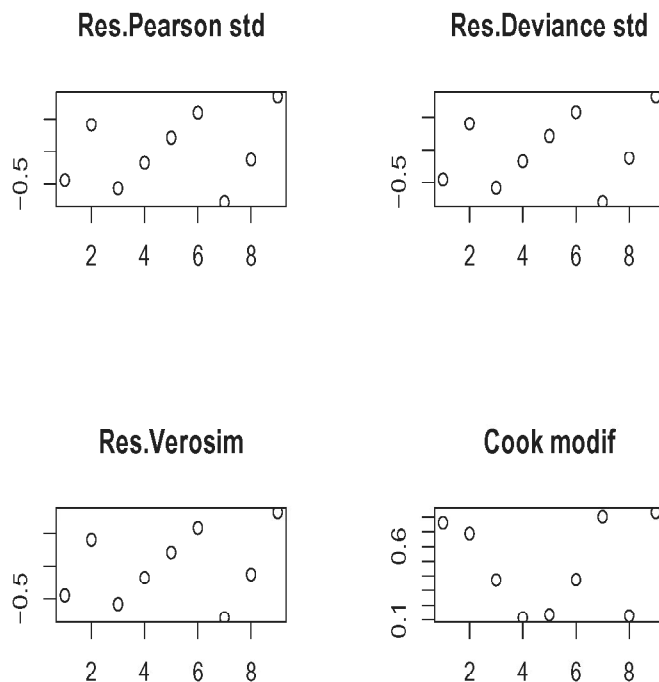


Figura 2.1: Residuos estandarizados para datos Poisson.

2.8. Análisis de la Covarianza (ANCOVA). Ejemplos

Una vez introducida toda la teoría general sobre modelos lineales generalizados, consideramos el ajuste de un modelo lineal normal como una particularización de los GLM. En concreto, a través de modelos de respuesta normal en los que algunas de las variables explicativas son continuas y otras son factores de clasificación. El objetivo básico del análisis de este tipo de datos, conocido usualmente como ANCOVA, es investigar si el factor (o factores) de clasificación influye en la relación entre la variable respuesta y las explicativas continuas.

Asumiremos normalidad en la repuesta para los ejemplos que se presentan a continuación.

Ejemplo 2.24 (Pesos al nacer) *Los datos en la Tabla 2.3 (Dobson, 1990, pag.17) representan los pesos al nacer (en gramos) y los períodos de gestación estimados (en semanas) para 24 bebés –12 varones y 12 hembras–, nacidos en cierto hospital. Las edades gestacionales medias son casi las mismas para los dos sexos, pero el peso medio al nacer para varones resulta mayor que el de hembras. La representación de los datos sugiere cierta tendencia lineal para explicar el peso al nacer con la edad gestacional. La pregunta de interés es si la razón de incremento es igual para varones y para hembras. Ver Figura 2.2, donde los varones están identificados por 'H' y las hembras por 'M'. La normalidad de la respuesta queda patente con el histograma.*

Varones		Hembras	
Edad (semanas)	Peso al nacer (g)	Edad (semanas)	Peso al nacer (g)
40	2968	40	3317
38	2795	36	2729
40	3163	40	2935
35	2925	38	2754
36	2625	42	3210
37	2847	39	2817
41	3292	40	3126
40	3473	37	2539
37	2628	36	2412
38	3176	38	2991
40	3421	39	2875
38	2975	40	3231

Tabla 2.3: Peso al nacer y edad gestacional para bebés varones y hembras

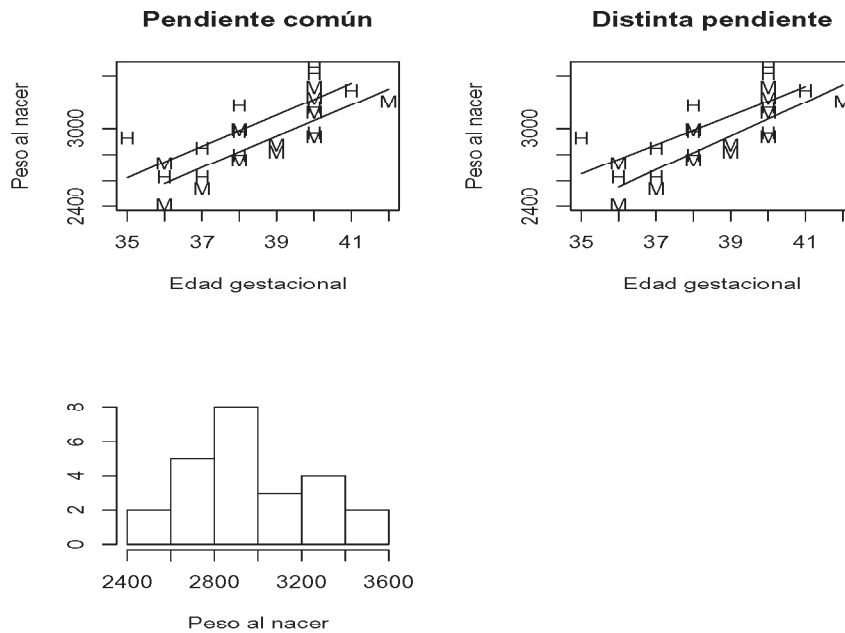


Figura 2.2: Datos de Peso al nacer.

MODELIZACIÓN:■ *Variables:*

– Respuesta $\rightsquigarrow Y_{ij} \equiv$ peso al nacer del individuo j en el grupo $i = 0$ (varones), 1 (hembras).

– Predictores:

Factor $\rightsquigarrow$ sexo: $i = 0$ (varones), $i = 1$ (hembras).

Covariable $\rightsquigarrow x_{ij} \equiv$ edad estimada (en semanas) al nacer del individuo j en el grupo i .

■ *Hipótesis distribucional:*

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2).$$

■ *Modelos propuestos:*

$$\begin{aligned} \text{misma recta} &\rightsquigarrow M1 : \mu_{ij} = \mu + \beta x_{ij} \\ \text{rectas paralelas} &\rightsquigarrow M2 : \mu_{ij} = \mu + \alpha_i + \beta x_{ij} \\ \text{rectas distintas} &\rightsquigarrow M3 : \mu_{ij} = \mu + \alpha_i + \beta_i x_{ij}. \end{aligned}$$

■ *Ajustes (en R):*

```
M1 <- glm(peso ~ edad)
M2 <- glm(peso ~ sexo + edad)
M3 <- glm(peso ~ sexo + edad + sexo : edad).
```

■ *Comparación de modelos:*

anova(M1, M2, M3, test = "F")						
	Res.Df	Res.Dev	Df	Deviance	F	Pr(> F)
edad	22	816074				
sexo + edad	21	658771	1	157304	5	0.04006
sexo * edad	20	652425	1	6346	1.945e-01	0.6639

El contraste $H_0 : M1$ versus $H_1 : M2$ tiene un p -valor asociado de 0.040, con lo que se rechaza H_0 a favor de H_1 , demostrando la significatividad de la variable sexo para distinguir entre el peso medio de los varones y el de las hembras al nacer.

El contraste $H_0 : M2$ versus $H_1 : M3$, plantea concluir sobre si tienen igual pendiente las dos rectas con las que predecir el peso al nacer de varones y hembras. El p -valor asociado vale 0.664, con lo que no se consigue significatividad para rechazar H_0 y se concluye a favor del modelo de rectas paralelas $M2$.

Ejemplo 2.25 (Consumo de azúcar) En la Tabla 2.4 (Dobson, 1990, pag.81) se muestran los promedios aparentes de consumo per capita de azúcar (en kg por año) en Australia. Están diferenciadas las cifras según el consumo sea en azúcar refinado o dedicado a la manufactura de alimentos (del Australian Bureau of Statistics, publication 4306.0).

Consumo de azúcar	Años					
	36-39	46-49	56-59	66-69	76-79	83-86
Refinado	32.0	31.2	27.0	21.0	14.9	9.9
En manufacturas	16.3	23.1	23.6	27.7	34.6	34.5

Tabla 2.4: Consumo de azúcar como refinado o para manufacturas de alimentos.

La cuestión de interés básica es si el consumo de azúcar ha variado con el paso del tiempo de forma distinta según el distinto uso que se le dé (como refinado o para manufacturas). Ver Figura 2.3.

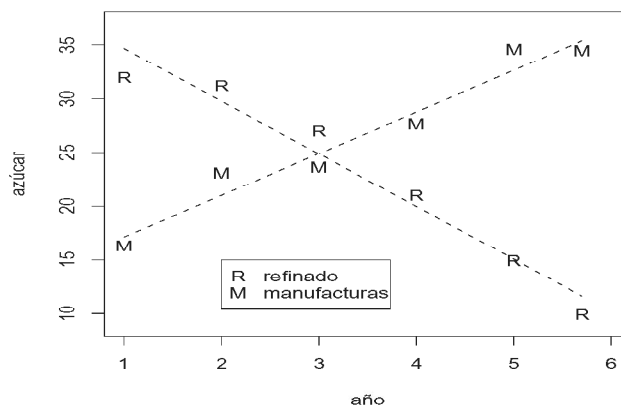


Figura 2.3: Datos del Consumo de Azúcar.

MODELIZACIÓN:

■ Variables:

– Respuesta $\rightsquigarrow Y_{ij} \equiv$ consumo de azúcar en la época j , con uso como $i = 0$ (refinado), 1 (manufacturas).

– Predictoras:

Factor $\rightsquigarrow$ uso, referente al tipo de consumo: $i = 0$ (refinado), 1 (manufacturas).

Covariable $\rightsquigarrow$ época t_{ij} : definimos la variable temporal empezando por 'época=1' y considerando la separación entre los períodos observados; los primeros 5 períodos de tiempo distan entre sí 10 años, pero el último período dista del anterior sólo 7 años, es decir,

$$epoca < -c(1, 2, 3, 4, 5, 5 + 7/10).$$

- Hipótesis distribucional:

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2).$$

- Modelos propuestos:

$$M1 : \mu_{ij} = \mu + \beta t_{ij}$$

$$M2 : \mu_{ij} = \mu + \alpha_i + \beta t_{ij}$$

$$M3 : \mu_{ij} = \mu + \alpha_i + \beta_i t_{ij}.$$

- Ajustes (en R):

```
M1 <- glm(consumo ~ epoca, family = gaussian)
M2 <- glm(consumo ~ epoca + uso, family = gaussian)
M3 <- glm(consumo ~ epoca * uso, family = gaussian).
```

- Comparación de modelos:

anova(M1, M2, M3, test = "F")						
	Resid. Df	Resid. Dev	Df	Deviance	F	Pr(> F)
epoca	10	698.55				
epoca + uso	9	651.35	1	47.20	12.49	0.007693
epoca * uso	8	30.25	1	621.10	164.28	1.296e-06

La comparación de modelos favorece la modelización M3, en la que se ajusta una recta para predecir el consumo de azúcar refinado y otra para el de uso en manufacturas, como claramente exigía la visualización de los datos en la Figura 2.3.

2.9. Ejercicios

1. Comprobar que las siguientes distribuciones de probabilidad pertenecen a la familia exponencial:

- a) Distribución de Pareto: $f(y; \theta) = \theta y^{-\theta-1}$.
- b) Distribución Exponencial: $f(y; \theta) = \theta \exp(-y\theta)$.
- c) Distribución Binomial Negativa:

$$f(y; \theta) = \binom{y+r-1}{r-1} \theta^r (1-\theta)^y.$$

- d) Distribución Bernoulli: $f(y; \pi) = \pi^y (1-\pi)^{1-y}$.

2. Considera una v.a. $Y_i \sim Ga(\nu, \lambda_i)$, con función de densidad

$$f(y_i; \nu, \lambda_i) = \frac{\lambda_i^\nu}{\Gamma(\nu)} y_i^{\nu-1} \exp(-\lambda_i y_i).$$

- a) Expresa la función de densidad de la forma

$$\exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}.$$

- b) Identifica el link canónico para datos gamma y justifica tu respuesta.
- c) Calcula $E(Y_i)$ y $Var(Y_i)$ utilizando la relación que existe entre ellas y $b(\theta)$. ¿Cuál es la función de varianza para datos gamma?
- d) Considera una m.a. $y_1, \dots, y_n$ de variables gamma, que relacionamos con un predictor X a través del link inverso:

$$\eta_i = \frac{1}{E(Y_i)} = \beta_0 + \beta_1 x_i.$$

Obtén las expresiones de los *scores* $U_j, j = 0, 1$ a partir de los que se calculan las estimaciones máximo-verosímiles:

$$U_j = \frac{\partial l(\theta, y)}{\partial \beta_j} = 0, \quad j = 0, 1.$$

3. Para las siguientes distribuciones: a) Verificar los resultados (2.4) y (2.5) sobre la relación entre $E(Y)$ y $Var(Y)$; b) identificar las funciones $b(\theta)$ y $a_i(\phi)$ de la descomposición en términos de la familia exponencial; c) identificar el link canónico.

- a) $Y \sim N(\mu, \sigma^2)$
 - b) $Y \sim Po(\mu)$
 - c) $Y \sim Bi(n, \pi)$
 - d) $Y \sim Ga(\nu, \nu/\mu)$.
4. En un modelo binomial $Y_i \sim Bi(n_i, \pi_i)$, $i = 1, \dots, n$, dado un predictor x y el ajuste del modelo $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$, obtén las expresiones de los scores U_j . Escribe a continuación la expresión de las ecuaciones a resolver para obtener los estimadores máximo-verosímiles $\hat{\beta}$.
5. Considera $Y_1, \dots, Y_n$ variables aleatorias independientes con

$$Y_i = \beta_0 + \log(\beta_1 + \beta_2 x_i) + \epsilon_i,$$

donde $\epsilon_i \sim N(0, \sigma^2)$, $\forall i$. ¿Se trata éste de un modelo lineal generalizado? Justifica tu respuesta.

6. Considera observaciones independientes Poisson $Y_i \sim Po(\mu_i)$ y el link identidad como predictor lineal,

$$g(\mu_i) = \mu_i = \eta_i = \beta_1 + \beta_2 x_i.$$

- a) Obtén la expresión de la función score y de la matriz de información de Fisher.
 - b) Obtén la expresión de la ecuación a resolver por el método Scoring de Fisher (2.14).
 - c) Calcula las matrices W y Z en (2.15) y obtén la expresión de (2.18) para este modelo.
 - d) Determina el punto de arranque $\hat{\beta}^{(0)}$ para inicial el algoritmo ajustando $\hat{\mu}_i = y_i$ para un par de observaciones i cualquiera.
7. Considera un modelo en el que la respuesta $Y_i \sim \text{Exp}(\theta_i)$ está relacionada con una variable explicativa x a través de:

$$E(Y_i) = \frac{1}{\theta_i} = \frac{\beta_1}{\exp\{\beta_2 x_i\}}.$$

Justifica si dicho modelo es un GLM, e identifica en caso afirmativo, la función link y el predictor lineal.

8. Considera las variables respuesta $Y_i \sim \text{Exp}(\theta_i)$, $i = 1, \dots, n$, y el modelo lineal generalizado:

$$E(Y_i) = \frac{1}{\theta_i} = \exp\{\beta_1 + \beta_2 x_i\},$$

- a) Expresa la función de densidad de probabilidad de Y_i de la forma

$$\exp\left\{\frac{y_i \theta_i - b_i(\theta)}{a_i(\phi)} + c(y_i, \phi)\right\},$$

e identifica cada uno de los elementos de dicha expresión.

- b) ¿Cuál es la función de varianza? Identifica el link canónico $g(\mu_i)$ y obtén la relación que tiene con la función de varianza $V(\mu_i)$.
- c) Obtén las expresiones de los scores U_j . Escribe a continuación la expresión de las ecuaciones a resolver para obtener los estimadores máximo-verosímiles $\hat{\beta}$.
9. Considera una v.a. $Y_i \sim \text{Po}(\lambda_i t)$, con función de densidad

$$f(y_i; \lambda_i) = \frac{1}{y_i!} (\lambda_i t)^{y_i} \exp(-\lambda_i t),$$

donde t es una constante conocida.

- a) Expresa la función de densidad de la forma

$$\exp\left\{\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}.$$

- b) Identifica el link canónico para datos Poisson y justifica tu respuesta.
- c) Calcula $E(Y_i)$ y $\text{Var}(Y_i)$ utilizando la relación que previamente has de derivar, entre ellas y $b(\theta)$. ¿Cuál es la función de varianza para datos Poisson?
- d) Considera una m.a. $y_1, \dots, y_n$ de variables Poisson, que relacionamos con un predictor x a través del link raíz cuadrada:

$$\eta_i = \sqrt{E(Y_i)} = \beta_0 + \beta_1 x_i.$$

Obtén las expresiones de los scores U_j , $j = 0, 1$ a partir de los que se calculan las estimaciones máximo-verosímiles:

$$U_j = \frac{\partial l(\theta, y)}{\partial \beta_j} = 0, \quad j = 0, 1.$$

10. La distribución exponencial se utiliza con frecuencia para describir tiempos de supervivencia.

- a) Utilizando la expresión general de la densidad de una distribución de la familia exponencial, prueba que la distribución exponencial es un caso particular de la distribución gamma.

Nota:

$$Y \sim \text{Exp}(\theta) \rightsquigarrow f(y) = \theta \exp(-y\theta)$$

$$Y \sim \text{Ga}(\alpha, \beta) \rightsquigarrow f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y).$$

- b) Si asumes que la variable respuesta $Y_i \sim \text{Exp}(\theta_i)$ está relacionada con un predictor lineal a través de:

$$E(Y_i) = \frac{1}{\theta_i} = \exp\{\beta_1 + \beta_2 x_i\},$$

identifica la función link y obtén las ecuaciones a resolver para obtener los estimadores máximo-verosímiles de β_1 y β_2 .

11. Considera observaciones independientes binomiales $Y_i \sim \text{Bi}(\pi_i, n_i)$, $i = 1, \dots, n$, y el link canónico como predictor lineal,

$$g(\mu_i) = \log \frac{n_i \pi_i}{n_i - n_i \pi_i} = \eta_i = \mathbf{x}_i' \beta.$$

Obtén la expresión de las ecuaciones a resolver por el método de mínimos cuadrados ponderados iterados (2.15).

12. Demuestra que cuando $g(\mu_i)$ es el link canónico, entonces $H = -\mathcal{I}$ y los métodos de Newton-Raphson y Scoring de Fisher coinciden. Verifica el resultado sobre las observaciones binomiales del Ejercicio anterior.
13. Comprueba que para variables normales independientes $Y_1, \dots, Y_n$, con $E(Y_i) = \mu_i$ y $\text{Var}(Y_i) = \sigma^2$, el estimador corregido de los grados de libertad para ϕ (en la ecuación (2.19)) viene dado por $\tilde{\phi} = \sigma^2$.
14. Comprueba que cuando tenemos una muestra de variables independientes y normales, $Y_1, \dots, Y_n \sim N(\mu_i, \sigma^2)$, la distribución del estadístico score en (2.21) no es asintótica sino *exacta*.

15. Obtén el estadístico score para cuando disponemos de una muestra Bernoulli $X_1, \dots, X_n \sim Br(\pi)$. Calcula su distribución asintótica utilizando el resultado (2.21). Utiliza la verosimilitud de la variable $Y = \sum X_i \sim Bi(n, \pi)$.
16. Demostrar que cuando las variables respuesta $Y_1, \dots, Y_n$ se distribuyen independientemente según

$$Y_i \sim N(\mathbf{x}_i' \beta, \sigma^2),$$

donde σ^2 es una constante conocida, entonces la distribución del estadístico de Wald en (2.22) es exacta.

17. Verificar que para datos Gamma, la deviance no escalada tiene la forma

$$D = 2 \sum_i \left[-\log \left(\frac{y_i}{\hat{\mu}_i} \right) + \left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right) \right].$$

18. Se llevó a cabo un experimento para calibrar un viscosímetro Stormer. Este instrumento mide la viscosidad de un líquido considerando el tiempo (T) que tarda su cilindro interno en completar cierto número de revoluciones. Es posible también, variar el peso que controla el cilindro. Según los expertos, el tiempo medio (μ) depende de la viscosidad (x_1) y del peso (x_2) de la siguiente forma

$$\mu_i = \frac{\alpha_1 x_{1i}}{x_{2i} - \alpha_2}.$$

Se dispone de un conjunto de observaciones independientes del tiempo T_i , $t_1, \dots, t_n$, a diferentes condiciones de peso y viscosidad. No se sabe nada sobre la distribución de la respuesta T_i .

- ¿Cuáles son las hipótesis básicas (aparte de independencia entre las respuestas) de un Modelo Lineal Generalizado?
 - De acuerdo con lo anterior, ¿se trata éste de un modelo lineal generalizado? Justifica tu respuesta.
 - Asumiendo lo que toca sobre la distribución de la respuesta, ¿cómo puedes convertirlo en un GLM? Obtén una transformación de la media y unas variables explicativas (construidas a partir de x_1 y x_2) entre las que exista una relación lineal. Identifica el link.
19. En la Tabla 2.5 (Dobson, 1990, pag.82) tienes los datos correspondientes a la producción total de hierba y legumbres conjuntamente en función de la cantidad de fósforo (en kg/Ha) administrado a la tierra de cultivo.
- ¿Cuál es la relación entre la producción y la cantidad de fertilizante?

fósforo	producción	fósforo	producción	fósforo	producción
0	1753.9	5	3184.6	15	4184.6
15	3107.7	5	3046.2	40	4461.5
10	2400.0	30	4892.3	0	2723.1
40	4923.1	10	3538.5	40	4692.3
30	4415.4	0	2553.8	20	4215.4
5	2861.6	40	4784.6	50	4784.6
50	5246.2	30	4000.0	20	3600.0
50	4938.4	10	3323.1	40	4153.9
40	3723.0	20	3184.6	15	3169.3

Tabla 2.5: Producción de hierbas y legumbres versus la cantidad de fosfatos incluidos en el abono

20. Los datos de la Tabla 2.6 (Dobson, 1990, pag.69) corresponden a un experimento que se llevó a cabo para investigar el efecto de una dieta alta en carbohidratos en diabéticos dependientes de insulina. Este experimento se llevó a cabo durante 6 meses en los cuales 20 pacientes varones diabéticos e insulino-dependientes habían de seguir un régimen impuesto por el médico. La obediencia de los pacientes a la hora de respetar el régimen (algo que podemos medir de alguna forma con el porcentaje de calorías totales en cada paciente que sólo se pueden obtener de carbohidratos complejos: variable *carbohydrate*) se piensa que puede estar relacionada con la edad (en años), el peso (relativo a peso ideal para la estatura del individuo), y otras componentes de la dieta como el porcentaje de calorías proteínicas.

¿Cómo explican las variables *edad*, *peso* y *proteínas* el porcentaje de calorías por hidratos de carbono?

21. Es bien conocido que la concentración de colesterol en la sangre aumenta con la edad. Sin embargo, está menos claro que el nivel de colesterol esté asociado también con el peso corporal. En la Tabla 2.7 (Dobson, 1990, pag.83) tienes los datos, para 30 mujeres, de su concentración de colesterol (en milimoles por litro), edad (en años) y un índice de masa corporal (peso dividido por altura al cuadrado, con el peso medido en kg. y la altura en metros).

El colesterol, ¿está asociado con el índice de masa corporal cuando la edad también se incluye en el modelo?

22. La Tabla 2.8 presenta los datos de *cholestyramine* (Barnhart y Sampson, 1995,

Carbohidrato (y)	Edad (x_1)	Peso (x_2)	Proteína (x_3)
33	33	100	14
40	47	92	15
37	49	135	18
27	35	144	12
30	46	140	15
43	52	101	15
34	62	95	14
48	23	101	17
30	32	98	15
38	42	105	14
50	31	108	17
51	61	85	19
30	63	130	19
36	40	127	20
41	50	109	15
42	64	107	16
46	56	117	18
24	61	100	13
35	48	lis	18
37	28	102	14

Tabla 2.6: Obediencia a un régimen alto en carbohidratos de 20 diabéticos varones dependientes de la insulina

también en Lindsey, 1997, pag.67). Ciento sesenta y cuatro hombres fueron sometidos a una prueba en la que se comprometían a tomar al día 6 paquetes de un medicamento llamado *cholostyramine*, al cual se le atribuía la capacidad de reducir el nivel de colesterol. A pesar del compromiso, se temía que realmente algunos de estos hombres tomaran al día mucho menos de la cantidad prescrita de medicamento. La variable z mide la *obediencia* como el porcentaje de la dosis propuesta que realmente se tomaba. Esto se medía contando el número de paquetes no consumidos que cada hombre devolvía. La variable y representa la *mejora*: el decrecimiento total del nivel de colesterol en el plasma sanguíneo desde el comienzo hasta el fin del experimento.

El medicamento, ¿realmente hacía disminuir el nivel de colesterol?

23. Considera los datos que hemos representado en la Figura 2.4 (Powsner, 1935, también en McCullagh y Nelder, 1989, pag.307). En el estudio en el que se recogieron dichos datos se pretendía relacionar la duración promedio del período embrionario de la mosca de la fruta (*Drosophila Melanogaster*) con la temperatura. En cada prueba se sometía a cierto número de huevos (n) a una tem-

Colesterol	Edad	Masa Corporal	Colesterol	Edad	Masa Corporal
5.94	52	20.7	6.48	65	26.3
4.71	46	21.3	8.83	76	22.7
5.86	51	25.4	5.10	47	21.5
6.52	44	22.7	5.81	43	20.7
6.80	70	23.9	4.65	30	18.9
5.23	33	24.3	6.82	58	23.9
4.97	21	22.2	6.28	78	24.3
8.78	63	26.2	5.15	49	23.8
5.13	56	23.3	2.92	36	19.6
6.74	54	29.2	9.27	67	24.3
5.95	44	22.7	5.57	42	22.0
5.83	71	21.9	4.92	29	22.5
5.74	39	22.4	6.72	33	24.1
4.92	58	20.2	5.57	42	22.7
6.69	58	24.4	6.25	66	27.3

Tabla 2.7: Efecto del peso en el nivel de colesterol

peratura constante (T) durante 20-30 minutos, y se anotaba la duración media del período embrionario (en el grupo), $m.duration$, y su desviación estándar, $sd.duration$.

Identifica los elementos que se piden a continuación, justificando tu respuesta con los gráficos de la Figura 2.4 (1a) duración promedio versus desviación estándar, 1b) histograma de la duración promedio, 1c) duración promedio versus temperatura, 1d) logaritmo de la duración promedio versus la inversa de la temperatura):

- la variable respuesta y la variable explicativa;
- una distribución (de la familia exponencial) que te parezca aceptable para modelizar la variable respuesta;
- un link razonable para ligar la variable respuesta con el predictor lineal, especificando también la forma de dicho predictor lineal;
- una función de varianza aproximada.

z	y	z	y	z	y	z	y
0	-5.25	27	-1.50	71	59.50	95	32.50
0	-7.25	28	23.50	71	14.75	95	70.75
0	-6.25	29	33.00	72	63.00	95	18.25
0	11.50	31	4.25	72	0.00	95	76.00
2	21.00	32	18.75	73	42.00	95	75.75
2	-23.00	32	8.50	74	41.25	95	78.75
2	5.75	33	3.25	75	36.25	95	54.75
3	3.25	33	27.75	76	66.50	95	77.00
3	8.75	34	30.75	77	61.75	96	68.00
4	8.25	34	-1.50	77	14.00	96	73.00
4	-10.25	34	1.00	78	36.00	96	28.75
7	-10.50	34	7.75	78	39.50	96	26.75
8	19.75	35	-15.75	81	1.00	96	56.00
8	-0.50	36	33.50	82	53.50	96	47.50
8	29.25	36	36.25	84	46.50	96	30.25
8	36.25	37	5.50	85	51.00	96	21.00
9	10.75	38	25.50	85	39.00	97	79.00
9	19.50	41	20.25	87	-0.25	97	69.00
9	17.25	43	33.25	87	1.00	97	80.00
10	3.50	45	56.75	87	46.75	97	86.00
10	11.25	45	4.25	87	11.50	98	54.75
11	-13.00	47	32.50	87	2.75	98	26.75
12	24.00	50	54.50	88	48.75	98	80.00
13	2.50	50	-4.25	89	56.75	98	42.25
15	3.00	51	42.75	90	29.25	98	6.00
15	5.50	51	62.75	90	72.50	98	104.75
16	21.25	52	64.25	91	41.75	98	94.25
16	29.75	53	30.25	92	48.50	98	41.25
17	7.50	54	14.75	92	61.25	98	40.25
18	-16.50	54	47.25	92	29.50	99	51.50
20	4.50	56	18.00	92	59.75	99	82.75
20	39.00	57	13.75	93	71.00	99	85.00
21	-5.75	57	48.75	93	37.75	99	70.00
21	-21.00	58	43.00	93	41.00	100	92.00
21	0.25	60	27.75	93	9.75	100	73.75
22	-10.25	62	44.50	93	53.75	100	54.00
24	-0.50	64	22.50	94	62.50	100	69.50
25	-19.00	64	-14.50	94	39.00	100	101.50
25	15.75	64	-20.75	94	3.25	100	68.00
26	6.00	67	46.25	94	60.00	100	44.75
27	10.50	68	39.50	95	113.25	100	86.75

Tabla 2.8: Efectividad de la Cholestyramine para reducir el nivel de colesterol. Obediencia de los pacientes a las prescripción de medicamento

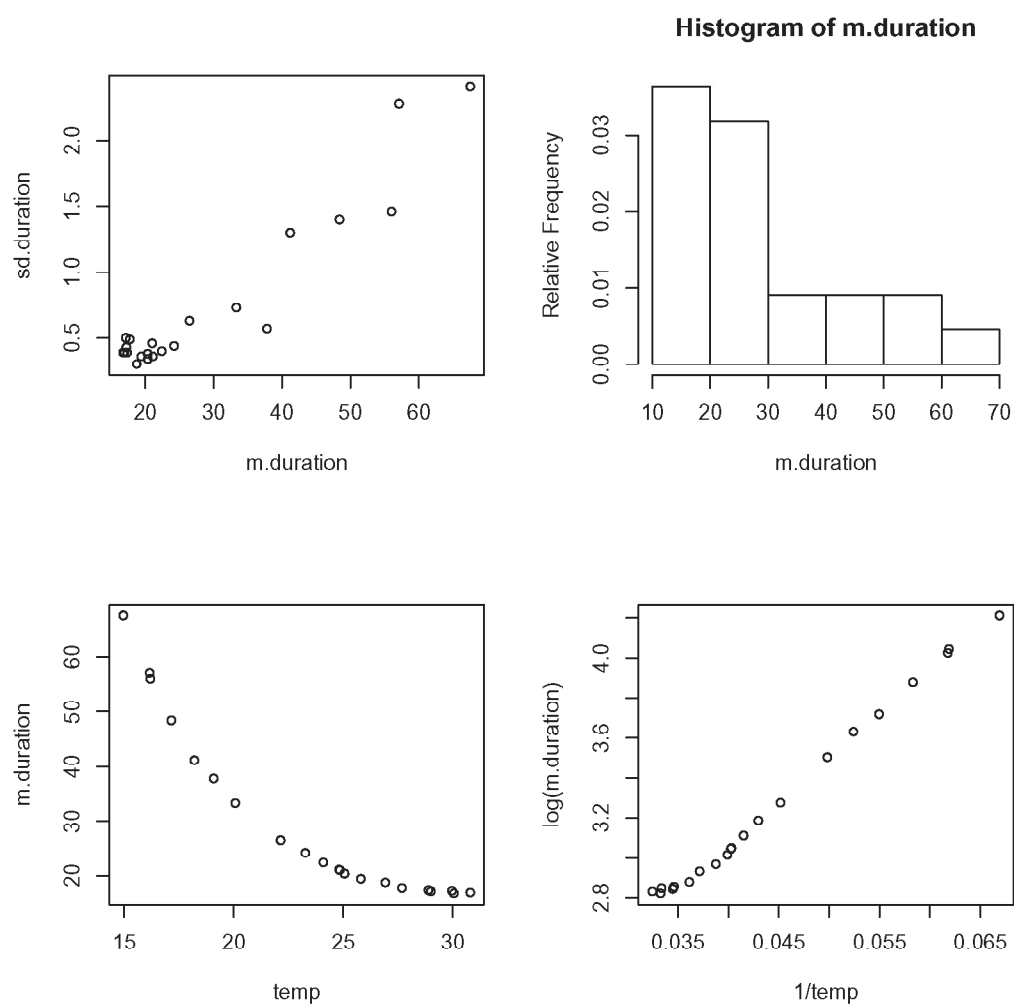


Figura 2.4: Duración promedio del período embrional versus temperatura.

3

GLM para Respuesta Binaria

3.1. Introducción. Ejemplos

Consideramos en este capítulo la modelización de respuestas binarias: para cada individuo, la variable respuesta puede tomar únicamente dos posibles valores, que denotaremos por 0 (fracaso) y 1 (éxito),

$$Pr(Z_i = 0) = 1 - \pi_i, \quad Pr(Z_i = 1) = \pi_i.$$

Si además se han observado una serie de covariables continuas o de tipo factor, el objetivo del análisis con el modelo lineal generalizado será predecir la probabilidad asociada al éxito (o equivalentemente al fracaso), en función de dichas covariables.

Ejemplo 3.1 (KIFOSIS) *En la Tabla 3.1 aparecen los datos de 81 niños que fueron intervenidos quirúrgicamente para corregirles problemas en la columna vertebral (en SPlus, Chambers y Hastie, 1993, pag.200). La variable binaria “kifosis” indica la presencia o ausencia de una deformidad postoperatoria en la columna, denominada kifosis. Las otras tres variables son ‘E’, edad del niño en meses, ‘Nro’, número*

de vértebras intervenidas en la operación, e 'Ini', que define la primera vértebra involucrada en la operación. Los datos están representados en la Figura 3.1.

Es de interés en el análisis investigar cómo están relacionadas variables como la edad del niño, el número de vértebras dañadas y el número de vértebra en la que empieza la deformidad a la hora de predecir la incidencia de la kifosis en el postoperatorio.

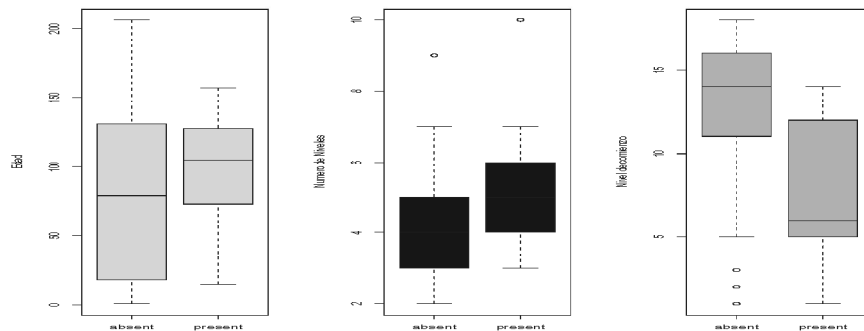


Figura 3.1: Datos de Kifosis

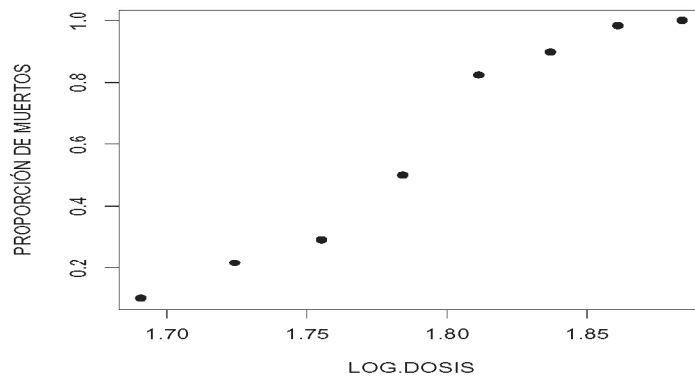


Figura 3.2: Datos de mortalidad de cucarachas.

kifosis	E	Nro	Ini	kifosis	E	Nro	Ini
no	71	3	5	no	35	3	13
no	158	3	14	no	143	9	3
si	128	4	5	no	61	4	1
no	2	5	1	no	97	3	16
no	1	4	15	si	139	3	10
no	1	2	16	no	136	4	15
no	61	2	17	no	131	5	13
no	37	3	16	si	121	3	3
no	113	2	16	no	177	2	14
si	59	6	12	no	68	5	10
si	82	5	14	no	9	2	17
no	148	3	16	si	139	10	6
no	18	5	2	no	2	2	17
no	1	4	12	no	140	4	15
no	168	3	18	no	72	5	15
no	1	3	16	no	2	3	13
no	78	6	15	si	120	5	8
no	175	5	13	no	51	7	9
no	80	5	16	no	102	3	13
no	27	4	9	si	130	4	1
no	22	2	16	si	114	7	8
si	105	6	5	no	81	4	1
si	96	3	12	no	118	3	16
no	131	2	3	no	118	4	16
si	15	7	2	no	17	4	10
no	9	5	13	no	195	2	17
no	8	3	6	no	159	4	13
no	100	3	14	no	18	4	11
no	4	3	16	no	15	5	16
no	151	2	16	no	158	5	14
no	31	3	16	no	127	4	12
no	125	2	11	no	87	4	16
no	130	5	13	no	206	4	10
no	112	3	16	no	11	3	15
no	140	5	11	no	178	4	15
no	93	3	16	si	157	3	13
no	1	3	9	no	26	7	13
si	52	5	6	no	120	2	13
no	20	6	9	si	42	7	6
si	91	5	12	no	36	4	13
si	73	5	1				

Tabla 3.1: Datos de Kifosis

Ejemplo 3.2 (CUCARACHAS) En un experimento se sometió a cierto número de cucarachas a cinco horas de exposición a disulfato de carbono caseoso a varias concentraciones (Bliss, 1935, también en Dobson, 1995, pag.109). Se pretendía investigar la relación existente entre la dosis de disulfato administrada y la resistencia de los insectos; si existe tal relación, determinar la dosis a la cual es posible garantizar el exterminio del 50% de los insectos. Los datos aparecen en la Tabla 3.2, y representados en la Figura 3.2.

log.dosis (x_i)	No. insectos (n_i)	No. muertos (y_i)
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Tabla 3.2: Datos de mortalidad de cucarachas

Ejemplo 3.3 (POLILLAS) Collet (1991) presenta un experimento sobre la toxicidad de distintas dosis (en microgramos) del piretroide trans-cipemetrín en los capullos del gusano del tabaco (*Heliothis virescens*). Se había comenzado a detectar resistencia de estas polillas a dicho tóxico (a determinadas dosis). El experimento consistía en exponer, durante tres días y a distintas dosis de tóxica, a series de 20 polillas de cada sexo (M=macho, H=hembra). Se anotó el número de polillas muertas en cada serie. Los resultados aparecen en la Tabla 3.3, y representados en la Figura 3.3.

Es de interés en el análisis investigar si efectivamente se demostraba tal resistencia al tóxico (en machos y hembras) y a partir de qué dosis.

sexo	dosis					
	1	2	4	8	16	32
M	1	4	9	13	18	20
H	0	2	6	10	12	16

Tabla 3.3: Datos sobre toxicidad con polillas del tabaco.

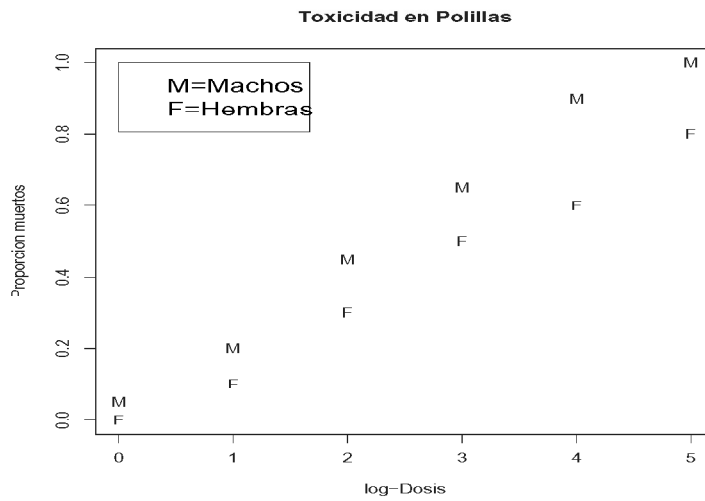


Figura 3.3: Datos sobre toxicidad con polillas del tabaco.

3.2. Hipótesis y Links para variables binarias

Cuando trabajemos con respuesta binaria, los datos observados vendrán, bien individualizados (con los valores $Z = 0/1$ como en el Ejemplo de Kifosis), bien resumidos en tablas de contingencia, a través de variables binomiales, $Y = \sum_i Z_i$ (como en los Ejemplos de Cucarachas (3.2) y Polillas (3.3)). En general, podemos plantear un problema con respuesta binaria representando los datos observados, $Y_1, Y_2, \dots, Y_N$ a través de una tabla de la forma:

	Pruebas			
éxitos	Y_1	Y_2	$\dots$	Y_N
fracasos	$n_1 - Y_1$	$n_2 - Y_2$	$\dots$	$n_N - Y_N$
totales	n_1	n_2	$\dots$	n_N

Cada prueba vendrá determinada por unas determinadas condiciones de experimentación (variables predictoras: continuas y de clasificación) que servirán para modelizar y predecir la probabilidad de éxito π_i asociada.

Las hipótesis a verificar a la hora de modelizar respuestas binarias con modelos lineales generalizados son:

- independencia entre las observaciones (entre los individuos si las respuestas son 0/1 o entre las pruebas binomiales)
- linealidad entre transformaciones de la proporción de éxitos y de las variables explicativas continuas
- consistencia entre la modelización y la interpretación física.

Si π_i es la probabilidad de éxito asociado al i -ésimo experimento binomial con respuesta Y_i , y condiciones de experimentación observadas en las variables explicativas $X_1, X_2, \dots, X_p$, la formulación de un GLM vendrá descrita en términos de la función link:

$$g(\pi_i) = \eta_i = \mathbf{x}_i' \beta.$$

El caso más simple a considerar sería el **Modelo Lineal con link identidad**:

$$\pi_i = \mathbf{x}_i' \beta. \quad (3.1)$$

Si bien este modelo se utiliza en algunas aplicaciones prácticas, tiene la desventaja de que, mientras que π_i está restringido al dominio $[0, 1]$, el predictor lineal $\mathbf{x}_i' \hat{\beta}$ podría caer fuera del intervalo probabilístico, invalidando la predicción para determinados valores de las variables explicativas. En estos modelos, la extrapolación de predicciones es, cuando menos, peligrosa.

Para asegurar que la predicción de π_i quede dentro del intervalo $[0, 1]$, es razonable considerar funciones link que provengan de funciones de distribución de probabilidad, de la forma

$$\pi_i = g^{-1}(\mathbf{x}_i' \beta) = \int_{-\infty}^t f(s) ds,$$

donde $f(s) \geq 0$ y $\int_{-\infty}^{\infty} f(s) ds = 1$ define una distribución de probabilidad. A la función de densidad de probabilidad $f(s)$ se la denomina en este ámbito **función de tolerancia**.

Los links más habituales en modelos lineales generalizados con respuesta binaria, son:

$$\begin{aligned} \text{logit}(\pi_i) &= \log(\pi_i/(1 - \pi_i)) \\ \text{probit}(\pi_i) &= \Phi^{-1}(\pi_i) \\ \text{cloglog}(\pi_i) &= \log(-\log(1 - \pi_i)) \\ \text{loglog}(\pi_i) &= -\log(-\log(\pi_i)). \end{aligned}$$

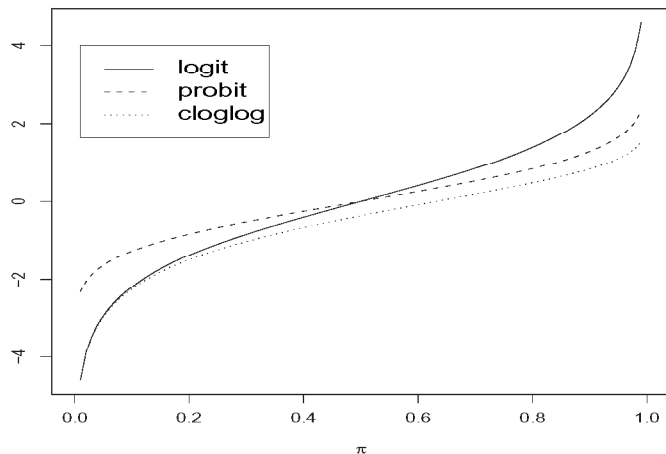


Figura 3.4: Links usuales para respuesta binaria

Las diferencias básicas entre estos links con forma de “S” invertida, residen en el comportamiento en las colas, esto es, para valores de π próximos a 0 o 1, tal y como puede apreciarse en la Figura 3.4.

3.3. Modelos de dosis–respuesta

Un tipo de modelos muy frecuentes para datos binarios son los de dosis–respuesta. Se trabaja probando sobre grupos de individuos diversas dosis de fármacos (o tóxicos), y se pretende inferir sobre su eficacia utilizando el número de individuos en un total fijo, que experimentan alguna mejoría (o deterioro). Los Ejemplos de Cucarachas (3.2) y Polillas (3.3) admiten una modelización de este tipo.

Consideremos x la dosis de fármaco administrada en cierta prueba binomial en la que Y individuos (de n tratados) experimentan mejoría. Buscamos una función link g para predecir la probabilidad de mejora π a dosis x ,

$$g(\pi) = \beta_1 + \beta_2 x.$$

3.3.1. Modelo Lineal

La función de tolerancia es la uniforme: se modeliza la “probabilidad de efectividad” π a dosis x , razonando con que existe un nivel mínimo de dosis c_1 a la cual el fármaco no tiene ningún efecto; a partir de dicho nivel, más dosis implica mayor efectividad, con un incremento uniforme; y existe un nivel máximo o de saturación c_2 , a partir del cual el fármaco ha llegado a su efectividad máxima y ya no produce ninguna mejora adicional (ver Figura 3.5).

Es decir,

$$f(s) = \begin{cases} \frac{1}{c_2 - c_1}, & \text{si } s \in [c_1, c_2] \\ 0 & \text{en otro caso.} \end{cases}$$

La función de distribución uniforme, representada en la Figura 3.5, da lugar al link identidad,

$$\pi = \int_{-\infty}^x f(s)ds = \frac{x - c_1}{c_2 - c_1} = \beta_1 + \beta_2 x.$$

3.3.2. Modelo Probit

Se asume que la efectividad del fármaco se va dejando notar desde los primeros momentos (dosis bajas) y se incrementa de modo continuo a medida que se aumenta la dosis x de fármaco, siguiendo la forma de una función de distribución normal (ver Figura 3.5). A lo largo de este incremento de efectividad con la dosis, existe un nivel de dosis medio μ que proporciona un 50 % de garantías de mejoría (efectividad). Dicho nivel se interpreta en estudios de toxicología, como la dosis letal media ($LD50$), y se entiende como la dosis requerida para lograr una tasa de mortalidad del 50 %.

Así,

$$\begin{aligned} f(s) = N(s|\mu, \sigma^2) &\rightsquigarrow \pi = \int_{-\infty}^x f(s)ds = \Phi\left(\frac{x - \mu}{\sigma}\right) \\ &\rightsquigarrow g(\pi) = \Phi^{-1}(\beta_1 + \beta_2 x) \\ &\text{con } \beta_1 = -\mu/\sigma \text{ y } \beta_2 = 1/\sigma. \end{aligned} \tag{3.2}$$

3.3.3. Modelo Logit

La función de tolerancia para este modelo tiene la forma

$$f(s) = \frac{\beta_2 \exp\{\beta_1 + \beta_2 s\}}{[1 + \exp\{\beta_1 + \beta_2 s\}]^2},$$

que relaciona la dosis con la proporción de éxito según una curva de la forma descrita en la Figura 3.5 y que da lugar al link logit:

$$\pi = \frac{\exp\{\beta_1 + \beta_2 s\}}{1 + \exp\{\beta_1 + \beta_2 s\}} \rightsquigarrow g(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \beta_1 + \beta_2 x. \quad (3.3)$$

3.3.4. Modelo Complementario Log-Log

La relación entre la dosis y la proporción de individuos que experimentan mejora viene dada también por una curva en forma de “S” invertida (similar a la de los modelos probit y logit), tal y como aparece en la Figura 3.5. La función de tolerancia es la función de densidad de la distribución del valor extremo:

$$f(s) = \beta_2 \exp\{(\beta_1 + \beta_2 s) - \exp(\beta_1 + \beta_2 s)\}.$$

El link cloglog surge de dicha función:

$$\begin{aligned} \pi &= \int_{-\infty}^x f(s) ds = 1 - \exp\{-\exp(\beta_1 + \beta_2 s)\} \\ g(\pi) &= \text{cloglog}(\pi) = \log(-\log(1 - \pi)) = \beta_1 + \beta_2 x. \end{aligned} \quad (3.4)$$

3.4. Estimación y Bondad de Ajuste

Supongamos $Y_1, Y_2, \dots, Y_N$ el número de éxitos obtenidos en N experimentos binomiales de tamaños $n_1, n_2, \dots, n_N$, con probabilidades asociadas $\pi_1, \pi_2, \dots, \pi_N$, y condiciones de experimentación definidas en p variables explicativas $X_1, X_2, \dots, X_p$.

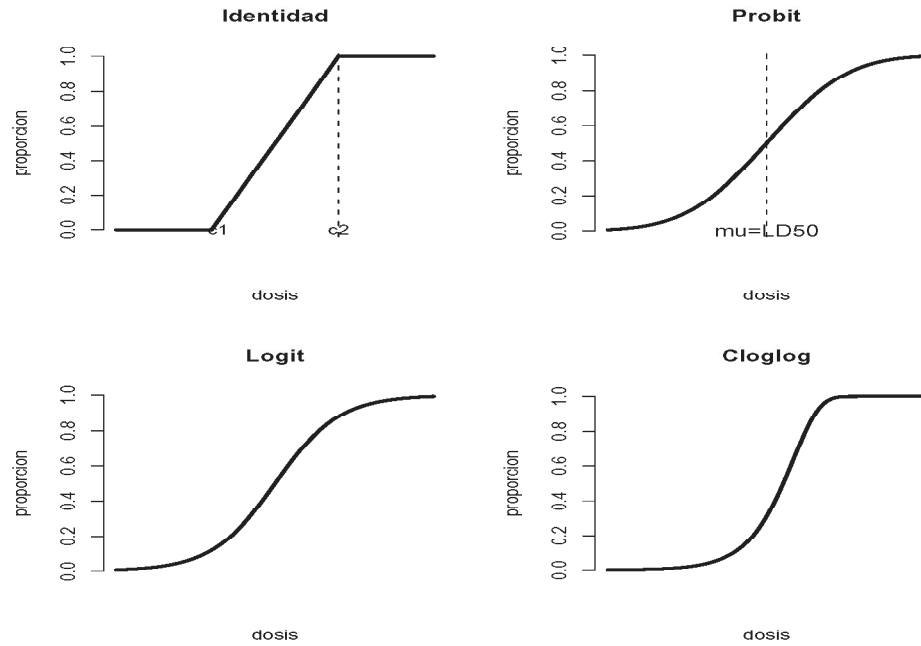


Figura 3.5: Funciones de tolerancia en Modelos de Dosis-Respuesta.

La función de log-verosimilitud conjunta viene dada por:

$$l(\pi; \mathbf{y}) = \sum_{i=1}^N y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i) + \log \binom{n_i}{y_i}.$$

Una vez especificada la forma del predictor lineal y elegido el link, $g(\pi_i) = \mathbf{x}_i' \beta$, las estimaciones de los parámetros β se obtendrán derivando la log-verosimilitud respecto de β e igualándola a 0.

Por otro lado, para el modelo saturado, las EMV de los π_i vienen dadas por las

proporciones observadas, $\hat{\pi}_{i(sat)} = y_i/n_i$, de donde la deviance resulta

$$\begin{aligned} S &= -2[l(\hat{\pi}; \mathbf{y}) - l(\hat{\pi}_{(sat)}; \mathbf{y})] \\ &= 2 \left\{ \sum_{i=1}^N y_i \log \left(\frac{y_i}{n_i \hat{\pi}_i} \right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i (1 - \hat{\pi}_i)} \right) \right\} \\ &= 2 \sum_{i=1}^{C=2N} O_i \log \left(\frac{O_i}{E_i} \right), \end{aligned} \quad (3.5)$$

donde O_i representa la frecuencia observada y E_i la frecuencia esperada en cada una de las $C = 2N$ celdas de la tabla de contingencia construida con los éxitos y los fracasos.

Puesto que en el modelo binomial el parámetro de dispersión es conocido, $\phi = 1$, la bondad del ajuste se podría contrastar directamente utilizando la aproximación asintótica a la distribución de la deviance,

$$S \sim \chi_{N-p}^2.$$

Sin embargo, existen ciertas restricciones para la utilización de S como criterio de bondad de ajuste, y son:

1. las observaciones han de ser independientes y sin sobredispersión, hipótesis que con frecuencia no se verifican en los datos;
2. si $\dim(\mathbf{y}) = N$ es fija, la aproximación asintótica es válida cuando el tamaño de cada experimento binomial es grande, $n_i \rightarrow \infty$, y por tanto $n_i \pi_i (1 - \pi_i) \rightarrow \infty$. Si siendo N grande, $n_i \pi_i (1 - \pi_i)$ permanece acotado, entonces: (i) la aproximación a una χ^2 no se dará; y (ii) S no será independiente de $\hat{\pi}$ (como debería ser), ni siquiera aproximadamente.

Conviene pues considerar la deviance como una medida no-absoluta de bondad de ajuste. Es preferible utilizarla para comparar modelos anidados, y elegir el mejor modelo a través de la inclusión/exclusión de términos. La aproximación asintótica a una χ^2 de la diferencia de deviances es, generalmente, más precisa que la de la deviance misma.

En general, es conveniente utilizar criterios de bondad de ajuste derivados de la comparación de modelos.

Otro criterio de bondad de ajuste para GLM's con respuesta binaria pasa por el estadístico χ^2 de Pearson, definido como

$$\chi^2 = \sum_{i=1}^{C=2N} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^N \frac{(y_i - n_i \hat{\pi}_i)^2}{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}. \quad (3.6)$$

El estadístico χ^2 es asintóticamente equivalente al del cociente de verosimilitudes (deviance), esto es,

$$S \sim \chi_{N-p}^2 \Rightarrow \chi^2 \sim \chi_{N-p}^2.$$

Algunos estudiosos comentan que el estadístico χ^2 parece funcionar mejor que la deviance como criterio de bondad de ajuste, debido a que la deviance S se ve indebidamente muy influenciada por frecuencias muy pequeñas en las celdas. Aun así, cualquier aproximación asintótica de la distribución de estos estadísticos de bondad de ajuste es pobre si las frecuencias esperadas son demasiado pequeñas.

Otra aproximación para evaluar la bondad del ajuste está basada en los errores de predicción. Si la predicción era el objetivo principal del análisis, la proporción de observaciones clasificadas correctamente sería un buen criterio para la comparación de modelos. Por ejemplo, podemos predecir éxito del ajuste en una prueba si la probabilidad estimada supera 0,5 y fracaso en otro caso; en ese caso cabe tabular las respuestas observadas y predichas, y calcular la proporción de casos predichos correctamente.

3.5. Residuos

El examen gráfico de los residuos es útil para evaluar la adecuación de un modelo propuesto para ajustar unos datos.

Los residuos de Pearson toman la forma

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(n_i - \hat{\mu}_i)/n_i}}, \quad (3.7)$$

donde $\hat{\mu}_i = \hat{\pi}_i n_i$ es el número esperado de éxitos para la prueba i -ésima. Recordemos que $\chi^2 = \sum_i (r_i^P)^2$.

Cuando los y_i corresponden a conteos de pruebas binomiales (datos agrupados), los residuos de Pearson se distribuyen aproximadamente según una normal con media 0 y desviación típica de 1. Cuando los datos vienen individualizados (los y_i 's son 0 y 1), dicha aproximación no es cierta. Con todo, ni siquiera con datos agrupados la distribución de los residuos ha de parecerse demasiado a una normal. Sin embargo, en ambos casos las observaciones con un residuo de Pearson mayor a 2 en valor absoluto, deberían ser miradas con mayor atención.

Cuando se representan estos residuos versus los niveles de factores y covariables continuas, no debería aparecer ningún patrón sistemático.

Los **residuos deviance** se definen como las raíces cuadradas de las contribuciones de las observaciones a la deviance,

$$r_i^D = \text{sgn}(y_i - \hat{\mu}_i) \sqrt{2 \left[y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - \hat{\mu}_i} \right) \right]}. \quad (3.8)$$

Pierce y Schafer (1986) probaron que los residuos deviance se distribuyen aproximadamente como normales si el modelo ajustado es bueno, y por lo tanto proporcionan residuos apropiados para propósitos de diagnóstico.

La deviance se obtiene con $S = \sum_i (r_i^D)^2$. Observaciones con un residuo deviance superior a 2 puede indicar falta de ajuste.

Las correspondientes **versiones estudentizadas** de los residuos se obtienen a partir de la matriz gorro H , dada en (2.39), con $W = \text{diag}\{(\mu_i/n_i)(n_i - \mu_i) g'(\mu_i)^2\}$, evaluada en el $EMV(\mu_i)$. Si el modelo ajustado es un modelo logit, se tiene $W^{-1} = \text{diag}\{\mu_i(n_i - \mu_i)/n_i\}$. Utilizando esta expresión, se puede probar que la varianza del residuo fila es, según una aproximación de primer orden,

$$\text{Var}(Y_i - \hat{\mu}_i) \approx (1 - h_{ii}) \text{Var}(Y_i),$$

donde h_{ii} es el valor leverage de la diagonal de H . Así, un residuo estandarizado se puede obtener dividiendo el residuo por la raíz cuadrada de $1 - h_{ii}$, como ya comentamos en la Sección 2.7, esto es,

$$r_i^{PS} = \frac{r_i^P}{\sqrt{1 - h_{ii}}} = \frac{y_i - \hat{\mu}_i}{\sqrt{(1 - h_{ii})\hat{\mu}_i(n_i - \hat{\mu}_i)/n_i}}, \quad (3.9)$$

y análogamente para los residuos deviance.

Las **distancias de Cook** permitirían identificar observaciones que tienen más influencia en el modelo ajustado. Pregibon (1981) demostró que las medidas

$$DC_i = (r_i^{PS})^2 \frac{h_{ii}}{(1 - h_{ii})p}, \quad (3.10)$$

con p el número de parámetros estimados en el modelo, pueden utilizarse como una aproximación de un paso a las distancias de Cook.

3.6. Modelización y Análisis. Ejemplos

Ejemplo 3.4 (Continuación del Ejemplo 3.1: Respuesta 0/1) Recordemos que el objetivo era investigar si la kifosis estaba relacionada con la edad de los niños (en meses), el número de vértebras afectadas y la posición de la primera vértebra afectada.

Identificación de variables

- *Respuesta:*

Y = presencia de kifosis, $= 0 \rightsquigarrow$ no hay kifosis, $= 1 \rightsquigarrow$ sí la hay.

- *Explicativas continuas:*

E = edad (en meses)

Nro = número de vértebras afectadas

Ini = vértebra de inicio de la afección.

Hipótesis distribucional y modelización

La variable respuesta representa si el individuo tiene o no kifosis. Denotando por π_i = probabilidad de que el i -ésimo individuo tenga kifosis, podemos asumir una distribución Bernoulli para Y_i . Si además optamos por el link logit, proponemos:

$$Y_i \sim Br(\pi_i)$$

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 E_i + \beta_2 Nro_i + \beta_3 Ini_i, \quad i = 1, \dots, 81.$$

Análisis

Para obtener el ajuste en $\mathcal{R}$, la variable respuesta ha de ser un vector de ceros (fracasos=ausencia de kifosis) y unos (éxitos=presencia de kifosis),

$$kifo.unos <- 1 * (kifosis == "si")$$

o directamente, utilizar como respuesta la variable '(kifosis == "si")'. Si usamos sin más el factor 'kifosis', hay que tener en cuenta que $\mathcal{R}$ codifica los niveles en orden alfabético, de forma que 'no-1' y 'si-2'. $\mathcal{R}$ identifica después el nivel 1 como fracaso y el nivel 2 como éxito. Así pues, en este caso no es necesaria la recodificación y es posible utilizar directamente el factor 'kifosis', con sus dos niveles de respuesta 'si' y 'no'. Se obtiene el ajuste logístico con:

$$fit <- glm(kifosis \sim E + Nro + Ini, binomial)$$

Las estimaciones que proporciona son

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.036708	1.443836	-1.411	0.15836
E	0.010929	0.006416	1.703	0.08849
Nro	0.410564	0.223772	1.835	0.06654
Ini	-0.206501	0.067498	-3.059	0.00222

Puestos a comprobar si todas las variables en el modelo son significativas:

1. Si atendemos a los contrastes basados en el estadístico de Wald (tabla anterior), tenemos que las variables edad (E) y número de vértebras (Nro) no resultan significativas al 95 %, aunque sí al 90 %.
2. Si atendemos a los contrastes basados en las deviances, comparando el ajuste actual 'fit' con todos los modelos que surgen excluyendo una a una cada una de las variables explicativas, tenemos:

<i>drop1(fit, test = "Chisq")</i>					
	<i>Df</i>	<i>Deviance</i>	<i>AIC</i>	<i>LRT</i>	<i>Pr(Chi)</i>
<i>none</i>		61.380	69.380		
<i>E</i>	1	64.536	70.536	3.157	0.075623
<i>Nro</i>	1	65.299	71.299	3.919	0.047739
<i>Ini</i>	1	71.627	77.627	10.247	0.001369

Así, la única variable no significativa al 95%, aunque sí al 90%, desde el punto de vista de las deviancias, es la variable edad (*E*).

3. Si atendemos al AIC, al excluir cualquiera de las variables anteriores no se consigue una reducción del AIC (como se aprecia en la tabla anterior), por lo que se recomienda conservarlas todas en el modelo.
4. La bondad del ajuste obtenido se juzga con el p-valor asociado a su deviance $D = 61,38$ con 77 g.l., que se obtiene con el comando:

```
1 - pchisq(fit$deviance, fit$df.residual)
```

y que resulta de $p = 0,90334$, con lo que no hay evidencias estadísticas para rechazar el modelo propuesto en el ajuste de los datos observados.

5. Si entendemos que 'predecir correctamente' equivale a que los ceros sean estimados con una $\pi < 0,5$ y los unos con una $\pi \geq 0,5$, podemos calcular el porcentaje de aciertos con:

```
fit.no <- -fitted(fit)[kifosis == "no"]
fit.si <- -fitted(fit)[kifosis == "si"]
table(fit.no < 0,5)/length(fit.no) * 100
table(fit.si >= 0,5)/length(fit.si) * 100
```

de donde obtenemos:

		<i>Predichos</i>	
		<i>no</i>	<i>si</i>
<i>Observados</i>	<i>no</i>	95.31	4.69
	<i>si</i>	58.82	41.18

Es decir, son 'bien' predichos un 95,3% de los que no tienen kifosis y un 41,18% de los que sí la tienen. En total, un 83,95% de los individuos en el ajuste son 'bien' predichos.

Ejemplo 3.5 (Continuación del Ejemplo 3.2: Varios Links para Dosis–Respuesta)

Pretendemos obtener un buen modelo de predicción para la proporción de cucarachas que mueren a un nivel de dosis tóxica x .

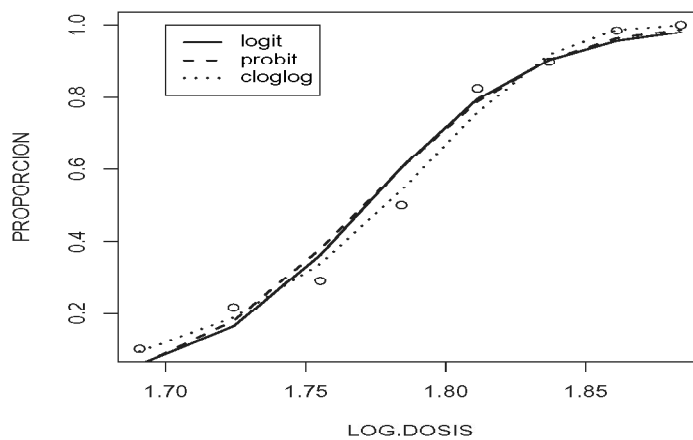


Figura 3.6: Ajustes logit, probit y cloglog para los datos de Cucarachas

Identificación de variables

■ *Respuesta:*

Y_i = número de cucarachas, de un total de n_i que mueren a una determinada dosis x_i

■ *Explicativa:*

ldosis_i = logaritmo de la dosis x_i .

Hipótesis distribucional y Modelización

Los datos consisten en Y_i número de éxitos (muertes) de un total de n_i insectos que han recibido la misma dosis x_i . Podemos asumir que la respuesta proviene de

una distribución binomial

$$Y_i \sim Bi(n_i, \pi_i) \rightsquigarrow E(Y_i) = \mu_i = n_i \pi_i,$$

donde π_i representa la proporción de insectos que mueren a dosis x_i .

En $\mathcal{R}$ hemos de introducir la respuesta binomial en una matriz cuya primera columna contiene el número de éxitos y cuya segunda columna contiene el número de fracasos:

```
ldosis <- -c(1,6907, 1,7242, 1,7552, 1,7842, 1,8113, 1,8369, 1,861, 1,8839)
muertos <- -c(6, 13, 18, 28, 52, 53, 61, 60)
total <- -c(59, 60, 62, 56, 63, 59, 62, 60)
Yres <- -cbind(muertos, vivos <- total - muertos)
```

Ajustaremos tres modelos válidos para datos de dosis-respuesta: *logit*, *probit* y *cloglog*. La Figura 3.6 contiene la representación gráfica de los datos (transformados en proporciones éxitos/fracasos) y los tres ajustes obtenidos.

- Modelo Logit

$$\log \left(\frac{\mu_i}{n_i - \mu_i} \right) = \text{logit}(\pi_i) = \eta_i = \alpha + \beta \text{ldosis}_i$$

$$\hat{\pi}_i = \frac{\exp\{\eta_i\}}{1 + \exp\{\eta_i\}}.$$

La sintaxis y resultados del ajuste en $\mathcal{R}$ son:

$$\text{glm}(Y_{res} \sim \text{ldosis}, \text{family} = \text{binomial}(\text{link} = \text{logit}))$$

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5941	-0.3944	0.8329	1.2592	1.5940

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-60.717	5.173	-11.74	< 2e - 016
ldosis	34.270	2.908	11.79	< 2e - 016

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.20 on 7 degrees of freedom

Residual deviance: 11.23 on 6 degrees of freedom

AIC: 41.43

Con lo que el modelo ajustado resulta:

$$\begin{aligned} \text{logit}(\hat{\pi}_i) &\cong -60,717 + 34,27 \text{ldosis}_i \\ \hat{\pi}_i &= \frac{\exp\{-60,717 + 34,27 \text{ldosis}_i\}}{1 + \exp\{-60,717 + 34,27 \text{ldosis}_i\}}. \end{aligned}$$

Si buscamos la dosis letal LD50 necesaria para garantizar un 50 % de muertes, basta resolver la ecuación:

$$\begin{aligned} \text{logit}(0,5) &= 0 = \alpha + \beta \log(\text{LD50}) \\ \text{LD50} &= \exp(60,717/34,27) = 5,881. \end{aligned}$$

- Modelo Probit

$$\Phi^{-1}(\pi_i) = \alpha + \beta \text{ldosis}_i \Rightarrow \pi_i = \Phi(\alpha + \beta \text{ldosis}_i).$$

La sintaxis y resultados en $\mathcal{R}$ son:

$\text{glm}(Y_{res} \sim \text{ldosis}, \text{family} = \text{binomial}(\text{link} = \text{probit}))$

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5714	-0.4702	0.7501	1.0632	1.3448

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-34.935	2.648	-13.19	< 2e - 016
ldosis	19.728	1.487	13.27	< 2e - 016

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.20 on 7 degrees of freedom

Residual deviance: 10.12 on 6 degrees of freedom

AIC: 40.32

Así, el modelo ajustado resulta:

$$\Phi^{-1}(\hat{\pi}_i) = -34,935 + 19,728 \text{ ldosis}_i \Rightarrow \hat{\pi}_i = \Phi(-34,935 + 19,728 \text{ ldosis}_i).$$

La dosis letal LD50 se obtiene con:

$$\begin{aligned}\Phi^{-1}(0,5) &= 0 = -34,935 + 19,728 \log(LD50) \\ LD50 &= \exp(34,935/19,728) = 5,87.\end{aligned}$$

• Modelo Cloglog

$$\begin{aligned}\log(-\log(1 - \pi_i)) &= \text{cloglog}(\pi_i) = \alpha + \beta \text{ ldosis}_i \\ \pi_i &= 1 - \exp\{-\exp\{\alpha + \beta \text{ ldosis}_i\}\}.\end{aligned}$$

La sintaxis y resultados obtenidos en $\mathcal{R}$ son:

$\text{glm}(Y_{res} \sim \text{ldosis}, \text{family} = \text{binomial}(\text{link} = \text{cloglog}))$				
Deviance Residuals:				
Min	1Q	Median	3Q	Max
-0.80327	-0.55136	0.03090	0.38315	1.28883
Coefficients:				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-39.573	3.237	-12.23	< 2e - 016
ldosis	22.041	1.797	12.26	< 2e - 016
(Dispersion parameter for binomial family taken to be 1)				
Null deviance: 284.202 on 7 degrees of freedom				
Residual deviance: 3.446 on 6 degrees of freedom				
AIC: 33.64				

Y así el modelo ajustado es:

$$\begin{aligned}\text{cloglog}(\hat{\pi}_i) &= -39,573 + 22,041 \text{ ldosis}_i \\ \hat{\pi}_i &= 1 - \exp\{-\exp\{-39,573 + 22,041 \text{ ldosis}_i\}\}.\end{aligned}$$

La dosis letal LD50 la obtenemos con:

$$\begin{aligned}\text{cloglog}(0,5) &= \log(-\log(0,5)) = -0,3665 = -39,573 + 22,041 \log(LD50) \\ LD50 &= \exp\{(39,573 - 0,3665)/22,041\} = 5,922.\end{aligned}$$

Conclusiones

Los tres modelos son significativos a la hora de predecir la proporción de insectos que sobreviven a una determinada dosis. Para concluir sobre la comparación de los tres modelos podemos utilizar el AIC y los residuos. Respecto al primer criterio, el modelo con menor valor del AIC es el modelo cloglog. También atendiendo a la magnitud de los residuos, este modelo supera a los otros dos al presentar la menor dispersión de los residuos deviance, como puede apreciarse del rango intercuartílico en la descriptiva de éstos.

El modelo más conservador respecto de la dosis LD50 es el modelo probit: predice una dosis de insecticida de 5.87 (en escala logarítmica, de 1.77) para con-

seguir un índice de mortalidad en cucarachas del 50 %. El modelo más exigente en dosis para matar a esa proporción de cucarachas es el modelo cloglog.

A continuación presentamos sendas tablas con los valores predichos para los π_i 's y los residuos obtenidos en cada uno de los tres modelos propuestos (Tablas 3.4 y 3.5). Como puede apreciarse, las predicciones son bastante similares en los tres ajustes; con todo, el ajuste cloglog parece capturar mejor las tendencias en los extremos (como puede apreciarse en la Figura 3.6). Las observaciones más influyentes en el cálculo de la deviance (resaltadas en negrita en la Tabla 3.5) para los modelos logit y probit son justamente las de dosis más extremas (por abajo y por encima); en el modelo cloglog son las correspondientes a dosis intermedias. Se vuelve a ratificar que el modelo cloglog ajusta mejor los datos extremos.

prop.obs	est.logit	est.probit	est.clog
0.102	0.059	0.057	0.095
0.217	0.164	0.179	0.188
0.290	0.362	0.379	0.338
0.500	0.605	0.604	0.542
0.825	0.795	0.788	0.758
0.898	0.903	0.904	0.918
0.984	0.955	0.962	0.986
1.000	0.979	0.987	0.999

Tabla 3.4: Proporción de cucarachas muertas observadas y estimadas por los modelos logit, probit y cloglog

ldosis	res.logit	res.probit	res.clog
1.6907	0.778	0.392	0.077
1.7242	0.383	0.145	0.169
1.7552	-0.311	-0.232	-0.174
1.7842	-0.441	-0.269	-0.118
1.8113	0.185	0.130	0.195
1.8369	-0.056	-0.032	-0.094
1.8610	0.667	0.262	-0.030
1.8839	1.016	0.388	0.140

Tabla 3.5: Residuos deviance en los modelos logit, probit y cloglog. Cucarachas.

Ejemplo 3.6 (Continuación del Ejemplo 3.3: Dosis–Respuesta con factor) El objetivo del análisis era detectar si las polillas presentan resistencia a determinadas dosis del tóxico (transcipemetrín), y si la resistencia de machos y hembras es la misma o no. La dosis se trabaja en escala logarítmica en base 2, por lo que ya utilizamos como variable explicativa esta transformación.

Identificación de variables

- *Respuesta:*

Y_{ij} = número de polillas de sexo i muertas de un total de n_i sometidas a dosis x_j

- *Explicativas:*

$ldosis$ = logaritmo en base 2 de la dosis x_j

$sexo_i$ = sexo de la polilla, $i = 1$: hembra; $i = 2$ macho.

Introducimos los datos en $\mathcal{R}$ creando una matriz respuesta con éxitos (vivos) en la primera columna y fracasos (muertos) en la segunda,

```
muertos <- -c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
vivos <- -20 - muertos
Y <- -cbind(vivos, muertos) ~> respuesta

ldosis <- -rep(0 : 5, 2)
sexo <- -factor(rep(c("M", "H"), c(6, 6)))
```

Hipótesis distribucional y Modelización

Se trata de pruebas binomiales de tamaño 20 y probabilidad de éxito (resistencia al tóxico) π_{ij} . Utilizaremos el link logit para relacionar la media $E(Y_{ij}) = \mu_{ij} = 20 \cdot \pi_{ij}$. La modelización logit general en función de las dos variables explicativas

es:

$$Y_{ij} \sim Bi(20, \pi_{ij}), \quad i = 1, 2; \quad j = 1, \dots, 6$$

$$\log \left(\frac{\mu_{ij}}{20 - \mu_{ij}} \right) = \text{logit}(\pi_{ij}) = \theta + \alpha_i + (\beta + \gamma_i) \text{ldosis}_j$$

Análisis.- Contestamos a las siguientes preguntas:

1. La relación entre la resistencia al tóxico y la dosis, ¿es similar para las polillas hembra que para las macho?

CONTESTACIÓN: Estamos planteando el contraste:

$$H_0 : \text{logit}(\pi_{ij}) = \theta + \alpha_i + \beta \text{ldosis}_j$$

$$H_1 : \text{logit}(\pi_{ij}) = \theta + \alpha_i + (\beta + \gamma_i) \text{ldosis}_j$$

en cuyo caso, $\gamma_2 - \gamma_1$ representa el diferente efecto que provoca incrementar la dosis en una unidad más sobre los machos y las hembras.

Resolvemos la cuestión a través de una comparación de los modelos en el contraste:

```
fit0 <- glm(Y ~ sexo + ldosis, family = binomial)
fit1 <- glm(Y ~ sexo * ldosis, family = binomial)
anova(fit0, fit1, test = "Chisq")
```

La diferencia de deviances de los dos modelos de $D_1 - D_0 = 1,7633$, y el p -valor asociado de $p = 0,1842$, con lo que no hay evidencias para rechazar que la dosis afecte de forma similar a hembras y machos. En consecuencia, nos preguntamos a continuación:

2. A una misma dosis de tóxico, ¿resisten igual las polillas hembra que las macho?

CONTESTACIÓN: Estamos planteando el contraste:

$$H_0 : \text{logit}(\pi_{ij}) = \theta + \gamma \text{ldosis}_j$$

$$H_1 : \text{logit}(\pi_{ij}) = \theta + \alpha_i + \gamma \text{ldosis}_j$$

en cuyo caso, $\alpha_2 - \alpha_1$ representaría la diferencia en resistencia entre machos y hembras, a $ldosis_j$.

Resolvemos la cuestión a través de una comparación de los modelos en el contraste:

```
fit0 <- glm(Y ~ ldosis, family = binomial)
fit1 <- glm(Y ~ sexo + ldosis, family = binomial)
anova(fit0, fit1, test = "Chisq")
```

La diferencia de deviances es de $D_1 - D_0 = 10,2270$, con un p -valor asociado de $p = 0,0014$. En consecuencia, se rechaza H_0 a favor de H_1 . Observando el resultado del ajuste del modelo de H_1 ,

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.4732	0.4683	7.417	1,20e - 13
sexoM	-1.1007	0.3557	-3.094	0.00197
ldosis	-1.0642	0.1310	-8.123	4,54e - 16

concluimos que, a una misma dosis de tóxico, los machos resisten menos que las hembras, en concreto, su probabilidad de resistencia es $\exp(-1,1) = 0,33$ veces menor:

$$\begin{aligned} \text{para hembras} &\rightsquigarrow \pi_j = \frac{\exp\{3,47 - 1,06 \, ldosis_j\}}{1 + \exp\{3,47 - 1,06 \, ldosis_j\}} \\ \text{para machos} &\rightsquigarrow \pi_j = \frac{\exp\{3,47 - 1,1 - 1,06 \, ldosis_j\}}{1 + \exp\{3,47 - 1,1 - 1,06 \, ldosis_j\}}. \end{aligned}$$

La tabla anterior demuestra la significatividad de la relación entre la dosis de tóxico administrada y la probabilidad de supervivencia de las polillas.

3. Puesto que se evidencia resistencia de las polillas a determinadas dosis, ¿cuál es la dosis de tóxico a partir de la cual la mitad de la población de polillas ya no resiste y muere?

CONTESTACIÓN: Obviamente habremos de determinar la dosis letal LD50 para hembras y la correspondiente para machos (no serán iguales), sabiendo

ya que la primera será mayor que la segunda, demostrada la mayor resistencia de las hembras. Basta con resolver la ecuación:

$$\text{logit}(0,5) = 0 = \theta + \alpha_i + \beta \log_2(LD50) \rightsquigarrow \log_2(LD50) = -\frac{\theta + \alpha_i}{\beta}$$

$$LD50 = \begin{cases} 9,604 & \text{para hembras} \\ 4,689 & \text{para machos} \end{cases}$$

Ejemplo 3.7 (Link Cloglog: Raíces de Gilligan) Gilligan (1983) describió un test (descrito en Garthwaite et al. (1995), pag. 253) para comprobar la aleatoriedad de la infección de raíces de plantas debida a patógenos nacidos en el suelo. En una serie de experimentos, se plantó un número fijo de plantas en muestras de suelo infectado con diferentes densidades de un inóculo. Cuanto mayor la densidad, mayor la probabilidad de infección. El objetivo del test era únicamente saber si cada raíz estaba infectada o no (sin necesidad de saber el número de infecciones por raíz, como era requerido en otros test desarrollados con anterioridad).

En un experimento, la observación en cada muestra de suelo es Y , el número de raíces (plantas) infectadas de un total plantado n ; esto es, se trata de una variable binomial cuya probabilidad de infección, π , depende de la densidad del inóculo.

Si X es el número de infecciones observadas en cada raíz, es razonable asumir que X tiene una distribución de Poisson con alguna media μ . Así, la probabilidad de que una raíz sufra x infecciones tiene probabilidad

$$\Pr(X = x) = \frac{1}{x!} (kd)^x \exp(-kd).$$

Podríamos asumir, por ejemplo, que para una determinada densidad d de inóculo, el número medio de infecciones por raíz es proporcional a d , es decir, $\mu = kd$, para alguna constante k . Si π es la probabilidad de que una raíz se infecte, entonces el hecho de asumir que $X \sim \text{Po}(\mu)$ nos lleva a afirmar que $\pi = 1 - \exp(-kd)$, dado que $\exp(-kd) = \Pr(X = 0)$ es la probabilidad Poisson de que una raíz no sufra ninguna infección.

Es decir:

$$-\log(1 - \pi) = kd,$$

o lo que es lo mismo,

$$\log(-\log(1 - \pi)) = \log(k) + \log(d).$$

Por tanto, en un GLM que relaciona π , la probabilidad de infección, con d , la densidad de inóculo, es natural considerar el link “complementario log-log” (cloglog).

Una forma más general de modelizar lo anterior es considerar:

$$\log(-\log(1 - \pi)) = \beta_0 + \beta_1 \log(d),$$

que implica que el número medio de infecciones en una raíz es proporcional a una potencia β_1 de la densidad, $\mu = k d^{\beta_1}$.

Es conveniente considerar toda la información que exista sobre el tipo de datos que tenemos para elegir el link a utilizar; el link cloglog surge del argumento de la Poisson. π podría haberse interpretado tanto como la probabilidad de infección como de no infección. La falta de simetría de la función cloglog significa que no obtendremos los mismos resultados si interpretamos π como probabilidad de infección que como probabilidad de no-infección.

En uno de los experimentos que llevó a cabo Gilligan, obtuvo los resultados que aparecen en el fichero roots.dat y en la Tabla 3.6 (no están publicados en Gilligan, 1983, pero aparecen en Garthwaite et al, 1995, pag.253). En este experimento hay dos tipos diferentes de suelo: conductivo (C) y supresivo (S), que se usaron en combinación con cinco densidades (crecientes) de inóculo, y cada densidad se utilizó en tres muestras independientes de cada tipo de suelo.

Identificación de variables

- Variable Respuesta $\rightsquigarrow$

Y_{ijk} = (infect), número de raíces infectadas en el k -ésimo experimento ($k \in \{1, 2, 3\}$), tratado con una densidad de inóculo (en escala logarítmica) $ldosis = j$, con $j \in \{0, 1, 2, 3, 4\}$, y plantada en suelo de tipo $i \in \{0(C), 1(S)\}$.

- Variables Explicativas:

factor $\rightsquigarrow$ suelo = tipo de suelo, con dos posibilidades: C o S.

covariable $\rightsquigarrow$ $ldosis = \log_2(inocden)$ densidad del inóculo en escala logarítmica en base 2. Toma los valores 0,1,2,3 y 4, correspondientes a los valores de $inocden = 1, 2, 4, 8, 16$.

tipo de suelo (suelo)	densidad inóculo (inocden)	infectadas (infect)	Nºraíces	prop infect.
C	1	6	23	0.26
C	2	9	24	0.38
C	4	20	25	0.8
C	8	24	24	1
C	16	21	22	0.95
C	1	1	23	0.04
C	2	14	25	0.56
C	4	17	24	0.71
C	8	21	23	0.91
C	16	24	24	1
C	1	11	24	0.46
C	2	14	24	0.58
C	4	20	26	0.77
C	8	19	23	0.83
C	16	22	23	0.96
S	1	4	23	0.17
S	2	9	21	0.43
S	4	15	24	0.63
S	8	23	23	1
S	16	24	24	1
S	1	7	23	0.3
S	2	16	26	0.62
S	4	23	25	0.92
S	8	17	18	0.94
S	16	25	25	1
S	1	5	22	0.23
S	2	8	24	0.33
S	4	23	25	0.92
S	8	23	23	1
S	16	26	26	1

Tabla 3.6: Datos de raíces infectadas de Gilligan.

Hipótesis distribucional y Modelización

Es claro que la respuesta tiene una distribución binomial (éxito=infección en la planta):

$$Y_{ijk} \sim Bi(n_{ijk}, \pi_{ijk}).$$

Según el razonamiento que hicimos arriba, es lógico considerar el link cloglog para ligar la probabilidad de infección con la densidad de inóculo. El tipo de suelo actuará como posible factor de clasificación para diferenciar respuestas.

A la hora de proponer modelos es conveniente llevar a cabo una inspección gráfica de los datos. Representamos los cloglog empíricos versus cada una de las variables explicativas. La función cloglog empírica viene dada por:

$$ecll = \log \left(-\log \left[1 - \frac{y + 1/2}{n + 1} \right] \right).$$

En la Figura 3.7 apreciamos cierta tendencia lineal entre $\text{cloglog}(\pi)$ y ldosis , además de diferencias según el tipo de suelo. Tiene sentido pues, plantear un modelo lineal generalizado que relacione de modo lineal ldosis con el $\text{cloglog}(\pi)$, considerando posibles diferencias respecto al tipo de suelo.

- *Modelo 1: Modelo Maximal.*

Interpretando el $\text{ldosis} = \log_2(\text{inocden})$ como un factor ordenado, en lugar de como una covariable continua, ajustamos un polinomio de orden 4 para cada uno de los tipos de suelo. Con eso, estimamos 10 parámetros y agotamos los 10 grados de libertad que nos proporcionan los 5 niveles de densidad de inóculo en cada tipo de suelo:

$$\text{cloglog} \left[E \left(\frac{Y_{ijk}}{n_{ijk}} \right) \right] = \sum_{l=0}^4 \alpha_{il} (\text{ldosis}_{ijk})^l.$$

Este ajuste lo obtenemos considerando ldosis como un factor ordenado,

```
fit1 <- glm(Y ~ suelo * ordered(ldosis),
            family = binomial(link = cloglog))
```

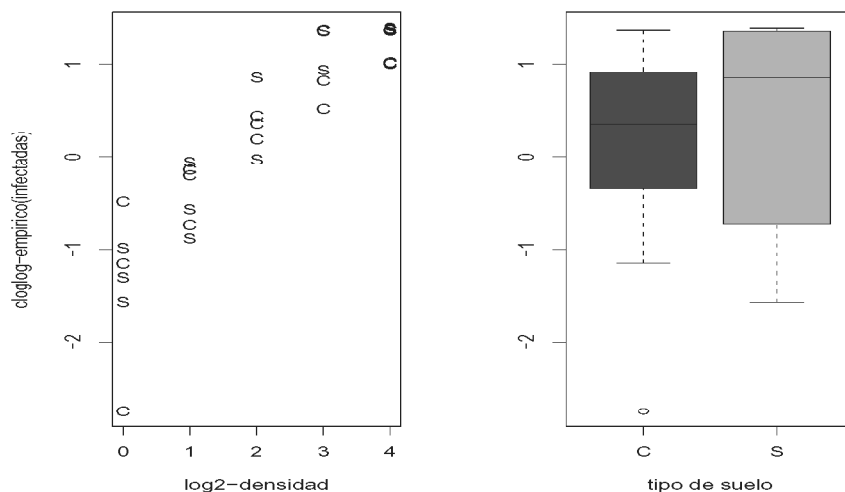


Figura 3.7: Datos de las raíces de Gilligan

Si el modelo maximal es un buen ajuste, entonces su deviance media $D/(n - p)$, debería estar cerca de 1, que es el parámetro de escala para la distribución binomial $(\phi - 1)$. Sin embargo,

$$D = 40,057 \text{ con } 20 \text{ g.l.},$$

de donde $\hat{\phi} = 2,00$, lo que sugiere que los datos están “sobre-dispersos”, es decir, que tienen una varianza mayor de la esperada para datos binomiales. Por lo tanto, para realizar la comparación de modelos será preferible usar los tests F a los χ^2 . La utilidad de este modelo es meramente la de identificar el problema de sobre-dispersión.

- **Modelo 2: Modelo Parsimonioso**

Nuestro objetivo, sin embargo, es obtener el modelo más simple posible que explique razonablemente bien la variabilidad de los datos, utilizando la estructura lineal que hemos comentado anteriormente para la relación entre los $\text{cloglog}(\pi)$ y la variable ldosis . Planteamos un modelo en el que pretendemos ajustar una curva distinta para cada tipo de suelo, con la que describir la relación entre ldosis y la

probabilidad de infección π :

$$\text{cloglog} \left[E \left(\frac{Y_{ijk}}{n_{ijk}} \right) \right] = \theta + \alpha_i + \beta_i \text{ldosis}_{ijk}.$$

La sintaxis en $\mathcal{R}$ para lograr dicho ajuste es:

```
fit2 <- glm(Y ~ suelo * ldosis, family = binomial(link = cloglog))
```

En la Figura 3.8 aparecen representadas las observaciones y el ajuste obtenido, tanto en la escala cloglog (rectas de la izquierda), como en la escala de las proporciones (curvas de la derecha).

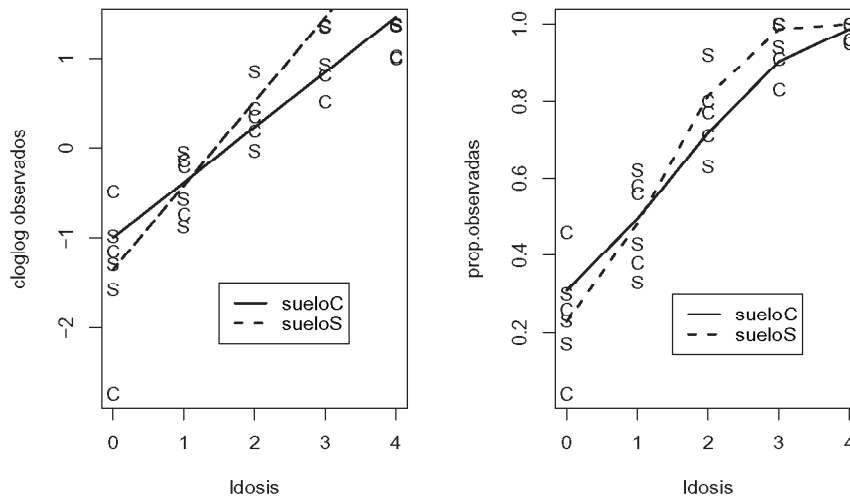


Figura 3.8: Ajuste GLM con el link cloglog de los Datos de Gilligan.

Conclusiones:

Contestamos a continuación a diversas preguntas sobre el ajuste realizado en función de los resultados obtenidos.

1. La diferencia entre lo que explica este modelo (*fit2*) y el maximal (*fit1*), ¿es significativa? Este modelo lineal, ¿es razonable para explicar la variabilidad de los datos?, ¿en base a qué?

```
anova(fit1, fit2, test = "F")
```

CONTESTACIÓN: La diferencia de deviances entre ambos modelos, *fit1* y *fit2*, es de 2,9, lo que da lugar a un estadístico $F = 0,483$, con un p -valor asociado de 0,8212 (obtenido de una $F_{6,10}$). La conclusión es que no se obtiene evidencia suficiente para rechazar el modelo más simple *fit2* a favor del más complejo *fit1*. La estimación de 10 parámetros en *fit1* no rentabiliza la ganancia en explicación de los datos, respecto a la conseguida con el modelo en que sólo se estiman 4 parámetros (*fit2*).

2. La densidad de inóculo, ¿afecta de forma diferente a la proporción de raíces infectadas para cada tipo de suelo?

```
anova(fit2, test = "F")
```

CONTESTACIÓN: El p -valor asociado al contraste del término de interacción $S : l_{dosis}$ resulta de 0,005801, claramente significativo. En consecuencia, se rechaza interacción nula a favor de que la dosis (*ldosis*) afecta de forma distinta a las plantas cultivadas en uno y otro tipo de suelo.

3. Queremos simplificar el modelo al máximo, y proponemos ajustar una misma curva a todos los datos, sea cual sea al tipo de suelo del que provienen. ¿Es razonable este ajuste?

```
fit3 <- glm(Y ~ ldosis, family = binomial(link = cloglog))
anova(fit2, fit3, test = "F")
```

CONTESTACIÓN: Prescindir del factor 'suelo' proporciona el ajuste *fit3*. El p -valor asociado a la comparación de los ajustes *fit2* y *fit3* (contraste F) es igual a 0,005317, lo que evidencia el rechazo del modelo más simple, *fit3*, a favor del más complejo, *fit2*. Esto es, el tipo de suelo es significativo para explicar la probabilidad de infección.

3.7. Ejercicios

1. En un ensayo clínico con tres tipos de insecticidas (A, B y C), se utilizaron grupos de aproximadamente 50 insectos y fueron expuestos a depósitos variables de cada insecticida. Las proporciones de insectos muertas después de la exposición durante seis días aparece en la Tabla 3.7 (Hewlett y Placket, 1950, también en Krzanowski, 1998, pag.199).

Insecticida	Cantidad de depósito (mg)					
	2.00	2.64	3.48	4.59	6.06	8.00
A	3/50	5/49	19/47	19/38	24/49	35/50
B	2/50	14/49	20/50	27/50	41/50	40/50
C	28/50	37/50	46/50	48/50	48/50	50/50

Tabla 3.7: Pruebas de insecticidas

2. Un radiólogo llevó a cabo un experimento con 14 platos de bacterias. Los platos fueron expuestos a varias dosis de radiación, y el investigador anotó la proporción de las células que sobrevivieron. Los datos están en la Tabla 3.8 (Efron y Tibshirani, 1993, pag.116).

Dosis mayores de radiación dieron lugar a proporciones más pequeñas de supervivencia, como se esperaba. Sin embargo, la respuesta en el plato 13 aportaba algo de incertidumbre a la conclusión que dio el investigador. ¿Cuáles son tus conclusiones?

3. Los datos de la Tabla 4.30 describen los patrones de comportamiento en el consumo de drogas psicotrópicas en una muestra de individuos del Oeste de Londres (Everitt y Dunn, 1984, pag.161).
 - a) Identifica las variables respuesta y explicativas, la distribución de la respuesta y el parámetro de interés. Justifica la elección del link que vas a utilizar.
 - b) ¿Hay diferencias por 'sexo' en el consumo de drogas? ¿Cómo influye la 'edad' para explicar el consumo de psicotrópicos? ¿Detectas un efecto cuadrático de la edad para explicar el consumo de drogas? ¿La edad influye igual en hombres y en mujeres?
 - c) Obtén la expresión y el valor de las predicciones sobre el consumo de psicotropos en hombres y en mujeres de 17 y 52 años con el modelo ajustado.

plato	dosis	prop.superv
1	1.175	0.44000
2	1.175	0.55000
3	2.350	0.16000
4	2.350	0.13000
5	4.700	0.04000
6	4.700	0.01960
7	4.700	0.06120
8	7.050	0.00500
9	7.050	0.00320
10	9.400	0.00110
11	9.400	0.00015
12	9.400	0.00019
13	14.100	0.00700
14	14.100	0.00006

Tabla 3.8: Efecto de la radiación en la supervivencia de las células

4. En la Tabla 3.10 (Breslow y Day, 1980, también en McCullagh y Nelder, 1989, pag.268) se presentan los datos de un estudio clínico realizado con la finalidad de investigar el efecto del consumo de alcohol sobre la incidencia del cáncer de esófago (la ratio del número de individuos con cáncer entre el de individuos sin cáncer). Puesto que es común encontrar que la incidencia de cáncer aumenta con la edad, también este factor se controló en el estudio a través de una estratificación por edades. En la Figura 3.9 está representado el logaritmo de la ratio de incidencia versus la edad, para los dos niveles de consumo de alcohol considerados (alto + y bajo 0).

El objetivo del análisis a realizar es el de modelizar y predecir la incidencia del cáncer de esófago en función de las variables controladas en el estudio: edad (categorizada en 6 grupos) y consumo de alcohol (alto (80+)/bajo (80-)).

- Identifica la variable respuesta y las variables explicativas. Especifica su tipo.
- Asumiendo fijos los totales marginales muestrales correspondientes a los factores edad y consumo de alcohol, ¿cuál es la distribución que se puede asumir para la variable respuesta? ¿Y el parámetro de interés?
- El cáncer de esófago, ¿guarda alguna relación con la edad y con el consumo de alcohol? El efecto del alcohol sobre la incidencia de cáncer, ¿es similar en todas las edades?

Sexo	Edad	Usa	No usa
hombre	16-29	21	683
hombre	30-44	32	596
hombre	45-64	70	705
hombre	65-74	43	295
mujer	16-29	46	738
mujer	30-44	89	700
mujer	45-64	169	847
mujer	65-74	51	196

Tabla 3.9: Uso de drogas psicotrópicas en Londres

edad	alcohol	cancer	nocancer
25-34	80+	1	9
	80-	1	106
35-44	80+	4	26
	80-	5	164
45-54	80+	25	29
	80-	21	138
55-64	80+	42	27
	80-	34	139
65-74	80+	19	18
	80-	36	88
75-84	80+	5	1
	80-	8	31

Tabla 3.10: Incidencia del cáncer de esófago versus la edad y el consumo de alcohol

- d) Comenta el ajuste obtenido utilizando los residuos.
 - e) Calcula las predicciones para la incidencia de cáncer de esófago en cada grupo de edad y en función de si el individuo consume mucho o poco alcohol.
5. Considera los datos de la Tabla 3.11, que corresponden al estatus de supervivencia tras tres años, *vivo-muerto*, de pacientes con cáncer de mama, clasificadas por su edad (en tres grupos) y por el tipo de tumor (*maligno o benigno*).

Tratando a los números de pacientes que sobrevivieron en cada fila C_i (Y_{ij}) como observaciones binomiales independientes, y a 'edad' y 'maligno' como factores,

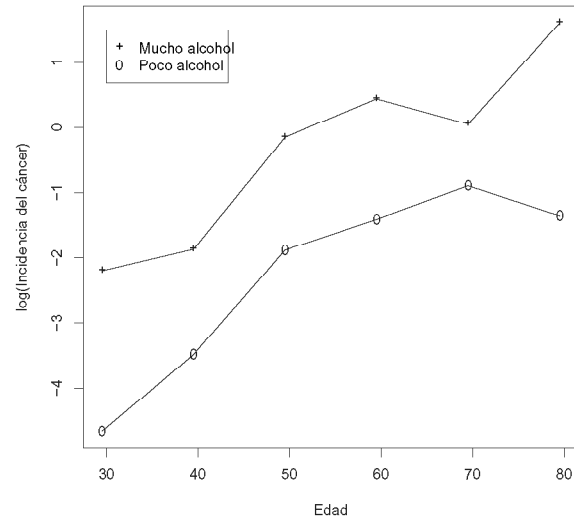


Figura 3.9: Logaritmo de la incidencia del cáncer en función de la edad y del consumo de alcohol

- a) Ajusta un buen modelo logístico para predecir supervivencia con la edad y el tipo de tumor. Formula explícitamente dicho modelo y comenta las conclusiones.
 - b) Escribe la expresión que utilizarías para estimar la probabilidad de que sobreviva una paciente de menos de 50 años y con un tumor benigno. Calcula esta probabilidad y también el número esperado (según este modelo) de pacientes que sobreviven, de las 87 que fueron clasificadas con menos de 50 años y un tumor no maligno.
6. Un estudio casos-control se llevó a cabo con el fin de identificar razones que motivaran las altas tasas de cáncer de pulmón entre los residentes masculinos de la costa de Georgia (Blot et al, 1978, también en Le, 1998, pag.3). Los casos fueron identificados a partir de estas fuentes:
 - a) diagnósticos desde 1970 en el único hospital grande en Brunswick,
 - b) diagnósticos durante 1975-76 en los tres hospitales más grandes en Savannah, y
 - c) certificados de muerte para el período 1970-74 en el área.

	edad	maligno	vive	
			si	no
C_1	< 50	no	77	10
C_2	< 50	si	51	13
C_3	50 – 69	no	51	11
C_4	50 – 69	si	38	20
C_5	> 70	no	7	3
C_6	> 70	si	6	3

Tabla 3.11: Supervivencia en pacientes de cáncer.

Los controles fueron seleccionados de las admisiones a los cuatro hospitales y de certificados de muerte en el mismo período, cuando los diagnósticos habían sido diferentes a cáncer de pulmón, de vejiga, o cáncer crónico de pulmón. Los datos fueron separados en fumadores y no fumadores y aparecen en la Tabla 3.12.

El objetivo del análisis es investigar si el hecho de haber trabajado en los astilleros durante la Segunda Guerra Mundial podría haber afectado a la alta tasa de cáncer de pulmón en la población masculina de Georgia.

Fuma	Astilleros	Casos	Controles
No	Sí	11	35
	No	50	203
Sí	Sí	84	45
	No	313	270

Tabla 3.12: Cáncer de pulmón en los astilleros de Georgia

- Se llevó a cabo un test citológico para el diagnóstico de cáncer cervical en mujeres (May, 1974, también en Le, 1998, pag. 17). Se consideró a un grupo de 24.103 mujeres de las cuales 23.724 tenían unas cervicales saludables, y las 379 restantes tenían alguna abnormalidad (respecto a la posibilidad de cáncer). Los datos se presentan en la Tabla 3.13.

El objetivo es estudiar la *sensibilidad* del test citológico (en el sentido de que si se hace el test una mujer con cáncer cervical, el resultado del test debería ser negativo).

- Durante muchos años se ha relacionado la pancreatitis con el estatus de fumador. Para demostrar estadísticamente esta relación, se llevó a cabo un es-

Enfermedad	Resultado del Test		
	Negativo (-)	Positivo (+)	Total
Negativo (-)	23.362	362	23.724
Positivo (+)	225	154	379

Tabla 3.13: Cáncer cervical y sensibilidad del test citológico.

tudio en el este de Massachusetts y Rhode Island entre 1975 y 1979 (Yen et al., 1982, también en Le, 1998, pag.25). Noventa y ocho pacientes que habían recibido el alta de un diagnóstico de pancreatitis en el hospital fueron incluidos en el estudio casos-contróles (no-pareado). El grupo control consistió en 451 pacientes admitidos en el hospital por otros motivos diferentes al páncreas o al tracto biliar. Se obtuvo información sobre el factor de riesgo a partir de una entrevista estandarizada con cada sujeto, llevada a cabo por un entrevistador entrenado. Los datos se presentan en la Tabla 3.14.

Uso de tabaco	Casos	Controles
Fumador habitual	38	81
Ex-fumador o no fumador	15	136
Total	53	217

Tabla 3.14: Hábitos tabaquísticos y pancreatitis.

9. En 1979, la Administración de Veteranos de los Estados Unidos llevó a cabo una encuesta de salud de los veteranos de guerra (True et al, 1988, también en Le, 1998, pag.26). Los datos que se presentan en la Tabla 3.15 se refieren al servicio en Vietnam y a problemas de sueño en los 1787 veteranos que entraron en el servicio militar entre 1965 y 1975. ¿Hay alguna relación entre el hecho de haber prestado servicio en Vietnam y tener posteriormente problemas de sueño?

Problemas de sueño	Servicio en Vietnam		
	Sí	No	Total
Sí	173	160	333
No	599	851	1450
Total	772	1011	1783

Tabla 3.15: Problemas de sueño en los veteranos de Vietnam

10. En Auckland, New Zealand, se llevó a cabo un estudio casos-contróles para investigar los efectos del consumo de alcohol en el infarto de miocardio no fatal y en la muerte coronaria durante las 24 horas siguientes al consumo de alcohol, entre bebedores habituales (Jackson et al, 1992, también en Le, 1998, pag.39). En la Tabla 3.16 se presentan los resultados de muerte coronaria para hombres. ¿Hay alguna relación?

Bebe en las últimas 24 horas	Casos	Controles
Sí	69	159
No	103	135
Total	172	294

Tabla 3.16: Muerte coronaria y alcohol.

11. Se llevó a cabo un estudio casos-contróles para investigar la relación entre el infarto de miocardio (IM) y el uso de contraceptivos orales (CO) (Shapiro et al., 1979, también en Le, 1998, pag.47). Los datos, estratificados por estatus de fumador, se presentan en la Tabla 3.17.

¿Hay alguna relación entre el IM y el uso de CO? El efecto de los CO sobre el IM, ¿es similar para las mujeres fumadoras y no fumadoras?

Fuma	utiliza OC	Casos	Controles
No	Sí	4	52
	No	34	754
Sí	Sí	25	83
	No	171	853

Tabla 3.17: Muerte coronaria y uso de contraceptivos.

12. Se llevó a cabo un estudio para investigar los efectos del consumo de café, tabaco y alcohol sobre el cáncer de ovarios (Whittemore et al., 1988, también en Le, 1998, pag.49). Los sujetos del estudio son 188 mujeres de la zona de la Bahía de San Francisco que habían sido diagnosticadas de cáncer de ovarios epitelial en 1986-85, y 539 mujeres control. De las 539 controles, 280 fueron hospitalizadas sin cáncer y 259 fueron elegidas de la población total por llamada telefónica aleatorizada. Los datos para el consumo de café se presentan en la Tabla 3.18.

¿Hay relación entre el cáncer de ovarios y el consumo de café si consideramos

sólo los controles del hospital? ¿Y si usamos los controles de la población? ¿Y con todos los controles?

Bebe café	Casos	Controles Hospital	Controles población	Total
Sí	177	249	233	659
No	11	31	26	68
Total	188	280	259	727

Tabla 3.18: Consumo de café y cáncer de ovarios.

13. En la Tabla 3.19 se presentan los resultados de una encuesta. A cada sujeto de una muestra de 300 adultos se le pidió que opinara sobre qué política consideraba adecuada implantar respecto al uso de tabaco en lugares públicos (Daniel, 1987, también en Le, 1998, pag.51). Los datos se presentan en la Tabla 3.19.

¿Hay alguna relación entre la actitud frente al tabaco y el nivel de estudios?

Nivel educacional	Actitud ante el uso de tabaco				Total
	sin restricciones	fumar sólo en áreas exclusivas	No fumar nunca	No opina	
Est.Superiores	5	44	23	3	75
Bachillerato	15	100	30	5	150
Grad. escolar	15	40	10	10	75
Total	35	184	53	18	300

Tabla 3.19: Estudios y tabaco en lugares públicos.

14. Un ejemplo que concierne al uso de cinturones de seguridad en automóviles (Le, 1998, pag.53). Cada accidente es clasificado en función de si se utilizó el cinturón de seguridad y la severidad de los daños recibidos: ninguno, menor, mayor o muerte. Los datos se presentan en la Tabla 3.20.

El uso del cinturón de seguridad, ¿reduce el riesgo de daños graves en un accidente?

15. En el Hospital de Yale-New Heaven, en Connecticut, se llevó a cabo un estudio para investigar la relación entre los nacimientos prematuros (el niño nazca antes de 37 semanas de gestación o su peso sea inferior a 2500 g.) y la edad de la madre (Berkowitz, 1981, también en Le, 1998, pag.56). La población de estudio consistió en 175 madres de niños nacidos únicos y prematuros, y 303 madres de niños no prematuros. Los datos se presentan en la Tabla 3.21.

Cinturón de seguridad	Severidad de los daños			
	ninguno	menor	mayor	muerte
Sí	75	160	100	15
No	65	175	135	25

Tabla 3.20: Cinturón de seguridad y daños en accidentes de tráfico.

¿Hay alguna relación entre la edad de la madre y el hecho de que un niño nazca prematuro? ¿Cuál es el grupo de edad con mayor riesgo?

Edad	Casos	controles	Total
14-17	15	16	31
18-19	22	25	47
20-24	47	62	109
25-29	56	122	178
≥ 30	35	18	113
Total	175	303	

Tabla 3.21: Edad de la madre y niños prematuros.

16. La relación entre uso de condón y embarazo en las trompas fue investigado en un estudio caso-control por el Group Health Cooperative de Puget Sound durante 1981-86 (Li, et al., 1990, también en Le, 1998, pag. 62). Los datos se presentan en la Tabla 3.22.

¿Existe alguna relación significativa?

	Casos		controles
	No	176	488
Usos de condón	Sí	51	186

Tabla 3.22: Uso de condón y embarazos.

17. La Tabla 3.23 proporciona los datos de un estudio sobre la asociación entre raza y uso de los servicios médicos en adultos con dolor de pecho durante el año 1989 (Strogatz, 1990, también en Le, 1998, pag.63). Estudiar la relación entre la utilización de servicios médicos y la raza.

Uso de servicios médicos	negros	blancos
sí el año pasado	35	67
sí, pero no el año pasado	45	38
nunca	78	39
Total	158	144

Tabla 3.23: Uso de servicios médicos y raza.

18. Estudia la sensibilidad de un test basado en el diagnóstico de la tuberculosis a través de una prueba basada en rayos X (Begg y McNeil, 1988, también en Le, 1998, pag. 63). Los datos aparecen en la Tabla 3.24.

Rayos X	Tuberculosis		
	No	Sí	Total
negativo	1739	8	1747
positivo	51	22	73
total	1790	30	1820

Tabla 3.24: Sensibilidad del diagnóstico de tuberculosis con rayos X.

19. A partir de los datos proporcionados por un estudio de Negri et al, (1988) (también en Le, 1998, pag. 65), en varias regiones italianas se recalculó la relación entre factores menstruales y reproductores sobre el cáncer de mama. Los datos aparecen resumidos para *edad a la menopausia* y *edad al primer parto vivo* en la Tabla 3.25.

Investiga la asociación entre estos factores y el cáncer de mama.

Edad en el primer parto (vivo)	Casos	Controles
< 22	621	898
22 – 24	795	909
25 – 27	791	769
≥ 28	1043	775
Edad a la menopausia	Casos	Controles
< 45	759	543
45 – 49	749	803
≥ 50	1378	1167

Tabla 3.25: Cáncer de mama, edad, reproducción y menopausia.

20. Las mujeres post-menopáusicas que desarrollan cáncer endometrial suelen ser, en general, más pesadas que las que no desarrollan la enfermedad. Una posible explicación es que las mujeres gruesas están más expuestas a estrógenos endógenos producidos tras la menopausia en las mujeres por conversión de precursores esteroides a estrógenos activos en la grasa periférica. Considerando varios niveles de producción de estrógeno endógeno, uno podría preguntarse si el potencial carcinogénico de los estrógenos exógenos residiría en la grasa periférica. Considerando diversos niveles de producción de estrógenos endógenos, uno se podría preguntar si el potencial carcinogénico de los estrógenos exógenos sería el mismo para todas las mujeres. Se llevó a cabo un estudio casos-contrroles para examinar la relación entre peso, terapia de reemplazamiento estrógeno y cáncer en el endometrio (Kelsey et al., 1982, también en Le, 1998, pag. 66), cuyos datos aparecen en la Tabla 3.26. ¿Se evidencian las relaciones hipotetizadas?

peso (kg)		Reemplazamiento de estrógeno	
		Sí	No
< 57	casos	20	12
	contrroles	61	183
57 – 75	casos	37	45
	contrroles	113	378
> 75	casos	9	42
	contrroles	23	140

Tabla 3.26: Cáncer en el endometrio, edad y estrógenos.

4

Tablas de Contingencia y Modelos Log-Lineales

4.1. Introducción y Ejemplos

Consideramos en este tema el análisis de tablas de contingencia, caracterizadas por:

- Las observaciones consisten en conteos o frecuencias de clasificación de los individuos observados en las celdas de una tabla de contingencia.
- Las variables respuestas y explicativas son todas categóricas, medidas en una escala nominal u ordinal.
- El objetivo básico en el análisis de tablas de contingencia es estudiar si existe alguna relación/asociación entre los factores de clasificación considerados. Posteriormente, si dicha relación existe, habrá que describir cómo es.
- Para cuantificar el grado de asociación entre los factores de clasificación, los modelos log-lineales proporcionarán predicciones de las frecuencias obser-

vadas en las celdas y las probabilidades asociadas a la clasificación, en función de dichos factores y la interacción entre ellos.

Ejemplo 4.1 (Melanoma) Los datos de la Tabla 4.1 (Roberts et al, 1981, también en Dobson, 1995, pag.121) provienen de un estudio de pacientes con una forma de cáncer de piel llamado melanoma maligno. En una muestra de $N = 400$ pacientes se recogió información sobre la localización del tumor y su tipo histológico. Los datos son el número de pacientes en cada combinación de tipo de tumor y localización.

Tipo de tumor	Cabeza/cuello	Localización		Total
		Tronco	Extremidades	
Melanoma de Hutchinson	22	2	10	34
Melanoma superficial	16	54	115	185
Nodular	19	33	73	125
Indeterminado	11	17	28	56
Total	68	106	226	400

Tabla 4.1: Melanoma maligno: frecuencias para tipo de tumor y localización.

El objetivo básico del análisis es investigar si determinados tipos de tumor están asociados a ciertas localizaciones. Ambas variables, tipo de tumor y localización, tienen pues, el mismo rol en el análisis, y sólo nos interesa determinar si están asociadas o no. Por otro lado, las frecuencias en las celdas se pueden contemplar como realizaciones de variables aleatorias cuya suma es N , el número total de individuos que se fijó en el estudio. El muestreo produce pues frecuencias multinomiales.

Ejemplo 4.2 (Vacuna de la gripe) En un estudio prospectivo sobre una nueva vacuna para la gripe, los pacientes fueron asignados aleatoriamente a dos grupos. A los pacientes de uno de los grupos se les trató con la nueva vacuna y a los otros se les dió un placebo salino. Las respuestas fueron los niveles de anticuerpos inhibidores de hemaglutinina (HIA) encontrados en la sangre seis semanas después de la vacunación. Los datos se encuentran en la Tabla 4.2 (Dobson, 1995, pag.125).

El objetivo del estudio es investigar el efecto de la nueva vacuna, esto es, comprobar si el hecho de dar placebo o vacuna provoca diferente respuesta HIA. Así, la variable 'respuesta HIA' es la variable a explicar en función del tipo de tratamiento que ha recibido el paciente (variable explicativa). Las frecuencias de las celdas en las filas de la Tabla 4.2 están restringidas a sumar el número de sujetos en cada

	Respuesta			
	Pequeño	Moderado	Grande	Total
placebo	25	8	5	38
vacuna	6	18	11	35

Tabla 4.2: Ensayo clínico sobre la vacuna de la gripe.

grupo de tratamiento (35 y 38 respectivamente); esto es, las frecuencias de cada fila constituyen realizaciones de distribuciones multinomiales independientes.

Ejemplo 4.3 (Aspirina) En un estudio retrospectivo de casos–controles, se reunió a un grupo de pacientes con úlcera y se buscó a otro grupo de individuos, de similares características a los anteriores en edad, sexo y estatus socio–económico, sobre los que no se sabía que tuvieran úlcera péptica. Los pacientes con úlcera fueron clasificados de acuerdo a la localización de la úlcera: gástrica o duodenal. Se les preguntó a todos si consumían aspirina. Los resultados aparecen en la Tabla 4.3 (Duggan et al, 1986, también en Dobson, 1995, pag. 125).

	Consumo de Aspirina		Total
	No consume	Consume	
Úlcera gástrica			
Casos	39	25	64
Controles	62	6	68
Úlcera duodenal			
Casos	49	8	57
Controles	53	8	61

Tabla 4.3: Relación entre las úlceras gástrica - duodenal y el consumo de aspirina.

El objetivo del análisis es averiguar si existe alguna relación entre el consumo de aspirina y la existencia de algún tipo de úlcera. Si existe tal asociación, querríamos comprobar si el consumo de aspirina es más habitual en los pacientes con úlcera gástrica que en los de úlcera duodenal. Interesa predecir el consumo de aspirina (variable a explicar) en función de si un individuo tiene úlcera o no, y si la tiene, de qué tipo es. Para conseguir la muestra de individuos se fijó un tiempo para la realización del estudio durante el que se reclutó a todos los pacientes/sanos de un hospital disponibles y que cumplieran los requisitos especificados. No se fijó pues, ningún tamaño de muestra, por lo que podemos asumir que los conteos que se obtuvieron

proviene de una distribución de Poisson.

En los ejemplos comentados hemos distinguido las condiciones básicas que podemos encontrar al analizar una tabla de contingencia: respecto a la identificación de variables a explicar y variables explicativas, y respecto al tipo de muestreo que se ha llevado a cabo para obtener las frecuencias de las celdas. En este tema planteamos, a través de modelos lineales generalizados, y en particular de modelos log-lineales, la búsqueda de asociaciones entre factores de clasificación y la predicción de la frecuencia media en cada una de las celdas, en función de dichos factores y las diferencias entre sus niveles.

4.2. Tipos de muestreo y distribución de frecuencias

Consideramos una tabla de contingencia de dos vías (con dos factores de clasificación), F_1 y F_2 , con I y J niveles de clasificación respectivamente. Sea Y_{ij} la frecuencia observada (número de individuos clasificados) en el nivel $i \in \{1, \dots, I\}$ del factor F_1 y en la categoría $j \in \{1, \dots, J\}$ del factor F_2 . Dicha variable constituirá la variable respuesta en los modelos que consideramos.

Puesto que cada individuo en el estudio es clasificado en la tabla independientemente de todos los demás, podemos hablar de *independencia* en las observaciones. El tipo de muestreo utilizado determinará la distribución de las frecuencias Y_{ij} . Distinguimos básicamente tres situaciones:

Muestreo Poisson. Ningún total es fijado previamente al estudio. Se observa durante un cierto período de tiempo y se clasifica a los individuos a los que se ha podido acceder en ese tiempo. Entonces se puede considerar que las frecuencias observadas provienen de una distribución de Poisson,

$$Y_{ij} \sim Po(\mu_{ij}) \rightsquigarrow E(Y_{ij}) = \mu_{ij},$$

y la probabilidad de que el número de individuos clasificados en la celda (i, j) sea y_{ij} viene dada por

$$Pr(Y_{ij} = y_{ij}) = \mu_{ij}^{y_{ij}} \exp(-\mu_{ij}) / y_{ij}!$$

La verosimilitud conjunta para todas las $I \times J$ celdas es:

$$L(\mu; \mathbf{y}) = \prod_{i=1}^I \prod_{j=1}^J \frac{\mu_{ij}^{y_{ij}} \exp(-\mu_{ij})}{y_{ij}!},$$

de donde la log-verosimilitud se puede expresar como:

$$l(\mu; \mathbf{y}) = \text{cte} + \sum_{i=1}^I \sum_{j=1}^J y_{ij} \log[E(Y_{ij})] - E(Y_{ij}). \quad (4.1)$$

Muestreo Multinomial. Inicialmente se fija el tamaño muestral total n a observar. Los n individuos a observar son clasificados en las $I \times J$ celdas de la tabla, de donde las frecuencias se pueden considerar como observaciones multinomiales,

$$(Y_{11}, \dots, Y_{IJ}) \sim MN(n, (\pi_{11}, \dots, \pi_{IJ})) \rightsquigarrow E(Y_{ij}) = n\pi_{ij},$$

donde π_{ij} representa la probabilidad de clasificar a un individuo en la celda (i, j) .

La verosimilitud resulta:

$$L(\pi; \mathbf{y}) = \frac{n!}{\prod_i \prod_j y_{ij}!} \prod_{i=1}^I \prod_{j=1}^J \pi_{ij}^{y_{ij}},$$

de donde la log-verosimilitud se puede escribir como:

$$l(\mu; \mathbf{y}) = \text{cte} + \sum_{i=1}^I \sum_{j=1}^J y_{ij} \log[E(Y_{ij})]. \quad (4.2)$$

Nota 4.2.1 La relación entre estas probabilidades condicionadas multinomiales π_{ij} y las frecuencias medias esperadas (marginales) en un muestreo Poisson se obtiene de considerar el total de observaciones n , como la realización de una variable Poisson con media $\mu = \sum_i \sum_j \mu_{ij}$ (al ser suma de variables Poisson independientes),

$$\pi_{ij} = \mu_{ij} / \mu.$$

Muestreo Producto Multinomial. Inicialmente se fija algún margen, esto es, el número de individuos a observar en cada una de las categorías de algún o algunos

factores de clasificación.

Supongamos que se fija el margen F_1 , esto es, el número de individuos a observar en cada una de las I categorías del factor F_1 , $y_{1+}, \dots, y_{I+}$. Así, las observaciones de cada fila i se pueden considerar realizaciones de una multinomial de tamaño y_{i+} y probabilidades $\{\rho_{ij}, j = 1, \dots, J\}$, donde ρ_{ij} representa la probabilidad condicionada de clasificar a un individuo en la categoría j del factor F_2 sabiendo que ya ha sido clasificado en el nivel i de F_1 ,

$$\rho_{ij} = Pr(F_2 = j | F_1 = i) = \pi_{j|i}.$$

Para cada fila,

$$(Y_{i1}, \dots, Y_{iJ}) \sim MN(y_{i+}, (\rho_{i1}, \dots, \rho_{iJ})) \rightsquigarrow E(Y_{ij}) = y_{i+}\rho_{ij},$$

es decir, la verosimilitud de cada fila vendrá dada por:

$$L(\pi_i; \mathbf{y}_i) = \frac{y_{i+}!}{\prod_j y_{ij}!} \prod_{j=1}^J \rho_{ij}^{y_{ij}},$$

y la conjunta de todas las filas será la producto multinomial:

$$L(\pi; \mathbf{y}) = \prod_i \frac{y_{i+}!}{\prod_j y_{ij}!} \prod_{i=1}^I \prod_{j=1}^J \rho_{ij}^{y_{ij}}.$$

Así, la log-verosimilitud la podremos escribir como:

$$l(\mu; \mathbf{y}) = cte + \sum_{i=1}^I \sum_{j=1}^J y_{ij} \log[E(Y_{ij})]. \quad (4.3)$$

Nota 4.2.2 La relación entre estas probabilidades condicionadas ρ_{ij} en el muestreo producto multinomial y las del muestreo multinomial, π_{ij} , viene dada por $\rho_{ij} = \pi_{ij}/\pi_{i+}$, siendo π_{i+} la probabilidad marginal de clasificar a un individuo en el nivel i de F_1 , independientemente de cuál sea su clasificación respecto del factor F_2 .

Generalizando a tablas de contingencia con k factores de clasificación, $F_1, F_2, \dots, F_k$, las fórmulas anteriores quedan:

Poisson: nada fijo

$$Y_{ij\dots k} \sim Po(\mu_{ij\dots k}) \rightsquigarrow E(Y_{ij\dots k}) = \mu_{ij\dots k}$$

con

$$Pr(Y_{ij\dots k} = y_{ij\dots k}) = \frac{\mu_{ij\dots k}^{y_{ij\dots k}} e^{-\mu_{ij\dots k}}}{y_{ij\dots k}!}.$$

– Verosimilitud:

$$L(\mu; \mathbf{y}) = \prod_{i=1}^I \prod_{j=1}^J \dots \prod_{k=1}^K \frac{\mu_{ij\dots k}^{y_{ij\dots k}} e^{-\mu_{ij\dots k}}}{y_{ij\dots k}!}.$$

– Log-verosimilitud:

$$l(\mu; \mathbf{y}) = cte + \sum_{i=1}^I \sum_{j=1}^J \dots \sum_{k=1}^K y_{ij\dots k} \log[E(Y_{ij\dots k})] - E(Y_{ij\dots k}). \quad (4.4)$$

Multinomial: tamaño muestral total fijo.

$$(Y_{11\dots 1}, \dots, Y_{IJ\dots K}) \sim MN(n, (\pi_{11\dots 1}, \dots, \pi_{IJ\dots K}))$$

$$\rightsquigarrow E(Y_{ij\dots k}) = n\pi_{ij\dots k}.$$

– Verosimilitud:

$$L(\pi; \mathbf{y}) = \frac{n!}{\prod_{i=1}^I \prod_{j=1}^J \dots \prod_{k=1}^K y_{ij\dots k}!} \prod_{i=1}^I \prod_{j=1}^J \dots \prod_{k=1}^K \pi_{ij\dots k}^{y_{ij\dots k}}.$$

con $\pi_{ij\dots k}$ la probabilidad de clasificar a un individuo en la celda $ij\dots k$ de la tabla.

– Log-verosimilitud:

$$l(\pi; \mathbf{y}) = cte + \sum_{i=1}^I \sum_{j=1}^J \dots \sum_{k=1}^K y_{ij\dots k} \log[E(Y_{ij\dots k})]. \quad (4.5)$$

Producto multinomial: algún total marginal fijo; por ejemplo, si se fijan los totales $y_{++\dots k}$ del factor F_k para sus K categorías de clasificación, y $\rho_{ij\dots k}$ es la probabilidad condicional de que un individuo que está en el nivel k de F_k sea también clasificado en los niveles i de F_1 , j de F_2 , etc., esto es,

$$\rho_{ij\dots k} = \frac{\pi_{ij\dots k}}{\pi_{++\dots k}},$$

con $\pi_{++\dots k}$ la probabilidad marginal de que uno de n individuos esté en el nivel k de F_k , entonces tenemos K distribuciones multinomiales independientes:

$$(Y_{11\dots k}, \dots, Y_{IJ\dots k}) \sim MN(y_{++\dots k}, (\pi_{11\dots k}, \dots, \pi_{IJ\dots k}))$$

$$\rightsquigarrow E(Y_{ij\dots k}) = y_{++\dots k} \rho_{ij\dots k}, \quad \text{para } k = 1, \dots, K.$$

– Verosimilitud:

$$L(\pi, \mathbf{y}) = \prod_{k=1}^K \left[\frac{y_{++\dots k}!}{\prod_{i=1}^I \prod_{j=1}^J \dots y_{ij\dots k}!} \prod_{i=1}^I \prod_{j=1}^J \dots \rho_{ij\dots k}^{y_{ij\dots k}} \right].$$

– Log-verosimilitud:

$$l(\mu; \mathbf{y}) = cte + \sum_{i=1}^I \sum_{j=1}^J \dots \sum_{k=1}^K y_{ij\dots k} \log[E(Y_{ij\dots k})].$$

Resulta pues, que la forma que tenemos para la log-verosimilitud es similar en estos tres casos, muestreo Poisson (P), Multinomial (M) o Producto Multinomial (PM). Si C es el número total de celdas en la tabla y $E(Y_i) = \mu_i$ la frecuencia esperada en la celda $i \in \{1, \dots, C\}$, la log-verosimilitud para una tabla de contingencia tiene la forma:

$$\begin{aligned} \text{(P)} \quad l(\mu; \mathbf{y}) &= cte + \sum_{i=1}^C y_i \log(\mu_i) - \mu_i \\ \text{(M) y (PM)} \quad l(\mu; \mathbf{y}) &= cte + \sum_{i=1}^C y_i \log(\mu_i). \end{aligned} \quad (4.6)$$

Esto es, depende únicamente de las frecuencias observadas en las celdas, y_i y de sus valores esperados $E(Y_i)$. Así, la modelización y el ajuste con GLM's de las tablas de contingencia será único para los tres tipos de datos. Cuando fijemos algunos totales, serán precisas ciertas consideraciones sobre el modelo (Sección 4.5).

4.3. Modelos Log-Lineales

Las cuestiones básicas de interés en el análisis de tablas de contingencia son detectar asociación entre los factores de clasificación y predecir las frecuencias medias

observadas. Veamos cómo se pueden resolver estas cuestiones a partir de la modelización log-lineal de las frecuencias esperadas. Consideramos, para presentar estos modelos, únicamente dos vías (factores) de clasificación, F_1 y F_2 .

- En **datos Poisson**, la hipótesis de independencia entre los factores de clasificación nos llevaría a establecer que el número de individuos que esperamos clasificar en la celda (i, j) se calcula a través del producto de los sucesos independientes “número esperado de individuos clasificados en el nivel i de F_1 ”, $E(Y_{i+}) = \mu_i$, y “número esperado de individuos clasificados en el nivel j de F_2 ”, $E(Y_{+j}) = \mu_j$, esto es,

$$\text{independencia} \equiv E(Y_{ij}) = \mu_i \mu_j.$$

Asociación equivaldrá a no-independencia entre ambos sucesos, $E(Y_{ij}) \neq \mu_i \mu_j$.

El efecto multiplicativo de los factores para predecir la frecuencia observada da lugar, de modo natural, a modelos logarítmicos aditivos (modelos log-lineales) de la forma:

$$\begin{aligned} H_0 : \log[E(Y_{ij})] &= \mu + \alpha_i + \beta_j \\ H_1 : \log[E(Y_{ij})] &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \end{aligned} \quad (4.7)$$

donde α_i representa el efecto principal del factor fila F_1 , β_j el correspondiente al factor columna F_2 y $(\alpha\beta)_{ij}$ el efecto de interacción entre ambos.

Contrastar independencia versus asociación será equivalente a contrastar interacción nula versus no nula en el modelo log lineal (4.7).

Puesto que en un modelo con interacción estimamos $1 + (I - 1) + (J - 1) + (I - 1)(J - 1) = IJ$ parámetros independientes, dicho modelo resulta ser el modelo saturado (tantos datos –celdas– como parámetros a estimar). Así, para contrastar asociación entre dos únicos factores hemos de ajustar el modelo con sólo efectos principales propuesto en H_0 y calcular la bondad del ajuste (que es, al fin y al cabo, compararlo con el modelo saturado –el que considera interacción no nula).

- En **datos multinomiales**, con un total n fijo, la hipótesis de independencia entre los factores significará que la probabilidad conjunta de clasificar a un individuo en el nivel i de F_1 y en el j de F_2 será igual al producto de la probabilidad marginal de clasificarlo según F_1 en el nivel i , π_{i+} , y la de clasificarlo según F_2 en el nivel j , π_{+j} , esto es,

$$\text{independencia} \equiv \pi_{ij} = \pi_{i+} \pi_{+j} \Rightarrow E(Y_{ij}) = n \pi_{i+} \pi_{+j}.$$

De ahí surge de forma natural un modelo log-lineal con sólo efectos principales. La asociación entre los factores, $E(Y_{ij}) = n\pi_{ij}$, implicará interacción no nula en dicho modelo:

$$\begin{aligned} H_0 : \log[E(Y_{ij})] &= \mu + \alpha_i + \beta_j, \\ H_1 : \log[E(Y_{ij})] &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}. \end{aligned} \quad (4.8)$$

• En **datos producto multinomial**, si los totales fila $y_{1+}, \dots, y_{I+}$ son fijos, la hipótesis de independencia implica que las probabilidades de las celdas son similares en todas las filas, esto es,

$$\text{independencia} \equiv \rho_{ij} = \rho_{+j}, \forall i \Rightarrow E(Y_{ij}) = y_{i+}\rho_{+j},$$

con lo que, de nuevo, la hipótesis de interés se puede expresar como un modelo multiplicativo en el que las frecuencias esperadas en las celdas vienen dadas por productos de probabilidades marginales y frecuencias marginales totales fijas. Tomando logaritmos, el contraste de independencia se convierte en una comparación de modelos aditivos –log-lineales– como en (4.8).

En conclusión, reconocer asociación entre los factores de clasificación de una tabla de contingencia es equivalente a rechazar interacción nula entre ellos en el modelo log-lineal con el que se predicen las frecuencias observadas en las celdas.

4.4. Estimación máximo-verosímil. Bondad de Ajuste

Dada la similaridad de las log-verosimilitudes para los tres tipos de muestreo mencionados, (4.6), y su dependencia exclusiva de las frecuencias observadas y esperadas, el ajuste se realizará a través de un modelo de respuesta Poisson con link logaritmo, y la estimación se obtendrá de la forma habitual por máxima verosimilitud (por mínimos cuadrados ponderados iterados, en (2.15)).

Considerando C el número total de celdas en la tabla y $E(Y_i) = \mu_i, i = 1, \dots, C$, en los tres tipos de datos (P, M y PM) el modelo log-lineal saturado proporciona como estimaciones de las frecuencias en las celdas las propias frecuencias observadas

$y_1, \dots, y_C$. La log-verosimilitud para el modelo saturado es de la forma:

$$l(\hat{\beta}_{sat}, \mathbf{y}) = cte + \sum_{i=1}^C y_i \log(y_i).$$

En el modelo Poisson, la deviance toma la forma:

$$D = 2 \sum_{i=1}^C [y_i \log(y_i / \hat{\mu}_i) - (y_i - \hat{\mu}_i)], \quad (4.9)$$

mientras que para los otros dos tipos de muestreo resulta:

$$D = 2 \sum_{i=1}^C y_i \log(y_i / \hat{\mu}_i), \quad (4.10)$$

Para muestras grandes, si el modelo ajusta bien los datos, D sigue una distribución chi-cuadrado con los grados de libertad dados por el número de celdas con frecuencias observadas distintas de cero (C si todas las celdas tienen frecuencias mayores que cero) menos el número de parámetros independientes (distintos de cero) estimados en el modelo.

El estadístico chi-cuadrado es más común que D para analizar tablas de contingencia,

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \quad (4.11)$$

donde $O_i = y_i$ y $E_i = E(Y_i) = \mu_i$. Ambos estadísticos son asintóticamente equivalentes y por lo tanto, para muestras grandes χ^2 tiene una distribución con los mismos grados de libertad que D .

La forma del estadístico chi-cuadrado sugiere que los residuos estandarizados para cada celda se puedan definir como:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}. \quad (4.12)$$

Estos residuos se corresponden con los residuos de Pearson para datos Poisson.

La adecuación del modelo se puede investigar a través de los residuos de la forma habitual. Residuos muy alejados del cero ($|r_i| > 3$) o la observación de ciertos patrones o tendencias de comportamiento no aleatorio podrían sugerir otros modelos más apropiados.

4.5. Consideraciones en el análisis de Tablas de Contingencia

1. En una tabla de contingencia es importante identificar la(s) variable(s) a explicar, es decir, aquellas en las que hay azar y por tanto probabilidad de ocurrencia a predecir, y las explicativas, que sirven para predecir la primera. Las preguntas a resolver en el análisis y la forma de plantear el modelo dependerán de esta caracterización. Si todas las variables se interpretan como variables a explicar, interesará concluir sobre entre cuáles existe asociación. Si sólo alguna lo es, interesará predecirla en función de los restantes factores de clasificación.
2. En los modelos log-lineales, los términos de orden más alto surgen tras la inclusión de términos de orden más bajo. Esto significa que al ajustar un modelo, si consideramos un término de interacción entre dos factores, necesariamente han de aparecer los efectos principales de ambos factores, aun no siendo individualmente significativos.
3. En el análisis de datos en tablas de contingencia, las cuestiones principales generalmente se refieren a investigar si hay asociaciones entre variables, es decir, el objetivo es contrastar la significatividad de las interacciones. Esto significa que los modelos que utilicemos para el contraste de asociación (independencia) han de involucrar términos de interacción y todos los efectos principales correspondientes.
4. Como en los modelos de Anova, al surgir tantos parámetros a estimar, hemos de considerar condiciones de identificabilidad. Así, si F_1 tiene I niveles, las diferencias entre ellos sólo nos permitirán estimar $I - 1$ parámetros independientes; si además F_2 tiene J niveles, sólo podremos estimar $(I - 1)(J - 1)$ parámetros independientes correspondientes al término de interacción $F_1 : F_2$.
5. A pesar de todos los supuestos distintos de muestreo que dan lugar a las tablas de contingencia, tanto el muestreo multinomial (tamaño total fijo) como el producto-multinomial (algún margen fijo) son equivalentes al muestreo Poisson independiente (nada fijo) respecto a los estimadores máximo verosímiles y a los estadísticos de contraste bajo una componente sistemática log-lineal, siempre, por supuesto, que se respeten ciertas condiciones.

Estas condiciones son que, si algunos totales marginales se asumen fijos, entonces los términos correspondientes a esos márgenes han de ser incluidos en

el predictor lineal. En los modelos multinomial y producto-multinomial, ciertos términos (tamaños muestrales respecto a algunos factores) son constantes fijas. Esto significa que los parámetros correspondientes han de ser incluidos SIEMPRE en cualquier modelo log-lineal que se proponga.

Así, cuando se asume tamaño muestral total fijo siempre habremos de incluir un efecto global μ . Si el margen de un factor (tamaño muestral en cada una de las categorías del factor) se asume fijo los términos que representan el efecto principal de ese factor, han de incluirse siempre en cualquier modelo log-lineal. Si asumimos fijos los márgenes de dos vías (F1 y F2), entonces TODOS los términos que representen a los efectos principales F1 y F2, así como el término de interacción (F1:F2), han de ser incluidos en el predictor lineal.

Es decir, siempre habrá un modelo minimal intrínseco a cada marco muestral, desde el cual hemos de llegar a “nuestro mejor modelo de predicción”.

6. Otras cuestiones de interés a resolver sobre los datos de tablas de contingencia, aparte de conclusiones sobre asociación, pueden ser:
 - ¿qué efectos son importantes para explicar las frecuencias que se dan en la muestra observada?
 - ¿cuál es la magnitud de los efectos importantes?
 - ¿cuál es el modelo más parsimonioso (simple) que proporciona un buen ajuste de los datos observados?

Todas estas cuestiones habrán de resolverse a través de contrastes sobre términos específicos, interacciones, análisis de deviance, inspección de medidas diagnósticas como los residuos, etc.

7. A veces, en una tabla de contingencia hay algunas celdas en las que las frecuencias observadas son cero, $y_{ij} = 0$. Estos ceros pueden ser de dos tipos: *ceros muestrales* y *ceros estructurales*. Un cero muestral surge cuando la frecuencia de esa clase es pequeña y simplemente no la hemos observado en la muestra; estos ceros son perfectamente aceptables y deberán ser incluidos en el análisis estándar. Un cero estructural, por el contrario, surge cuando esa clase es imposible, de forma que su probabilidad es cero; tales celdas deberán excluirse enteramente del análisis.

Ejemplo 4.4 (Continuación del Ejemplo 4.1) En los datos del melanoma (Tabla 4.1), simplemente interesaba estudiar la asociación entre los dos factores de clasificación: tipo de tumor (tipo) y localización (local).

Creamos variables con las frecuencias e identificación de grupos:

```
frec <- c(22, 2, 10, 16, 54, 115, 19, 33, 73, 11, 17, 28)
tipo <- -gl(4, 3, 12, labels = c("hutchinson", "superficial",
                                "nodular", "indeterminado"))
local <- -gl(3, 1, 12, labels = c("cabeza/cuello", "tronco",
                                   "extremidades"))
```

El resumen de los datos se presenta en la Tabla 4.4 en términos de proporciones observadas, de la que ya se pueden extraer conclusiones previas sobre qué localizaciones son más y menos frecuentes para cada tipo de tumor. Así Hutchinson es más frecuente en la cabeza/cuello, el melanoma superficial y el nodular en extremidades; en el tronco son más frecuentes los melanomas superficial y nodular.

Tipo	Localización			Total Fila	
	cabeza/cuello	tronco	extremidades		
Hutchinson	22	2	10	34	N
	0.647	0.059	0.294	0.085	N/RowTotal
	0.324	0.019	0.044		N/ColTotal
	0.055	0.005	0.025		N/Total
superficial	16	54	115	185	N
	0.086	0.292	0.622	0.462	N/RowTotal
	0.235	0.509	0.509		N/ColTotal
	0.040	0.135	0.288		N/Total
nodular	19	33	73	125	N
	0.152	0.264	0.584	0.312	N/RowTotal
	0.279	0.311	0.323		N/ColTotal
	0.048	0.082	0.182		N/Total
indeterm.	11	17	28	56	N
	0.196	0.304	0.500	0.140	N/RowTotal
	0.162	0.160	0.124		N/ColTotal
	0.028	0.042	0.070		N/Total
ColTotl	68	106	226	400	
	0.17	0.26	0.56		

Tabla 4.4: Datos sobre el Melanoma. Resumen de la Tabla 4.1.

El tipo de muestreo es multinomial, pues el tamaño total $n = 400$ fue fijado previamente.

Identificando el subíndice i con el factor $F_1 = \text{tipo}$ ($i = 1, 2, 3, 4$), y j con $F_2 = \text{local}$ ($j = 1, 2, 3$), e Y_{ij} con el número de individuos con un tumor de tipo i en la localización j , el contraste sobre si existe o no asociación entre el tipo y la localización de los tumores sería:

$$H_0 : E(Y_{ij}) = n\pi_{i+}\pi_{+j} \Leftrightarrow \log[E(Y_{ij})] = \mu + \alpha_i + \beta_j$$

$$H_1 : E(Y_{ij}) = n\pi_{ij} \Leftrightarrow \log[E(Y_{ij})] = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

Para resolverlo no hay más que ajustar el modelo de H_0 ,

$$\text{fit} < -\text{glm}(\text{frec} \sim \text{tipo} + \text{local}, \text{family} = \text{poisson})$$

y contrastar la bondad del ajuste (comparación con el modelo saturado $\equiv$ interacción no nula),

$$1 - \text{pchisq}(\text{fit\$deviance}, \text{fit\$df.residual})$$

que da un p -valor asociado de $2,050453 \cdot 10^{-9}$, con lo cual se rechaza independencia a favor de asociación entre los factores tipo y localización de tumores. Efectivamente, determinados tipos de tumor se presentan en ciertas localizaciones con más frecuencia que en otras.

Resultado 4.5.1 Interacciones e Independencia con tres factores de clasificación. Cuando se tienen tres factores de clasificación, la interpretación de las interacciones de orden 2 y 3 en un modelo log-lineal viene descrita a continuación.

Está claro que el modelo saturado es:

$$\eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}.$$

Ahora bien,

- Falta $(\alpha\beta\gamma)_{ijk} \rightsquigarrow$ asociación parcial entre cada par de factores.
- Faltan $(\alpha\beta\gamma)_{ijk}$ y $(\alpha\beta)_{ij} \rightsquigarrow F_1$ y F_2 son independientes en cada nivel de F_3 , pero ambas están asociadas con F_3 ; son condicionalmente independientes.
- Faltan $(\alpha\beta\gamma)_{ijk}$, $(\alpha\beta)_{ij}$ y $(\alpha\gamma)_{ik} \rightsquigarrow F_2$ y F_3 están asociadas, pero F_1 es completamente independiente de F_2 y F_3 .
- Faltan $(\alpha\beta\gamma)_{ijk}$, $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ y $(\beta\gamma)_{jk} \rightsquigarrow F_1$, F_2 y F_3 son completamente independientes entre sí.

Ejemplo 4.5 (Continuación del Ejemplo 4.3) Para los datos de la aspirina (Tabla 4.3) hemos de proponer un modelo de clasificación con tres vías. Las variables de clasificación involucradas son:

$A = \text{aspirina} \rightsquigarrow \text{no usa}(i = 1), \text{ sí usa}(i = 2)$
 $L = \text{local} \rightsquigarrow \text{duodenal}(j = 1), \text{ gástrica}(j = 2)$
 $U = \text{úlcera} \rightsquigarrow \text{no}(k = 1), \text{ sí}(k = 2)$

Por supuesto, además hemos de definir la variable que contiene las frecuencias de las celdas Y_{ijk} ; a dicha variable la llamamos 'frec'.

En el estudio se controlaron las variables úlcera y localización. A los pacientes clasificados en cada una de las combinaciones de estas dos variables se les preguntó sobre el consumo de aspirina, que es la variable aleatoria y por tanto la variable a explicar en función de las otras dos. Se trata pues de un muestreo producto-multinomial con los totales marginales fijados para los factores L y U , $y_{+11} = 64$, $y_{+12} = 68$, $y_{+21} = 57$, $y_{+22} = 61$. Así, ambos factores habrán de estar incluidos

en cualquier modelo log-lineal que propongamos para predecir las frecuencias de las celdas, esto es,

$$E(Y_{ijk}) = y_{+jk} \cdot \rho_{ijk} \Leftrightarrow \log[E(Y_{ijk})] = \theta_{jk} + \varepsilon_{ijk}.$$

El objetivo básico es buscar relaciones entre el consumo de aspirina y la existencia de úlcera péptica. En consecuencia, los parámetros de interés son las probabilidades condicionadas:

$$Pr(A = i | L = j, K = k) = \rho_{ijk}.$$

El predictor lineal para los modelos log-lineales minimal y saturado viene dado por:

$$\begin{aligned} \text{minimal: } \eta_{ijk} &= \underbrace{\mu + \beta_j^L + \gamma_k^U + (\beta\gamma)_{jk}^{LU}}_{\theta_{jk}^{LU}} \\ \text{saturado: } \eta_{ijk} &= \mu + \beta_j^L + \gamma_k^U + (\beta\gamma)_{jk}^{LU} + \\ &\quad + \underbrace{\alpha_i^A + (\alpha\beta)_{ij}^{AL} + (\alpha\gamma)_{ik}^{AU} + (\alpha\beta\gamma)_{ijk}^{ALU}}_{\varepsilon_{ijk}^{ALU}}, \end{aligned}$$

con $\eta_{ijk} = \log[E(Y_{ijk})]$ y donde ε_{ijk}^{ALU} explica la relación entre el consumo de aspirina y los factores controlados en la investigación (localización y úlcera).

Con el fin de dejar más clara la identificación de los efectos del modelo (la variable de la que proceden) hemos incluido como superíndices las letras A, L y U correspondientes a cada uno de los factores de clasificación involucrados en la Tabla de datos. Esta notación la utilizaremos alguna otra vez.

Cuestiones de interés:

1. En el grupo gástrico ($j = 2$), ¿la úlcera está asociada con el consumo de aspirina?
2. En el grupo duodenal ($j = 1$), ¿la úlcera está asociada con el consumo de aspirina?

CONTESTACIÓN: Para ambas cuestiones estamos planteando si, dado j fijo, el consumo de aspirina (A) está relacionado con la existencia de úlcera (U); la hipótesis de independencia/asociación entre aspirina y úlcera vendrá dada,

para un j fijo, por:

$$\begin{aligned} \text{indep } H_0 : Pr(A = i | L = j, U = k) &= Pr(A = i | L = j) \\ &\Leftrightarrow \log[E(Y_{ijk})] = \theta_{jk}^{LU} + \alpha_i^A + (\alpha\beta)_{ij}^{AL} \\ \text{asoc } H_1 : Pr(A = i | L = j, U = k) &= Pr(A = i | L = j, U = k) \\ &\Leftrightarrow \log[E(Y_{ijk})] = \theta_{jk}^{LU} + \alpha_i^A + (\alpha\beta)_{ij}^{AL} + (\alpha\gamma)_{ik}^{AU} + (\alpha\beta\gamma)_{ijk}^{ALU}. \end{aligned}$$

Para realizar este contraste, comparamos el modelo que define H_0 con el saturado (H_1), es decir, testamos la bondad del ajuste de fit1 , dado por:

$$\boxed{\text{fit1} < -\text{glm}(\text{frec} \sim U * L + A + A : L, \text{family} = \text{poisson})}$$

Este modelo tiene una deviance de $D = 17,697$ para una chi-cuadrado con 2 g.l., lo que da un p -valor de $p = 0,00014$. Claramente se rechaza independencia entre el consumo de aspirina y la existencia de úlcera a favor de asociación, tanto en el grupo de pacientes con úlcera gástrica como en el de úlcera duodenal.

3. Una vez reconocida como significativa la asociación entre aspirina y el desarrollo de úlcera, cabe preguntarse si dicha relación es similar en los dos grupos de pacientes: gástricos y duodenales.

CONTESTACIÓN: Que fuera similar significaría que

$$Pr(A = i | L = j, U = k) = Pr(A = i | L = j', U = k) \text{ para } j \neq j' \in \{1, 2\},$$

esto es,

$$Pr(A = i | L = 1, U = k) = Pr(A = i | L = 2, U = k),$$

es decir, que $\rho_{ijk} = \rho_{i \cdot k}$, $\forall i, k$, y por lo tanto el contraste a resolver sería:

$$\begin{aligned} \text{asoc. similar } H_0 : \log[E(Y_{ijk})] &= \theta_{jk}^{LU} + \alpha_i^A + (\alpha\gamma)_{ik}^{AU} \\ \text{asoc. distinta } H_1 : \log[E(Y_{ijk})] &= \theta_{jk}^{LU} + \varepsilon_{ijk}^{ALU}. \end{aligned}$$

Para resolver el contraste, basta con ajustar y contrastar la bondad del ajuste:

$$\boxed{\text{fit2} < -\text{glm}(\text{frec} \sim U * L + A + A : U, \text{family} = \text{poisson})}$$

que, con una deviance de $D = 10,538$ para una chi-cuadrado con 2 g.l., da un p -valor $p = 0,0051$, que implica el rechazo de similaridad (H_0), a favor de la hipótesis de que el consumo de aspirina es diferente en pacientes con úlcera gástrica que en los de úlcera duodenal.

4.6. Modelos Logit–Multinomial

Si en una tabla de contingencia identificamos uno de los factores de clasificación como variable a explicar y el resto como explicativas, sin duda, más que simplemente concluir que existe asociación entre ellos, interesará modelizar los datos para predecir la probabilidad de que se dé una u otra respuesta en la variable a explicar en función de las distintas combinaciones en las variables explicativas. Existe un conjunto de modelos log-lineales, útiles para investigar asociación, que son equivalentes a modelos logit, útiles para predecir una variable a través de otras; son los **modelos logit–multinomial**.

Consideremos una tabla de contingencia en la que están involucradas como variables de clasificación, una variable a explicar R con I niveles de clasificación, y dos variables explicativas A y B , con J y K niveles de clasificación respectivamente. Denotemos por $\pi_{i|jk}$ las probabilidades de respuesta $R = i$ dada cierta clasificación en las variables explicativas, $(A = j, B = k)$, esto es,

$$\pi_{i|jk} = Pr(R = i | A = j, B = k).$$

Los modelos logit–multinomial vendrán definidos en términos de los *log-odds* (logaritmo del cociente) de las probabilidades condicionadas de dos niveles de respuesta R , dadas las diferentes combinaciones de niveles de las variables explicativas. Al predecir cocientes de probabilidades, se suele elegir uno de los niveles de la variable a explicar como nivel de referencia, generalmente el primero o el último. Si se toma como categoría de referencia la última de la variable a explicar R , los *log-odds* a predecir serán:

$$\log \left(\frac{\pi_{i|jk}}{\pi_{I|jk}} \right).$$

Si el interés es, sin embargo, comparar categorías adyacentes, los *log-odds* a predecir serán:

$$\log \left(\frac{\pi_{i+1|jk}}{\pi_{i|jk}} \right).$$

El ajuste de estos modelos logit–multinomiales se realiza a través de modelos log–lineales equivalentes (con verosimilitud Poisson) con los que se predicen las frecuencias observadas Y_{ijk} en la tabla de contingencia. Es preciso puntualizar que la equivalencia entre estos modelos logit y los log–lineales es cierta asumiendo que

los totales marginales correspondientes a las variables explicativas son fijos. Básicamente es decir que las variables explicativas ha sido controladas en el estudio con el fin principal de predecir la variable a explicar R . Al asumir esto, obligamos a que los efectos asociados a dichas variables explicativas hayan de aparecer siempre (junto con sus interacciones) en cualquier modelo log–lineal que ajustemos.

Ejemplo 4.6 (Aspiraciones universitarias, estatus social e incentivos familiares)

En la Tabla 4.5 se muestran una parte de los resultados de un estudio en los Estados Unidos con el que se pretendía investigar el grado de asociación entre las aspiraciones de los estudiantes de Bachillerato por proseguir con estudios universitarios y su entorno social, medido en términos del estatus socio–económico de su familia y el hecho de que recibieran apoyo en su familia para continuar estudiando (Fienberg, 1978, pag.101)

Interpretando ‘aspiraciones’ (A) como la variable a explicar en función de ‘estatus’ (E) y ‘motivación’ (M), queremos predecir la proporción de individuos que pretenden continuar en la universidad en función del estatus económico de la familia y el apoyo que reciben en ella, o lo que es lo mismo, los log–odds:

$$\log \left[\frac{Pr(A = sí|E = j, M = k)}{Pr(A = no|E = j, M = k)} \right].$$

Estatus Social	Motivación familiar	Aspiraciones Universitarias		Total
		No	Sí	
Baja	Baja	749	35	784
	Alta	233	133	366
Media-Baja	Baja	627	38	665
	Alta	330	303	633
Media-Alta	Baja	420	37	457
	Alta	374	467	841
Alta	Baja	153	26	179
	Alta	266	800	1066
Total		3152	1938	4991

Tabla 4.5: Aspiraciones universitarias, estatus social e incentivos familiares

Ejemplo 4.7 (Contraceptivos) *En la Tabla 4.6 (Rodríguez, 1998) aparecen los datos de un estudio sobre distintos usos contraceptivos en diferentes grupos generacionales.*

Entendiendo la variable 'método contraceptivo' (M) como variable a explicar y la variable 'edad' (E) como explicativa, una cuestión de interés puede ser predecir el uso anticonceptivo de una mujer en función de su edad. Esto lo podemos hacer comparando la proporción de mujeres que usa cada uno de los usos anticonceptivos con la de aquellas que no usan ninguno, esto es, a través de los log-odds $\log(\pi_{i|j}/\pi_{3|j})$, para $i = 1, 2$ (esterilización y otros métodos), y $j = 1, 2, \dots, 7$ (grupos de edad).

Edad	Método contraceptivo			Total
	Esterilización	Otros	Ninguno	
15-19	3	61	232	296
20-24	80	137	400	617
25-29	216	131	301	648
30-34	268	76	203	547
35-39	197	50	188	435
40-44	150	24	164	338
45-49	91	10	183	284
Total	1005	489	1671	3165

Tabla 4.6: Usos contraceptivos y edades en mujeres casadas

Para presentar los modelos logit multinomial consideramos en primer lugar el caso más simple, en el que la variable a explicar R tiene únicamente dos niveles de clasificación, $I = 2$, esto es, se trata de una variable binaria. Posteriormente generalizamos el modelo logit al modelo logit-multinomial, para cuando R tiene más de dos niveles de respuesta.

4.6.1. Variable a explicar Binaria

Si R es una variable binaria (éxito/fracaso) y el propósito del análisis es predecir la probabilidad de éxito en función de las variables explicativas A y B controladas en el estudio, podemos plantear el modelo **logit**:

$$\log \left[\frac{E(Y_{1jk})}{E(Y_{2jk})} \right] = \log \left(\frac{\pi_{2|jk}}{\pi_{1|jk}} \right) = \lambda + \lambda_j^A + \lambda_k^B, \quad (4.13)$$

que es equivalente al modelo **log-lineal** dado por:

$$\log[E(Y_{ijk})] = \mu + \alpha_i^R + \beta_j^A + \gamma_k^B + (\alpha\beta)_{ij}^{RA} + (\alpha\gamma)_{ik}^{RB} + (\beta\gamma)_{jk}^{AB}, \quad (4.14)$$

donde:

$$\begin{aligned}\lambda &= \alpha_1^R - \alpha_2^R \\ \lambda_j^A &= (\alpha\beta)_{1j}^{RA} - (\alpha\beta)_{2j}^{RA} \\ \lambda_k^B &= (\alpha\gamma)_{1k}^{RB} - (\alpha\gamma)_{2k}^{RB},\end{aligned}\tag{4.15}$$

y añadimos superíndices a los efectos con el fin de identificar con claridad la(s) variable(s) de la(s) que provienen. Utilizaremos esta formulación en lo que queda del tema.

Para probarlo, basta con calcular, a partir de (4.14), las diferencias de logaritmos:

$$\log \left[\frac{E(Y_{1jk})}{E(Y_{2jk})} \right] = \log[E(Y_{1jk})] - \log[E(Y_{2jk})].$$

Por supuesto, a la hora de estimar los parámetros del modelo log-lineal será necesario imponer (como siempre) ciertas condiciones de identificabilidad, como por ejemplo, $\sum_i \alpha_i^R = 0$, $\sum_i (\alpha\beta)_{ij} = 0$ y $\sum_i (\alpha\gamma)_{ik} = 0$.

Ejemplo 4.8 (Continuación del Ejemplo 4.6) Consideramos los datos de la Tabla 4.5. Recordemos el objetivo del análisis: i) estudiar si existe asociación entre el hecho de que un estudiante pretenda proseguir sus estudios en la Universidad y el entorno social del que proviene, y si hay relación ii) predecir la proporción de individuos con aspiraciones de proseguir en la universidad, en función de su entorno.

La variable a explicar 'aspiraciones' (A) tiene dos niveles de respuesta: 'Sí' (éxito) y 'No' (fracaso). El entorno social viene descrito en términos de las variables explicativas 'motivación' (M) y 'estatus' (E), para las cuales asumimos los totales marginales fijos. Al asumir que se han controlado los dos factores explicativos, los efectos principales asociados a éstos y su interacción han de aparecer por defecto en cualquier modelo log-lineal que planteemos, esto es, los términos $M * E = M + E + M : E$. Para resolver el análisis planteamos un ajuste log-lineal equivalente a un logit, con el que predecir las probabilidades condicionadas de interés.

Obviamente, la forma de introducir los datos para realizar los dos análisis independientemente en $\mathcal{R}$ es distinta: en el logit, la variable a predecir es una matriz de dos columnas con los éxitos (aspiraciones=sí) y fracasos (aspiraciones=no); en

el log-lineal, es un vector con las frecuencias de todas las celdas. En función de cómo se introduce la respuesta se han de introducir los valores correspondientes de las variables explicativas; la longitud de los vectores que contienen los valores de las variables explicativas en el modelo log-lineal será doble a la de los del modelo logit. Aquí presentamos la modelización log-lineal. El estudiante puede realizar alternativamente la modelización logit y comparar:

Introducimos los datos en $\mathcal{R}$:

```

exitos <- c(35, 133, 38, 303, 37, 467, 26, 800)
fracasos <- c(759, 233, 627, 330, 420, 374, 153, 266)
frec <- c(exitos, fracasos)

M <- -gl(2, 1, 16, labels = c("baja", "alta"))
E <- -gl(4, 2, 16, labels = c("bajo", "medio-bajo", "medio-alto", "alto"))

```

y contestamos diversas preguntas:

1. Sea cual sea el estatus social del que provenga un individuo, ¿están relacionados la motivación familiar que recibe en casa el estudiante con sus aspiraciones de comenzar la universidad?

CONTESTACIÓN: basta contrastar la bondad del ajuste del modelo:

```
fit1 <- glm(frec ~ (M * E) + A + A : E, family = poisson)
```

que da un p-valor de 0, con lo concluimos a favor de asociación entre aspiraciones y motivaciones en cualquier estatus socio-económico.

2. La relación entre motivaciones y aspiraciones, ¿es la misma sea cual sea el estatus social en el que el individuo se desenvuelve?

CONTESTACIÓN: si no fuera la misma significaría que existe la interacción triple; así, para contestar la pregunta basta ajustar el modelo con todas las interacciones de orden 2 (H_0) y testar la bondad del ajuste:

$$\text{fit2} < -\text{glm}(\text{frec} \sim (M + E + A)^2, \text{family} = \text{poisson})$$

que da un p -valor de 0,66497, con lo que no rechazamos significativamente la hipótesis de que el efecto de la motivación en la proporción de individuos que aspiran a ir a la Universidad sea el mismo para cualquier estatus socio-económico.

En principio, asumiendo como válido el último modelo propuesto, fit2 ,

$$\log[E(Y_{ijk})] = \mu + \beta_j^M + \gamma_k^E + (\beta\gamma)_{jk}^{ME} + \alpha_i^A + (\alpha\beta)_{ij}^{AM} + (\alpha\gamma)_{ik}^{AE},$$

obtenemos las estimaciones de la Tabla siguiente:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.59850	0.05886	95.116	< 2e - 16
M.baja	-0.59471	0.09234	-6.441	1,19e - 10
E.bajo	-0.16542	0.08573	-1.930	0,05366
E.medio-alto	0.32331	0.07664	4.219	2,46e - 05
E.medio-bajo	0.20372	0.07841	2.598	0,00937
A.si	1.08107	0.06731	16.060	< 2e - 16
M.baja:E.bajo	1.78588	0.11444	15.606	< 2e - 16
M.baja:E.medio-alto	0.71532	0.11136	6.424	1,33e - 10
M.baja:E.medio-bajo	1.23178	0.10987	11.211	< 2e - 16
M.baja:A.si	-2.68292	0.09867	-27.191	< 2e - 16
E.bajo:A.si	-1.59311	0.11527	-13.820	< 2e - 16
E.medio-alto:A.si	-0.85460	0.09259	-9.230	< 2e - 16
E.medio-bajo:A.si	-1.17298	0.09803	-11.965	< 2e - 16

3. ¿Cuál es la probabilidad de que un individuo tenga planes universitarios si es de clase baja y recibe poca motivación en su familia para continuar con sus estudios?

CONTESTACIÓN: Se pide $\Pr(A = si | M = baja, E = bajo)$. Despejando del logit,

$$\begin{aligned} \text{logit}[\Pr(A = si | M, E)] &= \log \left[\frac{\Pr(A = si | M, E)}{\Pr(A = no | M, E)} \right] = \lambda + \lambda_j^M + \lambda_k^E \\ \Rightarrow \Pr(A = si | M, E) &= \frac{\exp\{\lambda + \lambda_j^M + \lambda_k^E\}}{1 + \exp\{\lambda + \lambda_j^M + \lambda_k^E\}}. \end{aligned}$$

Mirando la tabla anterior con las estimaciones de los parámetros, tenemos

$$\lambda = 1,08107, \quad \lambda_j^{M=baja} = -2,68292, \quad \lambda_k^{E=bajo} = -1,59311,$$

de donde

$$Pr(A = si|M = baja, E = bajo) = \frac{\exp\{1,08107 - 2,68292 - 1,59311\}}{1 + \exp\{1,08107 - 2,68292 - 1,59311\}} = 0,03935.$$

Concluimos pues, que son muy pocos (4%) los que tienen aspiraciones universitarias cuando provienen de un nivel socio-económico bajo y no reciben el apoyo de la familia.

4. Y en el colectivo de individuos de clase media-baja con motivaciones familiares altas, ¿cuál es la proporción estimada de estudiantes que aspiran a estudiar en la Universidad?

CONTESTACIÓN: Tenemos que estimar $Pr(A = si|M = alta, E = medio-bajo)$. De nuevo, mirando la tabla de estimaciones,

$$\lambda = 1,08107, \lambda_j^{M=alta} = 0, \lambda_k^{E=medio-bajo} = -1,17298,$$

obtenemos

$$Pr(A = si|M = baja, E = bajo) = \frac{\exp\{1,08107 - 1,17298\}}{1 + \exp\{1,08107 - 1,17298\}} = 0,4770.$$

En la clase media-baja, cuando los individuos reciben apoyo para estudiar, alrededor de un 50% aspiran a estudiar en la Universidad.

4.6.2. Variable a explicar Multinomial

Generalizando a la situación en que la variable a explicar R tiene más de dos categorías de clasificación, los modelos logit para el cociente éxito/fracaso pasan a convertirse en log-odds de comparación de cada nivel de respuesta frente a uno de referencia. Si consideramos como nivel de referencia el último de la variable R , I , el modelo **logit-multinomial** vendrá dado por:

$$\log\left(\frac{\pi_{i|jk}}{\pi_{I|jk}}\right) = \lambda_i + \lambda_{ij}^A + \lambda_{ik}^B, \quad (4.16)$$

que resultará equivalente al modelo log-lineal que predice las frecuencias en las celdas Y_{ijk} mediante:

$$\log[E(Y_{ijk})] = \mu + \alpha_i^R + \beta_j^A + \gamma_k^B + (\beta\gamma)_{jk}^{AB} + (\alpha\beta)_{ij}^{RA} + (\alpha\gamma)_{ik}^{RB}, \quad (4.17)$$

con

$$\begin{aligned}\lambda_i &= \alpha_i^R - \alpha_I^R \\ \lambda_{ij}^A &= (\alpha\beta)_{ij}^{RA} - (\alpha\beta)_{Ij}^{RA} \\ \lambda_{ik}^B &= (\alpha\gamma)_{ik}^{RB} - (\alpha\gamma)_{Ik}^{RB}.\end{aligned}$$

Los parámetros de este modelo proporcionan los parámetros para los logits de cualquier par de categorías respuesta, dado que

$$\log\left(\frac{\pi_{i|jk}}{\pi_{i'|jk}}\right) = \log\left(\frac{\pi_{i|jk}}{\pi_{I|jk}}\right) - \log\left(\frac{\pi_{i'|jk}}{\pi_{I|jk}}\right).$$

Ejemplo 4.9 (Continuación del Ejemplo 4.7) Recordemos los objetivos del análisis con los datos de contraceptivos en la Tabla 4.5: i) estudiar si existe asociación entre el método contraceptivo empleado y el grupo de edad de una población de mujeres casadas, y si existe, ii) predecir la proporción de mujeres que optan por un método concreto en función de su edad.

Una vez introducidos los datos en $\mathcal{R}$,

```
esteril <- -c(3, 80, 216, 268, 197, 150, 91)
otros <- -c(61, 137, 131, 76, 50, 24, 10)
ninguno <- -c(232, 400, 301, 203, 188, 164, 183)
frec <- -c(esteril, otros, ninguno) ~ (conteos de las celdas)
metodo <- -factor(rep(1 : 3, rep(7, 3)), levels = 1 : 3,
  labels = c("esteril", "otros", "ninguno"))
edad <- -factor(rep(1 : 7, 3), levels = 1 : 7,
  labels = c("15 - 19", "20 - 24", "25 - 29", "30 - 34",
    "35 - 39", "40 - 44", "45 - 49"))
```

contrastar independencia entre método y edad es equivalente a contrastar interacción nula, esto es, testar la bondad del ajuste del modelo

$$\boxed{\text{fit.ind} < - \text{glm}(\text{frec} \sim \text{metodo} + \text{edad}, \text{family} = \text{poisson})}$$

para el que se obtiene una deviance de $D = 521,1$ con 12 g.l., y un p -valor asociado de 0. Se rechaza pues, el modelo de independencia frente a algún tipo de asociación entre el método contraceptivo utilizado y la edad de la mujer.

Ahora bien, si utilizamos exclusivamente la información nominal de los factores de clasificación, la asociación sólo la podemos describir en términos de la interacción entre 'método' y 'edad', con lo que se agotan todos los grados de libertad disponibles y llegamos al modelo saturado (que no es un ajuste válido).

4.6.3. Modelos Logit Generalizados

El problema que encontramos en el Ejemplo anterior es relativamente frecuente. Cuando trabajamos con factores de clasificación que contienen información ordinal, cuantitativa de algún modo, es posible utilizarla para describir mejor la asociación entre factores sin llegar al modelo saturado. Veremos a continuación cómo explicar asociación entre factores cuando se dispone de información cuantitativa (ordinal), como es el caso en la variable 'edad' en el Ejemplo anterior.

Los modelos logit-multinomial para tablas de contingencia son una particularización de los modelos logit generalizados.

Los **modelos logit generalizados**, como los modelos de regresión usuales, pueden contener variables explicativas continuas o discretas. Denotemos por $\pi_{i|j} \equiv \pi_i(\mathbf{x}_j)$, para $i = 1, 2, \dots, I$ y $j = 1, 2, \dots, J$, la probabilidad de responder en el nivel i de la variable respuesta R , cuando tenemos la combinación de valores de p variables explicativas $\mathbf{x}_j = (x_{j1}, \dots, x_{jp})'$.

En términos de las probabilidades de respuesta, el modelo logit generalizado viene dado por:

$$\eta_{ij} = \log \frac{\pi_{i|j}}{\pi_{I|j}} = \alpha_i + \beta_i' \mathbf{x}_j, \quad (4.18)$$

con la restricción de identificabilidad $\beta_I = 0$. El modelo anterior 4.18 permite despejar las probabilidades condicionadas de interés como:

$$\pi_{i|j} = \frac{\exp(\eta_{ij})}{\sum_{k=1}^I \exp(\eta_{kj})}, \quad i = 1, \dots, I, \quad (4.19)$$

expresión que obtenemos tomando exponencial en (4.18), $\pi_{i|j} = \pi_{I|j} \exp(\eta_{ij})$, y teniendo en cuenta que $\sum_i \pi_{i|j} = 1$.

Para obtener estimaciones máximo-verosímiles, habríamos de maximizar la verosimilitud multinomial conjunta de todas las observaciones independientes, sujeta a la restricción que impone (4.19). Los estimadores tendrían distribuciones asintóticas normales, con errores estándar obtenidos a partir de la inversa de la matriz de información de Fisher. Al ser cóncava la log-verosimilitud, un procedimiento iterativo como el de Newton-Raphson proporcionaría las estimaciones.

Sin embargo, los modelos logit multinomiales también se pueden ajustar por máxima-verosimilitud utilizando un modelo log-lineal equivalente y la verosimilitud Poisson. Específicamente, si $\{Y_{ij}; i = 1, \dots, I; j = 1, \dots, J\}$ representan los conteos de variables Poisson en las celdas, tenemos el modelo log-lineal equivalente a (4.18), dado por:

$$\log[E(Y_{ij})] = \mu + \lambda_j^X + \lambda_i^R + \gamma_i' \mathbf{x}_j, \quad (4.20)$$

con las restricciones de identificabilidad usuales.

Comentar que λ_j^X corresponde al efecto de estar clasificando a un individuo en las condiciones experimentales definidas por $\mathbf{x}_j$, es decir, es el término de restricción que corresponde a fijar los totales marginales y_{+j} de las observaciones multinomiales y_{ij} . Dicho término asegura la reproducción exacta de los márgenes multinomiales y_{+j} . La equivalencia entre los modelos (4.18) y (4.20) se obtiene al tomar diferencias entre los logaritmos de las frecuencias medias en las categorías de respuesta $R = i$ y una de referencia, como por ejemplo la I ,

$$\log(\mu_{ij}) - \log(\mu_{Ij}) = \underbrace{(\lambda_i^R - \lambda_I^R)}_{\alpha_i} + \underbrace{(\gamma_i - \gamma_I)' \mathbf{x}_j}_{\beta_i'} \quad (4.21)$$

Además, el modelo log-lineal incluye un parámetro λ_i^R para cada categoría de la variable a explicar R , lo que permite que los conteos esperados varíen de una a otra. Por último, el modelo incluye términos de interacción $\gamma_i' \mathbf{x}_j$ para captar los efectos de las covariables $\mathbf{x}_j$ sobre los *log-odds* de la variable a explicar.

En general, para convertir un modelo logit multinomial en uno Poisson log-lineal, consideremos una variable R , a explicar en términos de factores $F_1, F_2, \dots$ y covariables continuas $X_1, X_2, \dots$. Se ha recogido un conjunto de observaciones en R ,

tomadas bajo k conjuntos de condiciones experimentales distintas, bajo las cuales estos factores y covariables continuas varían. Supongamos el modelo logit-multinomial definido por:

$$M(F_1, F_2, \dots, X_1, X_2, \dots).$$

Supongamos un factor N que identifica cada conjunto de condiciones experimentales bajo las que se ha observado la respuesta en R . Entonces el modelo Poisson log-lineal con fórmula:

$$R * M + N$$

es equivalente al modelo logit multinomial con fórmula M . Las estimaciones de los parámetros en el modelo multinomial aparecerán como interacciones con R . Las estimaciones de los parámetros asociadas con N son completamente irrelevantes para la inferencia en el modelo multinomial, y generalmente serán muchas. De hecho, N es la interacción de todos los factores y covariables continuas categorizadas, excluyendo el factor a explicar R .

Ejemplo 4.10 (Continuación del Ejemplo 4.9) *Una vez reconocida la asociación entre el método contraceptivo empleado y la edad de la mujer, pasamos al segundo objetivo: predecir la proporción de mujeres que usan cada uno de los métodos contraceptivos en función del grupo de edad al que pertenece. Para ello vamos a emplear la información cuantitativa de la variable edad, contenida en las marcas de clase de los intervalos que definen los grupos de edad. Puesto que trabajamos con log-odds, tomamos como categoría de referencia la del grupo de mujeres que no usan ningún método contraceptivo, ($I = 3$). Si dibujamos los logits empíricos, definidos por $\log((y_{ij} + 1/2)/(y_{3j} + 1/2))$ para $i = 1, 2$ y $j = 1, \dots, 7$, versus la variable 'edad.x', definida con las marcas de clase de los intervalos (ver Figura 4.1), apreciamos cierta tendencia cuadrática. Planteamos pues, un modelo logit-multinomial cuadrático para describir la asociación método–edad:*

$$\log \left[\frac{\pi_{i|j}}{\pi_{I|j}} \right] = \eta_{ij} = \lambda_i + \delta_i x_j + \omega_i x_j^2,$$

donde x_j es la marca de clase para el nivel j de la variable 'edad', e $i = 1, 2$ corresponde a los grupos de mujeres que usan algún tipo de método contraceptivo: (E) esterilización u (O) otros.

El modelo Poisson log-lineal equivalente al anterior será:

$$\log[E(Y_{ij})] = \mu + \theta_j^E + \alpha_i^M + \beta_i^M x_j + \gamma_i^M x_j^2,$$

donde θ_j^E representa el efecto de estar clasificando a los individuos en el nivel j del factor explicativo (nominal) 'edad' (E) y α_i^M el correspondiente al nivel i del factor 'método' (M). Cada una de las condiciones de experimentación (grupo de

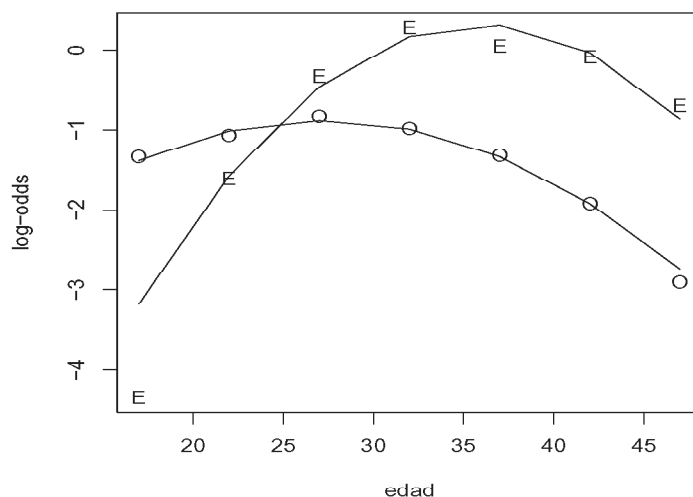


Figura 4.1: Relación entre el uso de contraceptivos y la edad en mujeres casadas.

edad) corresponde a un valor de la covariable continua que definimos a través de las marcas de clase de la edad, de ahí que al ajustar el modelo estamos considerando a la variable edad, por un lado como factor nominal (con efecto θ_j), y por otro lado como covariable numérica, con el valor correspondiente x_j .

Procedemos al ajuste en $\mathcal{R}$. La variable 'edad.x' con las marcas de 'edad' es:

```
edad.x <- rep(c(17, 22, 27, 32, 37, 42, 47), 3)
```

El ajuste del modelo log-lineal cuadrático viene dado por:


```
fit.aso <- glm(frec ~ edad + metodo * (edad.x + I(edad.x^2)),
              family = poisson)
```

que proporciona como estimaciones:

$$(E) \log[E(Y_{1j})] = 2,25559 + \theta_j^E$$

$$(O) \log[E(Y_{2j})] = 2,25559 + \theta_j^E + 7,84677 - 0,44067x_j + 0,00497x_j^2$$

$$(N) \log[E(Y_{3j})] = 2,25559 + \theta_j^E + 12,26572 - 0,69998x_j + 0,0097327x_j^2,$$

con $\theta_1^E = 0$, $\theta_2^E = 2,14255$, $\theta_3^E = 3,04271$, $\theta_4^E = 3,27917$, $\theta_5^E = 3,16419$, $\theta_6^E = 2,78484$ y $\theta_7^E = 2,13188$, correspondientes a los siete grupos de edad.

Este modelo log-lineal es equivalente al modelo logit generalizado (cuyas estimaciones corresponden a las líneas dibujadas en la Figura 4.1):

$$(E) \log(\pi_{1j}/\pi_{3j}) = -12,26572 + 0,69999x_j - 0,00973x_j^2$$

$$(O) \log(\pi_{2j}/\pi_{3j}) = 7,84677 - 12,26572 + (-0,44067 + 0,69999)x_j + (0,00497 - 0,00973)x_j^2, \quad \text{para } j = 1, 2, \dots, 7.$$

Comparando este modelo log-lineal de asociación con el de independencia entre edad y método ('fit.ind' en el Ejemplo 4.9),

```
anova(fit.ind, fit.aso, test = "Chisq")
```

obtenemos un p -valor prácticamente de cero, con lo que (rechazando independencia) evidenciamos significativamente la validez del modelo ajustado para explicar la relación entre las variables involucradas,

4.7. Modelos logit-multinomial y variables ordinales

Ya en la sección anterior, cuando hablamos de los modelos logit generalizados, describimos la forma de incluir en un modelo la información cuantitativa de los factores de tipo ordinal. Trataremos aquí unos cuantos casos particulares en tablas de contingencia en las que tenemos algún factor de clasificación de tipo ordinal.

Cuando los factores de una tabla de contingencia están relacionados entre sí y además algunos contienen información de tipo ordinal, es posible utilizar dicha información para modelizar asociaciones y conseguir mejores predicciones de las frecuencias de las celdas. De hecho, si todos los factores de una tabla están relacionados entre sí, describir la interacción con información exclusivamente nominal agota los grados de libertad disponibles dando lugar a un modelo saturado, inútil para predecir. Una aproximación simple y común es reemplazar las etiquetas ordinales por *scores* o puntuaciones numéricas, convirtiendo las medidas ordinales categóricas en medidas numéricas. Esto se puede entender como natural dado que las medidas ordinales realmente se derivan de alguna medición numérica que más tarde ha sido categorizada (por ejemplo, la categorización en grupos de edad, o en niveles de ingresos).

Las ventajas de tratar como ordinales a estos factores son: i) ciertos tipos de tendencias son mucho más fáciles de interpretar que en modelos nominales; ii) las asociaciones de variables ordinales a veces se pueden describir de forma más simple con un menor número de parámetros de regresión; iii) el enfocar los modelos ordinales de asociación de esta forma aumenta el número de grados de libertad residuales (menor número de parámetros estimados) y en consecuencia aumenta la potencia de los tests de asociación.

Así, cuando algún factor sea de tipo ordinal, construiremos nuevas variables “cuantitativas” con puntuaciones (*scores*) que contengan información numérica sobre sus diversas categorías de respuesta. Cuando el factor de clasificación proviene de la categorización de una variable numérica (como edad, ingresos,...), podemos utilizar las marcas de clase de cada intervalo como *scores*. Si la distancia entre dos categorías adyacentes es similar, es común tomar como puntuaciones números enteros consecutivos (por ejemplo, si $u_1, \dots, u_I$ son los *scores* atribuidos a las categorías de clasificación de un factor, una posibilidad es considerar $u_i = i$ para $i = 1, \dots, I$). Generalmente sin embargo, la definición de estas puntuaciones tiene algo de arbitraria; cuando se tienen dudas sobre la asignación de puntuaciones convendrá hacer algún análisis de sensibilidad sobre los resultados.

Ejemplo 4.11 (Afiliación e Ideología políticas) *Los datos que se presentan en la Tabla 4.7 (Hedlund, 1978, también en Agresti, 1990, pag.273) corresponden a un estudio realizado sobre una muestra de votantes de Wisconsin en las elecciones de 1976. Se pretendía investigar la relación entre la afiliación y la ideología políticas, ordenada esta última desde la más liberal a la más conservadora. Si existe asociación, un segundo objetivo es predecir la ideología de un individuo en función de*

su afiliación política.

Afiliación Política	Ideología Política			Total
	Liberal	Moderada	Conservadora	
Demócrata	143	156	100	399
Independiente	119	210	141	470
Republicano	15	72	127	214

Tabla 4.7: Ideología y afiliación políticas

Ejemplo 4.12 (Satisfacción e ingresos laborales) *Los datos que se presentan en la Tabla 4.8 (Agresti, 1990, pag. 262) corresponden a un estudio en el que se pretendía concluir sobre la relación entre el grado de satisfacción en el trabajo y los ingresos percibidos, ambas variables categorizadas y de tipo ordinal. Si existe asociación, un segundo objetivo es predecir el grado de satisfacción en el trabajo en función de los ingresos percibidos.*

Ingresos (en dólares)	Satisfacción en el trabajo			
	Muy Insatisfecho	Poco Insatisfecho	Moderadamente Satisfecho	Muy Satisfecho
< 6,000	20	24	80	82
6,000 – 15,000	22	38	104	125
15,000 – 25,000	13	28	81	113
> 25,000	7	18	54	92

Tabla 4.8: Satisfacción en el trabajo e ingresos

Consideramos a continuación la modelización de la asociación entre los factores de una tabla de contingencia de dos vías, en la que una de las variables ha de ser explicada en función de la otra, y alguna de ellas contiene información ordinal.

La variable a explicar la seguimos denotando por R , con I niveles de respuesta, y la variable explicativa por A , con J categorías de clasificación. Llamaremos y_{ij} al número de individuos clasificados en el nivel i de R y j de A .

En los modelos log-lineales que proponemos a continuación obviaremos mencionar las restricciones de identificabilidad necesarias en la estimación de los parámetros. No por ello son prescindibles.

4.7.1. Variable a explicar nominal y predictor ordinal

Supongamos una variable a explicar R de tipo nominal y una variable explicativa ordinal A . Para incorporar la información ordinal de A definimos una nueva variable con puntuaciones (*scores*) dados por $v_1 \leq \dots \leq v_J$. La asociación entre ambas variables podría venir descrita por el modelo log-lineal:

$$\log[E(Y_{ij})] = \mu + \lambda_i^R + \lambda_j^A + \theta_i^R v_j, \quad (4.22)$$

donde los θ_i son parámetros para describir el efecto de cada una de las categorías de respuesta. El modelo log-lineal anterior corresponde al modelo logit generalizado:

$$\log \left[\frac{\pi_{i|j}}{\pi_{I|j}} \right] = \alpha_i^R + \beta_i^R v_j, \quad (4.23)$$

donde $\alpha_i = \lambda_i^R - \lambda_i^R$ y $\beta_i^R = \theta_i^R - \theta_I^R$.

Es decir, se trata de un modelo logit lineal en los scores para el predictor A , con pendiente β_i^R que varía en las categorías de respuesta.

Que no exista asociación entre ambas variables es equivalente a decir que cuando nos movemos por los niveles de la variable explicativa A , no alteramos la respuesta R . Esto implicaría que $\theta_i^R = 0$, $\forall i = 1, 2, \dots, I$, y daría lugar al modelo de independencia:

$$\log[E(Y_{ij})] = \mu + \lambda_i^R + \lambda_j^A. \quad (4.24)$$

4.7.2. Variable a explicar ordinal y predictor nominal

Cuando la variable a explicar R contiene información ordinal y la variable explicativa A es puramente nominal, construimos una variable cuantitativa con *scores* para los niveles de respuesta de R , $u_1 \leq u_2 \leq \dots \leq u_I$. Un modelo de asociación uniforme entre los factores R y A viene dado por:

$$\log[E(Y_{ij})] = \mu + \lambda_i^R + \lambda_j^A + \theta_j^A u_i. \quad (4.25)$$

El modelo de asociación anterior (4.25) equivale a un modelo logit para categorías adyacentes de la forma

$$\log \left[\frac{\pi_{i+1|j}}{\pi_{i|j}} \right] = \alpha_i^R + \theta_j^A (u_{i+1} - u_i), \quad (4.26)$$

con $\alpha_i^R = \lambda_{i+1}^R - \lambda_i^R$. Cuando los niveles de R son equidistantes, $u_{i+1} - u_i = 1$, queda simplificado de la forma:

$$\log \left[\frac{\pi_{i+1|j}}{\pi_{i|j}} \right] = \alpha_i^R + \lambda_j^A. \quad (4.27)$$

Contrastar independencia entre ambos factores será equivalente a contrastar que moverse en los niveles del factor A no afecta a la respuesta de clasificación en los niveles de R , es decir, que todos los θ_j^A son iguales, $\theta_j^A = \theta$, $\forall j \in \{1, \dots, J\}$:

$$\log[E(Y_{ij})] = \mu + \alpha_i^R + \beta_j^A + \theta u_i. \quad (4.28)$$

Ejemplo 4.13 (Continuación del Ejemplo 4.11) Nos preguntamos en primer lugar si guarda alguna relación la tendencia ideológica de un individuo (variable 'ideol') con su afiliación política (variable 'afilia'). Si para ello queremos incorporar la información ordinal de la variable a explicar, 'ideol', construimos una variable de puntuaciones numéricas 'ideol.s', que da un 1 a ideología conservadora, un 2 a moderada y un 3 a liberal. Contrastar independencia resulta de comparar los modelos (4.25) y (4.28) para predecir las frecuencias de las celdas ('frec'), esto es,

```
fit.ind <- glm(frec ~ afilia + ideol + ideol.s, family = poisson)
fit.aso <- glm(frec ~ afilia + ideol + afilia : ideol.s, family = poisson)
anova(fit.ind, fit.aso, test = "Chisq")
```

que da un p -valor de $4,645 \cdot 10^{-23}$, lo que justifica el rechazo del modelo de independencia a favor del modelo de asociación considerado. De hecho, la bondad del ajuste para el modelo de asociación 'fit.aso', da un p -valor de 0.245 para una deviance de $D = 2,8149$ con 2 g.l.; es decir, el modelo fit.aso explica razonablemente bien la asociación entre afiliación e ideología.

Las estimaciones de los parámetros que proporciona este ajuste son

	<i>Estimate</i>	<i>Std. Error</i>
<i>(Intercept)</i>	1.2772	0.3745
<i>afiliaindependiente</i>	0.7135	0.1969
<i>afiliarepublicano</i>	1.5304	0.2352
<i>ideolmoderado</i>	1.4244	0.1409
<i>ideolconservador</i>	2.0488	0.2216
<i>afiliademocrata:ideol.s</i>	1.2134	0.1304
<i>afiliaindependiente:ideol.s</i>	0.9426	0.1260

Si $\pi_{i|j} = \Pr(\text{ideol} = i | \text{afilia} = j)$, el modelo ajustado nos lleva a la relación entre los logit dada por (4.27):

$$\frac{\pi_{2|j}}{\pi_{1|j}} = \exp\{1,42 - 2,049 + \theta_j\}$$

$$\frac{\pi_{3|j}}{\pi_{2|j}} = \exp\{-1,42 + \theta_j\}$$

con $\theta_1 = 1,21$ (demócrata), $\theta_2 = 0,943$ (independiente) y $\theta_3 = 0$ (republicano).

4.7.3. Variable a explicar ordinal y predictor ordinal

Cuando tanto la variable a explicar R como el predictor A , contienen información ordinal, para reflejar el orden de las categorías utilizaremos los *scores*:

$$u_1 \leq u_2 \leq \dots \leq u_I \rightsquigarrow R$$

$$v_1 \leq v_2 \leq \dots \leq v_J \rightsquigarrow A.$$

Un modelo log-lineal que recoge la asociación uniforme entre R y A es:

$$\log[E(Y_{ij})] = \mu + \lambda_i^R + \lambda_j^A + \theta u_i v_j, \quad (4.29)$$

donde λ_i^R y λ_j^A representan los efectos (nominales) de los factores de clasificación sobre las frecuencias de las celdas. El número de parámetros a estimar es $1 + (I - 1) + (J - 1) + 1$, con lo que los grados de libertad para los residuos son $IJ - I - J$.

El modelo log-lineal de asociación (4.29) equivale a un modelo logit para categorías adyacentes de la forma:

$$\log \left[\frac{\pi_{i+1|j}}{\pi_{i|j}} \right] = \alpha_i^R + \theta(u_{i+1} - u_i)v_j \quad (4.30)$$

donde $\alpha_i^R = \lambda_{i+1}^R - \lambda_i^R$. Cuando los niveles de la respuesta R son equidistantes, es decir, $u_{i+1} - u_i = 1$, el modelo logit anterior queda reducido a:

$$\log \left[\frac{\pi_{i+1|j}}{\pi_{i|j}} \right] = \alpha_i^R + \theta v_j. \quad (4.31)$$

La interpretación de θ podría ser: por cada unidad de cambio en la respuesta R (cambio de nivel, de i a $i + 1$), la diferencia de los logit de las probabilidades condicionadas para un cambio de nivel en el predictor A (de j a $j + 1$) es igual a θ .

Contrastar independencia equivale a contrastar $\theta = 0$. El modelo de independencia es pues,

$$\log[E(Y_{ij})] = \mu + \lambda_i^R + \lambda_j^A. \quad (4.32)$$

Ejemplo 4.14 (Continuación del Ejemplo 4.12) Introducimos los conteos en la variable 'frec', y creamos los factores de clasificación nominal sobre satisfacción en el trabajo, 'satisf' (a explicar), y nivel de ingresos percibidos, 'ingresos' (predictor). Para incluir la información ordinal de ambos, creamos unas variables de scores: 'satisf.s', con $u_i = i$, $i \in \{1, 2, 3, 4\}$ para los distintos grados de satisfacción, de menos a más, e 'ingresos.s', con $v_j = j$, $j \in \{1, 2, 3, 4\}$ los distintos niveles de ingresos, también de menos a más. La variable con la que tratamos de describir la según (4.29) es el producto de las puntuaciones $u_i v_j$:

$$\boxed{aso < -satisf.s * ingresos.s}$$

Contrastar independencia equivale a comparar los modelos (4.32) y (4.29), esto es, contrastar la significatividad de la variable 'aso' una vez ajustado el modelo de asociación:

```
fit1 <- glm(frec ~ ingresos + satisf + aso, family = poisson)
anova(fit1, test = "Chisq")
```

que da un p-valor de 0,00189 asociado al término 'aso' ($\theta = 0$ en (4.29)). Se rechaza pues el modelo de independencia a favor de la asociación descrita a través de la variable 'aso'.

Interpretando los resultados en términos del modelo logit (4.30), para el grupo de los que ganan más de 25000\$, la probabilidad de estar muy satisfecho en el trabajo es 11.65 veces la de estar muy insatisfecho; en el de los que ganan menos de 6000\$, la primera es sólo 4.25 veces mayor que la última.

4.8. Modelos de cambio

El análisis de datos procedentes de encuestas ha sido uno de los usos más importantes de los modelos log-lineales. Un tipo de encuestas particularmente interesantes es el de los *estudios de panel*. Las mismas cuestiones se preguntan al mismo grupo de individuos en dos o más épocas diferentes. Con las respuestas podemos estudiar la evolución de las poblaciones, esto es, los cambios en el tiempo.

A continuación estudiaremos algunos métodos útiles para analizar datos categóricos de muestras dependientes. Las muestras dependientes resultan de observaciones repetidas de la variable respuesta para un conjunto de individuos. También ocurren en *diseños pareados* (ver el Ejercicio 14 sobre movilidad ocupacional, en el que cada observación es el par constituido por la ocupación del padre y la del hijo). Este tipo de datos viene resumido en tablas de contingencia cuadradas en las que ambas clasificaciones (fila/columna) tienen el mismo número de categorías, $I = J$. Las tablas de contingencia cuadradas también se utilizan para hacer comparaciones pareadas; por ejemplo, comparar la calidad de dos vinos o el rendimiento de varios equipos deportistas. Presentaremos algunos análisis para este tipo de muestras dependientes.

Hay dos cuestiones básicas en el análisis de tablas cuadradas. En primer lugar, las probabilidades de las celdas o asociaciones pueden seguir un patrón de simetría alrededor de la diagonal principal de la tabla. En segundo lugar, las dos distribuciones marginales pueden diferir sistemáticamente. Para tablas 2×2 , el estudio de simetría es equivalente al estudio de las distribuciones marginales. Presentaremos modelos para estudiar estas cuestiones.

Consideramos pues, tablas cuadradas de dimensión $I \times I$ en las que estudiamos los cambios de estado a lo largo del tiempo, o entre la clasificación fila (F) y la clasificación columna (C). La predicción de las frecuencias Y_{ij} se realizará a través de modelos log-lineales. El modelo log-lineal saturado será:

$$\log[E(Y_{ij})] = \theta + \alpha_i^F + \beta_j^C + (\alpha\beta)_{ij}^{FC}, \quad i, j \in \{1, \dots, I\}.$$

en el que se estiman $(I - 1) \times (I - 1)$ parámetros para describir la asociación entre las respuestas sucesivas. El modelo minimal es el de independencia:

$$\pi_{ij} - \pi_{i+}\pi_{+j} \longleftrightarrow (\alpha\beta)_{ij}^{FC} = 0 \quad \forall i, j.$$

Modelos más interesantes están basados en *cadena de Markov*, en los que la respuesta actual depende simplemente de la respuesta previa. Estos modelos describen las probabilidades de transición para el cambio de un estado a otro entre dos instantes de tiempo.

4.8.1. Modelos Mover-Stayer

Ejemplo 4.15 (Movilidad) Consideremos los datos que se presentan en la Tabla 4.9 (Fingleton, 1984, también en Lindsey, 19997, pag. 142). Consisten en una muestra del 10 % del censo británico de 1971 sobre la gente que emigró entre 1966 y 1971 entre cuatro centros importantes de población de Gran Bretaña. Todos los individuos eran residentes británicos nacidos en la Nueva Commonwealth. El objetivo del análisis es modelizar la movilidad entre estos cuatro núcleos de población.

1966	1971			
	CC	ULY	WM	GL
CC	118	12	7	23
ULY	14	2127	86	130
WM	8	69	2548	107
GL	12	110	88	7712

Tabla 4.9: Cambio de residencia en la Commonwealth

El modelo *mover-stayer* parte del supuesto de considerar a los individuos de la encuesta divididos en dos grupos: la de los susceptibles de cambiar de opinión y la de los que no cambian. Si es ésta la situación, sabemos que todos los individuos fuera de la diagonal principal pertenecerán a la primera subpoblación (la de los 'susceptibles de cambio' o *movers*). Sin embargo, las frecuencias de la diagonal principal son algo más complejas de modelizar ya que contendrán tanto a individuos que no cambian nunca (*stayers*) como a algunos que no han llegado a cambiar dentro del período de observación pero son de los que cambian (*movers*). En este tipo de tablas a menudo

hay un alto número de individuos que dan la misma respuesta en los dos instantes de tiempo y son clasificados en la diagonal.

Planteemos un modelo simple: asumamos que las localizaciones de los *movers* en los dos instantes de tiempo observados son independientes. Si ignoramos la mezcla de individuos en la diagonal, podemos modelizar el resto de la tabla por **casi-independencia**, es decir, independencia en una tabla incompleta en la que la diagonal se ha perdido. Con esta hipótesis de independencia podemos obtener estimaciones del número de *movers* en la diagonal y, por lo tanto, del número de *stayers*.

Ejemplo 4.16 (Continuación del Ejemplo 4.15) *En el ejemplo de migración las frecuencias en la diagonal son muy grandes: hay mucha gente que no cambió su lugar de residencia durante el período considerado.*

Creamos dos factores que recogen el lugar de residencia en 1966 (resi1) y en 1971 (resi2). La variable 'frec' contiene las frecuencias de las celdas. Para concluir sobre si existe relación o no entre el lugar de residencia en los dos años de interés, contrastamos independencia ajustando un modelo log-lineal para predecir las frecuencias observadas en la tabla con los factores de clasificación:

$$fit1 < -glm(frec \sim resi1 + resi2, family = poisson)$$

El modelo tiene una deviance de $D = 19884$ con 9 g.l., que da un p-valor de 0, y en consecuencia evidencia significativamente la relación entre el lugar de residencia en 1966 y el de 1971.

A continuación buscamos un modelo con el que describir la asociación, y en concreto, cómo se ha producido el cambio de residencia entre 1966 y 1971. Proponemos un modelo 'mover-stayer' en el que asumimos independencia entre el lugar de residencia en 1971 y en 1966 para los individuos susceptibles de cambio, esto es, para los que no están en la diagonal principal:

$$fit2 < -update(fit1, subset - (resi1! - resi2))$$

con una deviance de $D = 4,367$ con 5 g.l., que da un p-valor de 0,497. En consecuencia, no hay evidencias de que los que cambiaron de residencia hayan seguido

algún patrón específico en sus desplazamientos (no se rechaza independencia para los individuos fuera de la diagonal). La dependencia detectada antes se explicaría pues, con los individuos que no cambiaron de residencia. El modelo mover-stayer podría ser un modelo válido para predecir el número de individuos susceptibles de cambio de los que entre 1966 y 1971 todavía no se habían movido. El número de 'movers' en la diagonal predichos por el modelo 'fit2' resulta de la sintaxis en R:

```
resi1.d <- -factor(c("CC", "ULY", "WM", "GL"))
resi2.d <- -factor(c("CC", "ULY", "WM", "GL"))
new.data <- -data.frame(resi1 = resi1.d, resi2 = resi2.d)
pred <- -predict(fit2, new.data, type = "response")
```

La variable 'pred' contiene las estimaciones de 'movers' en la diagonal de la tabla, que son (1,6, 95,2, 60,3, 154,6). La diferencia entre el número de individuos observados en la diagonal y el número de 'movers' predichos da el número de 'stayers' predichos. Así, la mayoría de la gente pertenece al grupo de los que no cambian de residencia.

Las estimaciones de todas las celdas se presentan en la Tabla 4.10 ('stayers' en la diagonal y 'movers' fuera de ella). Comparando con los valores observados en la Tabla 4.9, el modelo mover-stayer resulta bastante eficiente para estimar la movilidad en esas zonas.

	1971			
1966	CC	ULY	WM	GL
CC	116	13	11	18
ULY	12	2032	80	138
WM	9	72	2488	104
GL	13	107	90	7557

Tabla 4.10: Estimaciones del cambio de residencia en la Commonwealth con el modelo mover-stayer.

El modelo de *mover-stayer* da cabida a una probabilidad diferente de quedarse dentro de cada estado de respuesta: modeliza por un lado a los que se mueven y ajusta

con ellos las predicciones para los que no se mueven. Este modelo es sin embargo ilegítimo para calcular (estimar) probabilidades condicionales. La probabilidad de estar en un estado dado en el instante 2 depende de saber si el individuo es un *mover* o no, pero esto no se puede saber hasta que no se conoce el estado en el instante 2.

4.8.2. Simetría

Puesto que en los estudios de panel las variables respuesta se observan dos (o más) veces, podríamos esperar cierta simetría entre las mediciones realizadas.

Ejemplo 4.17 (Elecciones en Suecia) La Tabla 4.11 (Upton, 1978, pag.128, también en Lindsey, 1997, pag.35) contiene los datos de un estudio de panel de 1651 votantes en las elecciones en Suecia para estudiar la variación en el voto desde las elecciones de 1968 a las de 1970. Las opciones de voto son: SD=social demócrata, C=centro, P=independientes, Co=conservador. Para este tipo de datos cabe pensar que si se ha producido un cambio de voto, éste se ha dado hacia partidos de tendencias próximas (categorías adyacentes). De hecho, mirando la Tabla, se observa un cambio de voto hacia los partidos de centro.

	1970				
1968	SD	C	P	Co	Total
SD	850	35	25	6	916
C	9	286	21	6	322
P	3	26	185	5	219
Con	3	26	27	138	194

Tabla 4.11: Evolución en el voto de Suecia entre 1968 y 1970.

Hay **simetría** en una tabla de contingencia cuadrada cuando la probabilidad de cambio entre cualquier par de categorías es la misma en ambas direcciones:

$$\pi_{i|j} = \pi_{j|i} \quad \forall i, j.$$

Cuando todos los $\pi_{ij} > 0$, la simetría en una tabla de contingencia tiene una representación en términos de un modelo log-lineal para predecir las frecuencias observadas y_{ij} , de la forma:

$$\log[E(Y_{ij})] = \mu + \lambda_i + \lambda_j + \lambda_{ij}, \quad (4.33)$$

con la restricción adicional de que $\lambda_{ij} = \lambda_{ji}$. Nótese que los efectos principales fila y columna, λ_i y λ_j , son similares para garantizar la simetría sobre las frecuencias esperadas, $\mu_{ij} = \mu_{ji}$.

El modelo de simetría se puede ajustar a mano. Las ecuaciones de verosimilitud en una tabla de contingencia dan:

$$\hat{\mu}_{ij} + \hat{\mu}_{ji} = y_{ij} + y_{ji}, \quad (4.34)$$

donde $E(Y_{ij}) = \mu_{ij} = n\pi_{ij}$. Si además la solución ha de satisfacer la condición de simetría $\mu_{ij} = \mu_{ji}$, entonces ha de ser:

$$\hat{\mu}_{ij} = \frac{y_{ij} + y_{ji}}{2}, \quad \forall i \neq j, \quad \text{y } \hat{\mu}_{ii} = y_{ii}. \quad (4.35)$$

Es decir, el ajuste de los datos de la diagonal es perfecto, y para los datos fuera de la diagonal es el promedio de las frecuencias en celdas simétricas.

Para el modelo de simetría, los grados de libertad residuales son $I(I-1)/2$. El estadístico de Pearson para contrastar el ajuste del modelo de simetría se simplifica a:

$$\chi^2 = \sum \sum_{i < j} \frac{(y_{ij} - y_{ji})^2}{y_{ij} + y_{ji}}. \quad (4.36)$$

Para ajustar un modelo de simetría, basta con definir un factor con idénticas etiquetas para las posiciones de la tabla que son simétricas, y excluir en el ajuste los elementos de la diagonal.

Ejemplo 4.18 (Continuación del Ejemplo 4.17) Una vez introducidas las frecuencias observadas ('frec'), se define un factor de simetría que da etiquetas iguales a posiciones simétricas en la tabla. La introducción de datos en $\mathcal{R}$ viene dada por

```
frec <- c(850, 35, 25, 6, 9, 286, 21, 6, 3, 26, 185, 5, 3, 26, 27, 138)
voto68 <- factor(rep(c("SD", "C", "P", "CON"), rep(4, 4)))
voto70 <- factor(rep(c("SD", "C", "P", "CON"), 4))
simetria <- factor(c(1 : 4, 2, 5 : 7, 3, 6, 8, 9, 4, 7, 9, 10))
```

Para realizar el ajuste hemos de excluir los elementos de la diagonal (la simetría se contrasta para los elementos fuera de la diagonal principal). En $\mathcal{R}$ contrastar simetría es contrastar el modelo:

```
fit.sim <- glm(frec ~ simetria, family = poisson,
               subset = (voto68! = voto70))
```

que da una deviance de $D = 67,814$ con 6 g.l. y un $AIC = 131,34$. El término de simetría es significativo (p -valor de $2,55 \cdot 10^{-6}$ al compararlo con el modelo nulo), pero el p -valor asociado a la bondad del ajuste de $1,15 \cdot 10^{-12}$; esto es, la simetría completa no explica suficientemente bien el comportamiento del voto.

El modelo de simetría raramente ajusta bien. Por ejemplo, en cuanto hay heterogeneidad marginal, el modelo de simetría no se sustenta. Se dice que hay **homogeneidad marginal** si la distribución marginal de los estados es la misma en los diferentes momentos observados, esto es, si las probabilidades marginales son iguales en los dos instantes de tiempo:

$$\pi_{i+} = \pi_{+i}, \quad i = 1, \dots, I. \quad (4.37)$$

Cuando hay simetría, entonces $\pi_{i+} = \sum_j \pi_{ij} = \sum_j \pi_{ji} = \pi_{+i}$, luego también hay homogeneidad marginal. Para $I = 2$, simetría es equivalente a homogeneidad marginal.

Una forma de generalizar algo el modelo de simetría es borrar la restricción de igualdad de los parámetros de los efectos principales. El **modelo de casi-simetría** en una tabla cuadrada en la que los factores de clasificación fila y columna son, respectivamente F y C , viene dado por:

$$\log[E(Y_{ij})] = \mu + \alpha_i^F + \beta_j^C + (\alpha\beta)_{ij}^{FC}, \quad (4.38)$$

donde $(\alpha\beta)_{ij}^{FC} = (\alpha\beta)_{ji}^{FC}$.

Un modelo de casi-simetría proviene de la propiedad de **reversibilidad**, que implica probabilidades de transición iguales entre cada par de categorías respuesta, respetando las restricciones de que las probabilidades marginales han de coincidir con los valores observados.

El modelo de casi-simetría se puede ajustar, al igual que el de simetría, creando un factor de simetría, con pares de valores idénticos para las posiciones simétricas a ambos lados de la diagonal principal. Esta variable interviene en el ajuste junto con los dos parámetros marginales (factores fila y columna), excluyendo la diagonal principal.

Ejemplo 4.19 (Continuación del Ejemplo 4.17) *El modelo de casi-simetría viene dado por:*

```
fit.cs <- glm(frec ~ voto68 + voto70 + simetria,
              family = poisson, subset = (voto68 != voto70))
```

que da una deviance de $D = 2,4935$ con 3 g.l., un $AIC = 72,023$ y un p -valor de 0,4765 asociado a la bondad del ajuste. Resulta pues, un buen modelo para predecir el cambio de voto en las elecciones de Suecia 1968-1970. La predicción de las frecuencias viene dada en la Tabla 4.12.

	1970				
1968	SD	C	P	Co	Total
SD	850	37.29	23.68	5.03	916
C	6.71	286	23.34	5.95	322
P	4.32	23.66	185	6.02	219
Con	3.97	26.05	25.98	138	194

Tabla 4.12: Predicción de la evolución en el voto de Suecia entre 1968 y 1970 con un modelo de casi-simetría

4.9. Ejercicios

1. Considera una tabla de contingencia $I \times 2$ en la que los totales fila y_{i+} están fijos para $i = 1, 2, \dots, I$, tal y como se muestra en la Tabla 4.13.

	vivos	muertos	total
C_1	y_{11}	y_{12}	y_{1+}
C_2	y_{21}	y_{22}	y_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_I	y_{I1}	y_{I2}	y_{I+}

Tabla 4.13: Tabla de contingencia $I \times 2$.

- a) Si asumes que cada fila C_i constituye una realización independiente de un experimento binomial, donde la variable que recoge el número de éxitos es $Z_i = Y_{i1}$ y los parámetros son n_i y π_i :
 - 1) ¿Qué representa el parámetro π_i en cada prueba binomial C_i ? Utilizando la notación de la Tabla 1, da una estimación burda de π_i .
 - 2) ¿Cuál es el valor esperado del número de individuos 'vivos' en la prueba binomial C_i , (Y_{i1})? ¿Y el de 'muertos' (Y_{i2})? Con la notación de la tabla, obtén la expresión del estadístico de bondad de ajuste:

$$\chi^2 = \sum_{k=1}^N \frac{(o_i - e_i)^2}{c_i},$$

donde N representa el número de celdas en la tabla, o_i las frecuencias observadas y e_i las frecuencias esperadas.

- 3) Construye la verosimilitud para cada fila.
- 4) Construye la verosimilitud conjunta para toda la tabla.

La que has construido es la **distribución producto binomial**, que es un caso particular de la distribución producto multinomial.

- b) Prueba que el modelo log-lineal con:

$$\begin{aligned}\eta_{i1} &= \log[E(Z_i)] = x'_{i1}\beta \\ \eta_{i2} &= \log[E(n_i - Z_i)] = x'_{i2}\beta\end{aligned}$$

es equivalente al modelo logístico:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = x'_i\beta,$$

con $x_i = x_{i1} - x_{i2}$.

- En una tabla de contingencia 2×2 , el modelo log-lineal maximal se puede escribir como:

$$\begin{aligned}\eta_{11} &= \mu + \alpha + \beta + (\alpha\beta) & \eta_{12} &= \mu + \alpha - \beta - (\alpha\beta) \\ \eta_{21} &= \mu - \alpha + \beta - (\alpha\beta) & \eta_{22} &= \mu - \alpha - \beta + (\alpha\beta),\end{aligned}$$

donde η_{ij} es el predictor lineal, $E(Y_{ij}) = n\pi_{ij}$ y $n = \sum_i \sum_j Y_{ij}$.

Prueba que el término de interacción $(\alpha\beta)$ viene dado por:

$$(\alpha\beta) = 1/4 \log(\phi),$$

con $\phi = \pi_{11}\pi_{22}/\pi_{12}\pi_{21}$, y por lo tanto que $\phi = 1$ corresponde a un modelo sin interacción.

- Replantéate el Ejercicio 3 del Tema 3 y resuélvelo desde la perspectiva de los modelos log-lineales.
- Replantéate el Ejercicio del Tema 3 y resuélvelo desde la perspectiva de los modelos log-lineales.
- En la Tabla 4.14 (Lloyd, 1999, pag.301) aparecen los datos de una encuesta social en Dinamarca, en la que se clasificó a 165 individuos según su estatus marital (soltero, casado, o divorciado alguna vez) y su edad. Se pretendía establecer una relación entre las variables estatus marital y edad, con el fin de concluir sobre el comportamiento social en Dinamarca (ver Figura 4.2).

Edad	Marca-Edad (x_i)	Solteros	Casados	Divorciados	Total
17-21	19	17	1	0	18
21-25	23	16	8	0	24
25-30	27.5	8	17	1	26
30-40	35	6	22	4	32
40-50	45	5	21	6	32
50-60	55	3	17	8	28
60-70	65	2	8	6	16
70+	75	1	3	5	9
Total		58	97	30	185

Tabla 4.14: Encuesta danesa sobre edad y estatus marital.

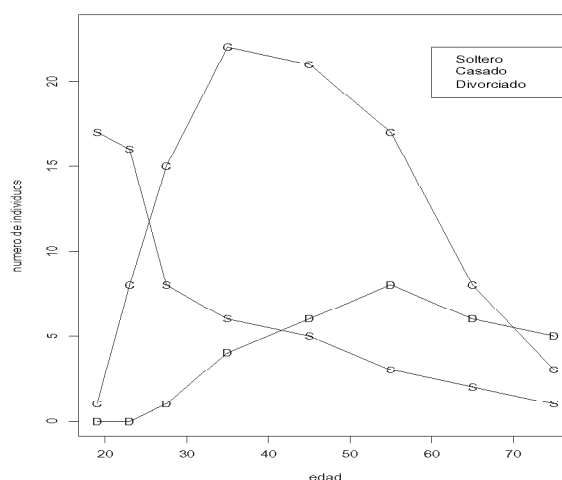


Figura 4.2: Encuesta danesa sobre edad y estatus marital.

- Identifica las variables involucradas en el análisis de estos datos. ¿Cuáles son de tipo cualitativo –nominal u ordinal– y cuáles cuantitativas?
- A la vista de los datos y del gráfico (Figura 4.2), ¿aprecias algún tipo de tendencia entre edad y estatus marital? El efecto de la edad, ¿es similar para los solteros, los casados y los divorciados?
- ¿Existe alguna relación entre el estatus marital y la edad? Tu conclusión, ¿es coherente con lo que comentaste en el apartado anterior?
- Si pretendemos llevar a cabo un análisis predictivo y explicar la proporción de individuos en cada uno de los estatus matrimoniales en función de la edad, ¿cuál es la variable a explicar y cuál la explicativa? Incorporando toda la información de que dispones, propón, formula y ajusta un modelo log-lineal (llámalo *fit2*) útil para concluir sobre la relación entre edad y estatus marital. Comenta las hipótesis que asumes y tus conclusiones.
- Considera que y_{ij} representa el número de individuos observados en el grupo de edad $i \in \{1, \dots, 8\}$ y con estatus marital $j \in \{1, 2, 3\}$; x_i representa la marca de clase del grupo de edad i . Construye la variable $lx_i = \log(x_i - 16)$ y ajusta el modelo log-lineal cuadrático *fit3*, dado por:

$$fit3: \log[E(Y_{ij})] = \theta + \alpha_i + \beta_j + \gamma_j lx_i + \delta_j lx_i^2. \quad (4.39)$$

¿Qué significa dicho modelo? ¿Qué tipo de relación implica entre la edad y el número de individuos en cada categoría de estatus marital?

Una vez hecho el ajuste, ¿cuáles son tus conclusiones sobre la relación entre edad y estatus marital?

Compara el ajuste *fit3* con el que tú propusiste (*fit2*) ¿Cuál es el mejor y en qué te basas? (inspecciona también los residuos).

- f) Para comparar las probabilidades condicionadas $\pi_{j|i} = Pr(estatus = j | edad = x_i)$ y concluir sobre cuál es el estado marital más común en cada grupo de edad podemos acudir al modelo logit multinomial equivalente al modelo log-lineal *fit3*. Deriva dicho modelo logit multinomial a partir de *fit3* para predecir $\log(\pi_{1|i}/\pi_{2|i})$ y $\log(\pi_{3|i}/\pi_{2|i})$.

Para el grupo de edad 40–50, ¿cuál es la relación que predices con *fit3* entre las proporciones de individuos en cada estatus marital? ¿Qué es lo más común en ese grupo de edad?

6. En la Tabla 4.15 (Muller y Mayhall, 1970, también en Krzanowski, 1998, pag. 230) tenemos datos de un estudio de investigación sobre la incidencia de una característica morfológica denominada *Torus Mandibularis* (una pequeña protuberancia cerca de la barbilla), en tres poblaciones esquimales, clasificadas también por edad y sexo.
 - a) Si en principio simplemente tratas de identificar relaciones entre las distintas variables observadas, busca un modelo que te permita obtener conclusiones sobre ellas. Para encontrar dicho modelo, plantea cuál es el modelo saturado y mediante un procedimiento de búsqueda por pasos consigue un buen modelo de predicción de las frecuencias de las celdas. Justifica las decisiones que vas tomando en dicha búsqueda.
 - b) Formula el modelo que acabas de ajustar. ¿Concluirías que para explicar la incidencia de *Torus Mandibularis* tienen algo que ver el grupo poblacional al que pertenecen los individuos y su edad? Justifica tu respuesta utilizando los resultados obtenidos en el ajuste.
 - c) Considera que ahora estamos interesados específicamente en predecir la incidencia de *Torus Mandibularis*, (probabilidad de que se dé la mencionada malformación) en función del grupo poblacional al que pertenece y la edad del individuo.
 - 1) Identifica la variable respuesta y las variables explicativas. ¿Cuál es el parámetro de interés? ¿Qué distribución asumirías para la variable respuesta?

			Grupos de edad						
Pop	Sex	Inc	1-10	11-20	21-30	31-40	41-50	51+	Tot
Iglook	H	+	4	8	13	18	10	12	65
		-	44	32	21	5	0	1	103
	M	+	1	11	19	13	6	10	60
		-	42	17	17	5	4	2	87
Beech	H	+	2	5	7	5	4	4	27
		-	17	10	6	2	2	1	38
	M	+	1	3	2	5	4	2	17
		-	12	16	6	2	0	0	36
Aleut	H	+	4	2	4	7	4	3	24
		-	6	13	3	3	5	3	33
	M	+	3	1	2	2	2	4	14
		-	10	7	12	5	2	1	37
Totales		146	125	112	72	43	43	541	

Tabla 4.15: Incidencia de Torus Mandibularis en tres poblaciones de esquimales. Presencia (+) y ausencia (-).

- 2) Consigue un modelo (el mejor posible) que permita predecir la probabilidad de padecer la malformación en función del grupo poblacional y la edad del individuo. Formula dicho modelo con los resultados obtenidos.
 - 3) En los individuos entre 31 y 40 años, ¿en qué grupo poblacional es más alta la incidencia del Torus Mandibularis? ¿Cuáles son las probabilidades (estimadas) asociadas en función de las cuales has contestado?
 - 4) En el grupo de los iglook, ¿tiene algún efecto la edad sobre la incidencia del Torus Mandibularis? Justifica tu respuesta utilizando el modelo ajustado y cuantificando las probabilidades asociadas.
7. Considera los datos de la Tabla 4.16, que corresponden al estatus de supervivencia, tras tres años, de pacientes con cáncer de mama, clasificadas por su edad (en tres grupos) y por el tipo de tumor (*maligno o benigno*).
- a) Formula al menos dos preguntas de interés científico para contestar a partir de los datos anteriores.
 - b) Formula el modelo log-lineal saturado adecuado a los datos de la Tabla 4.16. Denota con α el efecto de la 'edad', con β el efecto del tipo de

	edad	maligno	vive	
			si	no
C_1	< 50	no	77	10
C_2	< 50	si	51	13
C_3	50 – 69	no	51	11
C_4	50 – 69	si	38	20
C_5	> 70	no	7	3
C_6	> 70	si	6	3

Tabla 4.16: Supervivencia en pacientes de cáncer.

tumor, con γ el efecto del estatus de supervivencia, y con combinaciones de letras las interacciones.

- c) Formula el modelo a ajustar suponiendo que existe independencia completa entre los tres factores de clasificación. Contrasta la hipótesis de independencia completa.
- d) Formula el modelo correspondiente a la hipótesis de que el que una paciente sobreviva no tiene nada que ver con su edad o con el tipo de tumor que padece, y resuelve el contraste.
- e) Formula el modelo correspondiente a la hipótesis de que en cada grupo de edad, el que una paciente sobreviva o no es independiente del tipo de tumor, y resuelve el contraste.
- f) Consigue el mejor ajuste para explicar los datos de supervivencia al cáncer de mama. Justifícalo y formula explícitamente el modelo que has ajustado.
- g) ¿Cómo están relacionados la supervivencia de una paciente, con su edad y con el tipo de tumor que padece? Identifica los coeficientes que explican dicha relación.
- h) Según el modelo que has ajustado, calcula el número de pacientes que se espera que sobrevivan en función del grupo de edad al que pertenecen y el tipo de tumor que padecen.
- i) Compara los resultados que has obtenido en 'el mejor ajuste Poisson' con 'el mejor ajuste logístico' que obtuviste en el Ejercicio 5 del Tema 3. ¿Detectas alguna relación entre los coeficientes que explican la relación entre la probabilidad de supervivencia y los factores 'edad' y 'maligno' en sendos ajustes? ¿Podrías considerar equivalentes las conclusiones? ¿En qué te basas? (Utiliza la parte teórica del Ejercicio 1 de este Tema 4).

8. En la *Fiji Fertility Survey* de 1975 se entrevistó a 1607 mujeres casadas y fecundas. Fueron clasificadas por edad actual, nivel de estudios, deseo de tener más hijos y uso de contraceptivos. Los datos aparecen en la Tabla 4.17 (Rodríguez, 1998).

¿Están relacionados factores como la educación, la edad, o simplemente el deseo de no tener más hijos, con el uso de contraceptivos?

Edad	Educación	¿Desea más hijos?	Uso de contraceptivos	
			No	Si
< 25	baja	si	53	6
		no	10	4
	alta	si	212	52
		no	50	10
25–29	baja	si	60	14
		no	19	10
	alta	si	155	54
		no	65	27
30–39	baja	si	112	33
		no	77	80
	alta	si	118	46
		no	68	78
40–49	baja	si	35	6
		no	46	48
	alta	si	8	8
		no	12	31

Tabla 4.17: Datos sobre contraceptivos.

9. Considera los datos de la Tabla 4.18 (Wermuth, 1976, también en McCullagh y Nelder, 1989, pag. 190), que proporcionan el período de gestación y los ratios de mortalidad perinatal para un grupo de mujeres alemanas, muchas de las cuales estaban embarazadas por primera vez o habían tenido complicaciones en embarazos anteriores.
- Considera en primer lugar como variable respuesta la que contiene los conteos de 'vivos' y 'muertos' (variable 'Mortalidad Perinatal'). Al ser conteos es lógico considerar que dichas observaciones provienen de una distribución de Poisson:

$$Y.p < -c(50, 315, 9, 40, 41, 147, 4, 11, 24, 4012, 6, 459, 14, 1494, 1, 124)$$

período gestacional (días)	Edad de la madre	Cigarrillos fumados	Mortalidad Perinatal muertes	total nacim
197-260	< 30	≤ 5	50	365
		> 5	9	49
	30+	≤ 5	41	188
		> 5	4	15
261+	< 30	≤ 5	24	4036
		> 5	6	465
	30+	≤ 5	14	1508
		> 5	1	125

Tabla 4.18: Efecto de la edad de la madre y sus hábitos con el tabaco sobre el período gestacional y la mortalidad perinatal.

- Formula explícitamente el modelo log-lineal saturado para predecir los conteos Y_{ip} , denotando por α el efecto debido al “período de gestación”, β para la “edad de la madre”, γ para el “número de cigarrillos” consumidos diaria y habitualmente por la madre, y δ para “si nació vivo o muerto”.
Ajusta un modelo log-lineal excluyendo todas las interacciones de orden 3 y superior. Contrasta la bondad de dicho ajuste (es decir, compáralo con el modelo saturado). ¿Cuáles son las conclusiones que extraes de los resultados del ajuste?
- Investiga si el tiempo de gestación está relacionado con la edad de la madre. Plantea y resuelve el contraste correspondiente (especifica las dos hipótesis y obra en consecuencia comparando sus *deviances*). Explica tus conclusiones.
- Investiga si el tiempo de gestación está relacionado con el número de cigarrillos que consume la madre diariamente. Plantea y resuelve el contraste correspondiente. Explica tus conclusiones.
- Utilizando los ajustes anteriores, identifica algún factor o interacción sospechoso de no explicar nada sobre los datos y ajusta ‘el mejor modelo log-lineal’ posible. Formula el modelo que has conseguido y evalúa la bondad del ajuste conseguido. ¿Cuáles son las conclusiones que obtienes con el ajuste conseguido (especialmente, cómo interpretas los términos de interacción)?
- Con el modelo que has ajustado en el apartado anterior, ¿cuántos nacimientos vivos predices cuando la madre tiene más de 30 años y fuma más de

5 cigarrillos diarios? ¿Cuántos nacimientos prematuros (vivos) predices para el mismo tipo de madres? ¿Cuál es la probabilidad de que un niño nazca prematuro (y vivo) si su madre tiene más de 30 años y fuma más de 5 cigarrillos diarios? (Consulta la matriz de diseño para no meter la pata).

f) Repite el apartado anterior para el grupo de madres con menos de 30 años que fuman menos de 5 cigarrillos diarios? ¿Moraleja?

- Considera ahora como variable respuesta el número de nacidos muertos y vivos,

$$Y.b < -cbind(muertos, total - muertos)$$

- Ajusta un modelo probit simple (con interacciones a lo sumo de orden 2) que explique lo mejor posible la probabilidad de morir al nacer en función del período gestacional, la edad de la madre y sus hábitos de tabaco. Evalúa la bondad del ajuste y explica tus conclusiones.
 - ¿Cuál es la probabilidad (estimada) de que un niño prematuro muera al nacer cuando su madre tiene más de 30 años y fuma más de 5 cigarrillos diarios?
 - Repite los dos apartados anteriores utilizando un modelo logit.
10. Algunos escarabajos del género *Tribolium* son caníbales: los adultos se comen los huevos de su propia especie, así como los de especies afines. Este estudio investigaba si tres especies adultas de *Tribolium* (*castaneum*, *confusum* y *madens*) han desarrollado una ventaja evolutiva para reconocer y evitar comerse los huevos de su propia especie. En el experimento se aisló a un número de escarabajos adultos de la misma especie y se les enfrentó a cada uno a un total de 150 huevos (50 de cada tipo), habiendo antes mezclado los huevos para asegurar uniformidad en el conjunto. En la Tabla 4.19 se presenta el consumo en adultos de la especie *castaneum*; las tres últimas columnas representan el número de huevos *castaneum*, *confusum* y *madens* que se dejaron sin comer después de dos días expuestos ante los beetles adultos. La primera columna (*Replicaciones 1, 2 y 3*) corresponden a diferentes ocasiones en que el experimento se llevó a cabo. El experimento fue realizado en la Universidad de Chicago por Mr S. Teleky, del *Department of Evolutionary Biology* (McCullagh y Nelder, 1989, pag.143).

Los adultos de la especie *castaneum*, ¿muestran algún tipo de preferencia por los huevos de otras especies?

Replicación	Especie del los huevos		
	Castaneum	Confusum	Madens
1	25	24	15
	26	14	26
	26	24	32
2	29	14	32
	28	13	19
	27	19	16
3	26	10	13
	20	7	15
	14	14	23

Tabla 4.19: Número de huevos que no se comieron los escarabajos castaneum de un total de 50 huevos de cada especie en tres replicaciones distintas del experimento.

11. Considera y analiza los datos de la Tabla 4.20, tomados de Forthofer y Lehnen (1981) (también en Agresti, 1990, pag.280). Se trata de establecer asociaciones entre el estatus de fumador (F), los resultados en un test respiratorio (R) y la edad (E) de caucasianos en una planta industrial en Houston, Texas, durante 1974 y 1975. El fumar, ¿tiene alguna influencia sobre las deficiencias respiratorias de los trabajadores? ¿Y la edad?
12. La Tabla 4.21 (Madsen, 1976, también en Agresti, 1990, pag.341) contiene los datos de una muestra de 1681 residentes en 12 áreas de Copenague. Las variables son: tipo de casa (T), grado de contacto con otros residentes (C), sentimiento de la influencia del individuo sobre la administración de la vivienda (A) y satisfacción con las condiciones de la vivienda (S). ¿Tiene algo que ver el grado de satisfacción en la vivienda con el tipo de casa o el contacto más o menos fluido con los vecinos? Analiza estos datos y extrae las conclusiones pertinentes.
13. La Tabla 4.22 (Upton, 1978, pag.111, también en Lindsey, 1997, pag.44) contiene los datos relativos a cambios en el voto entre las elecciones británicas de 1964 y las de 1966 para un panel de votantes. Los miembros del panel son aquellos votantes que durante dicho período de 3 años han permanecido en un mismo panorama político definido por las tres opciones de voto: conservador, liberal y laborista.

¿Qué has descubierto sobre la 'lealtad al partido' y la distancia en el movimien-

Edad	Estatus de Fumador	Resultados en el test respiratorio		
		Normal	Borderline	Anormal
< 40	Nunca fumó	577	27	7
	Ex-fumador	192	20	3
	Fumador	682	46	11
40 – 59	Nunca fumó	164	4	0
	Ex-fumador	145	15	7
	Fumador	245	47	27

Tabla 4.20: Trabajadores en Houston clasificados por estatus de fumador, edad y resultados de un test respiratorio

to de votantes entre los partidos (desde la derecha más derecha a la izquierda más izquierda, pasando por la abstención)?

14. Un estudio clásico sobre la movilidad intergeneracional en Gran Bretaña (Bishop et al, 1975, p.100, también en Lindsey, 1997, pag.44) da los resultados presentados en la Tabla 4.23. La codificación de las categorías es: A: profesional, alto administrativo; B: directivo, ejecutivo, alto supervisor; C: bajo inspector, supervisor; D: rutina no-manual, obrero manual especializado, y E: obrero semi-especializado y obrero manual no especializado.

¿Cómo depende la profesión de un hijo de la de su padre? ¿Cuánto varía la escala profesional en una familia durante una generación? ¿Se aprecia alguna tendencia en un sentido o en el otro?

15. Los empleados de entre 30 y 39 años en fábricas de armamento (*Royal Ordnance*) en la época de 1943 a 1946 fueron sometidos a unas pruebas ópticas para investigar la visión de lejos sin ayuda de lentes (Stuart, 1953, from the Association of Optical Practitioners, también en Lindsey, 1997, pag.45). Los datos se presentan en la Tabla 4.24.

Estudia la relación entre los dos ojos, ¿tienen ambos generalmente una visión igual de buena? ¿Es tan probable que un ojo izquierdo sea peor que el derecho como lo contrario? ¿Las conclusiones son similares para los dos sexos?

16. En un estudio sobre adolescentes, a todos los estudiantes de 10 *high schools* en Estados Unidos se les formularon varias preguntas sobre pertenencia a bandas y la importancia de éstas. Entre otras cosas, se les preguntó si ellos se consideraban miembros de alguna banda (variable *Miembro*). Otra pregunta de actitud

Tipo de Vivienda	Satisf.	C=Poco Contacto			C=Mucho Contacto		
		A=Bajo	A=Medio	A=Alto	A=B	A=M	A=A
Rascacielos	Baja	21	21	28	14	19	37
	Media	34	22	36	17	23	40
	Alta	10	11	36	3	5	23
Apartamentos	Baja	61	23	17	78	46	43
	Media	43	35	40	48	45	86
	Alta	26	18	54	15	25	62
Chalets	Baja	13	9	10	20	23	20
	Media	8	8	12	10	22	24
	Alta	6	7	9	7	10	21
Casonas	Baja	18	6	7	57	23	13
	Media	15	13	13	31	21	13
	Alta	7	5	11	5	6	13

Tabla 4.21: Satisfacción en el hogar y condiciones de la vivienda

1966	1964			
	Conservador	Liberal	Laborista	Abstención
Conservador	157	4	17	9
Liberal	16	159	13	9
Laborista	11	9	51	1
Abstención	18	12	11	15

Tabla 4.22: Elecciones británicas en 1966. Comparación con los resultados de 1964.

era si consideraban que el ser miembro de una banda obligaba, en ocasiones, a ir en contra de sus principios personales (variable *Actitud*). Las respuestas fueron anotadas al principio y al final del curso escolar (Respuestas 1 y 2). Los resultados de la encuesta se presentan en la Tabla 4.25 (Coleman, 1964, pag.171, también en Lindsey, 1997, pag.45)

El considerarse miembro de una banda, ¿está relacionado con la actitud hacia hacer primar los intereses de la banda frente a los suyos propios? ¿Cómo dependen las respuestas al final del curso de las respuestas al principio? ¿Ha habido algún cambio? ¿En qué sentido?

17. En un estudio en seis ciudades de los Estados Unidos se pretendía investigar el efecto de la polución en el aire sobre la salud a lo largo de los años (Fitzmaurice y Laird, 1993). Parte de los datos de ese estudio conciernen a niños de dos

Hijo	Padre				
	A	B	C	D	E
A	50	45	8	18	8
B	28	174	84	154	55
C	11	78	110	223	96
D	14	150	185	714	447
E	3	42	72	320	411

Tabla 4.23: Estudio de movilidad social intergeneracional.

Grad. ojo dcho.	Grad. Ojo Izqdo.				
	4	3	2	1	Total
	Mujer				
4	1520	266	124	66	1976
3	234	1512	432	78	2256
2	117	362	1772	205	2456
1	36	82	179	492	789
Total	1907	2222	2507	841	7477
	Hombre				
4	821	122	85	35	1053
3	116	494	145	27	782
2	72	151	583	87	893
1	43	34	106	331	514
Total	1052	791	919	480	3242

Tabla 4.24: Tests sobre los ojos para visión de lejos sin ayuda de lentes

Actitud 1	Miembro 1	Miembro 2	Actitud 2	
			Favorable	Desfavorable
Favorable	Yes	Yes	458	140
Desfavorable			171	182
Favorable	No	Yes	184	75
Desfavorable			85	97
Favorable	Yes	No	110	49
Desfavorable			56	87
Favorable	No	No	531	281
Desfavorable			338	554

Tabla 4.25: Pertenencia a bandas y actitud en ellas

ciudades de USA, que fueron seguidos cada año desde que éstos tenían 7 años hasta los 10. Estos datos se presentan en la Tabla 4.26. Cada año se anotaba el estado respiratorio del niño (si padecía de 'pitos' o no). También se anotó si la madre del niño fumaba o no cuando el niño tenía 7 años (aunque luego hubiera cambiado la situación con el paso del tiempo).

Ignorando el hecho de que los datos son longitudinales, determina si hay heterogeneidad entre los niños incluidos en el estudio. El estado respiratorio del niño, ¿depende de si la madre fumaba o no cuando él tenía 7 años? ¿Hay diferencias entre las dos ciudades?

18. Las respuestas a tres preguntas sobre el aborto correspondientes a varias encuestas llevadas a cabo durante tres años se presentan en la Tabla 4.27 (Haberman, 1979, pag.482, también en Lindsey, 1997, pag. 46).

Las preguntas eran: Una mujer embarazada, debería poder acceder a un aborto legal (1) si está casada y no desea más niños, (2) si la familia tiene ingresos muy bajos y su economía no puede soportar más hijos, (3) si no está casada y no quiere casarse?

Las personas que contestaron a la encuesta eran cristianas y blancas (General Social Survey in the United States of America). No se trata de un panel, de forma que los individuos encuestados cada año son diferentes. ¿Hay cambios de actitud con el paso del tiempo?

19. Weisberg (1985, pag. 192-195) presenta un estudio sobre los salarios y otras características de los profesores de un pequeño colegio. Los datos incluyen las respuestas a seis variables de un total de 52 profesores y están representados en la Tabla 4.28. Estas variables son:

	Edad				Fumadora	
	7	8	9	10	No	Sí
C1	No	No	No	No	237	118
				Sí	10	6
				Sí No	15	8
				Sí	4	2
			Sí	No No	16	11
				Sí	2	1
				Sí No	7	6
				Sí	3	4
C2	No	No	No	No	24	7
				Sí	3	3
				Sí No	3	3
				Sí	2	1
		Sí	No	No No	6	4
				Sí	2	2
				Sí No	5	4
				Sí	11	7

Tabla 4.26: Efecto de la contaminación en los niños de dos ciudades de USA

SX= sexo (1=mujer, 0=hombre)

RK= rango (1=asistente, 2=asociado, 3=titular)

YR=número de años ejerciendo con el rango actual

DG= máxima graduación alcanzada (1=PhD y 0=licenciatura)

YD= número de años que hace que consiguió su graduación más alta

SL= salario académico anual en dólares.

¿Hay alguna diferencia entre hombres y mujeres en el salario recibido? ¿Existe tal diferencia si sólo tenemos en cuenta la graduación más alta conseguida y los años que hace que consiguieron dicha graduación?

20. Los datos en la Tabla 4.29 (Agresti, 1990, pag.124) presentan los datos obtenidos en un estudio de la Universidad de Arizona sobre los hábitos tabaquísticos de estudiantes en colegios de Arizona, en relación al comportamiento de sus padres.

El que los padres fumen o no, ¿está relacionado con el comportamiento de los hijos? ¿Cuál es la probabilidad media de que un hijo fume si ninguno de sus padres fuma?

21. Los datos de la Tabla 4.30 describen los patrones de comportamiento en el

Respuesta	Año		
	1972	1973	1974
SSS	334	428	413
SSN	34	29	29
SNS	12	13	16
SNN	15	17	18
NSS	53	42	60
NSN	63	53	57
NNS	43	31	37
NNN	501	453	430

Tabla 4.27: Actitud ante la despenalización del aborto

consumo de drogas psicotrópicas en una muestra de individuos del Oeste de Londres

¿Hay diferencias por sexo en el consumo de drogas? ¿Cómo influye la edad para explicar el consumo de psicotrópicos? El efecto de la edad, ¿es distinto para hombres y para mujeres?

22. En la Tabla 4.31 aparecen las cifras obtenidas en un estudio sobre adolescentes de 15 y 16 años (Morgan and Teachman, 1988, también en Agresti, 1990, pag.157). El suceso de interés es si han tenido alguna vez relaciones sexuales.
¿Hay diferencias entre individuos de raza blanca y de raza negra respecto al hecho de haber tenido relaciones sexuales antes de los 16 años? ¿Y entre chicos y chicas? Predice la proporción de individuos que han tenido relaciones sexuales antes de los 16 años en función de su sexo y raza.
23. Se realizó un estudio con el fin de investigar sobre la relación entre la salud mental y el nivel socioeconómico de los padres en una muestra de individuos. Se sospechaba que un estatus bajo estaba asociado con niveles inferiores de salud mental. ¿Puedes verificar la hipótesis? Los datos aparecen en la Tabla 4.32 (Srole et al., 1978, pag.289, también en Lloyd, 1999, pag.343).

SX	RK	YR	DG	YD	SL	SX	RK	YR	DG	YD	SL
0	3	25	1	35	36350	0	2	11	1	14	24800
0	3	13	1	22	35350	1	3	5	1	16	25500
0	3	10	1	23	28200	0	2	3	0	7	26182
1	3	7	1	27	26775	0	2	3	0	17	23725
0	3	19	0	30	33696	1	1	10	0	15	21600
0	3	16	1	21	28516	0	2	11	0	31	23300
1	3	0	0	32	24900	0	1	9	0	14	23713
0	3	16	1	18	31909	1	2	4	0	33	20690
0	3	13	0	30	31850	1	2	6	0	29	22450
0	3	13	0	31	32850	0	2	1	1	9	20850
0	3	12	1	22	27025	1	1	8	1	14	18304
0	2	15	1	19	24750	0	1	4	1	4	17095
0	3	9	1	17	28200	0	1	4	1	5	16700
0	2	9	0	27	23712	0	1	4	1	4	17600
0	3	9	1	24	25748	0	1	3	1	4	18075
0	3	7	1	15	29342	0	1	3	0	11	18000
0	3	13	1	20	31114	0	2	0	1	7	20999
0	2	11	0	14	24742	1	1	3	1	3	17250
0	2	10	0	15	22906	0	1	2	1	3	16500
0	3	6	0	21	24450	0	1	2	1	1	16094
0	1	16	0	23	19175	1	1	2	1	6	16150
0	2	8	0	31	20525	1	1	2	1	2	15350
0	3	7	1	13	27959	0	1	1	1	1	16244
1	3	8	1	24	38045	1	1	1	1	1	16686
0	2	9	1	12	24832	1	1	1	1	1	15000
0	3	5	1	18	25400	1	1	0	1	2	20300

Tabla 4.28: Salarios de profesores de colegios

Padres que fuman	Hijo que fuma	Número
ninguno	SI	188
ninguno	NO	1168
uno	SI	416
uno	NO	1823
ambos	SI	400
ambos	NO	1380

Tabla 4.29: Hábitos de tabaco en estudiantes y en sus padres.

Sexo	Edad	Usa	No usa
hombre	16-29	21	683
hombre	30-44	32	596
hombre	45-64	70	705
hombre	65-74	43	295
mujer	16-29	46	738
mujer	30-44	89	700
mujer	45-64	169	847
mujer	65-74	51	196

Tabla 4.30: Uso de drogas psicotrópicas en Londres.

		Relaciones Sexuales	
Raza	Sexo	Sí	No
Blanca	Varón	43	134
	Mujer	26	149
Negra	Varón	29	23
	Mujer	22	36

Tabla 4.31: Relaciones sexuales en adolescentes

Salud mental				
Estatus	Buena	Apacible	Moderada	Deteriorada
alto	121	188	112	86
medio	129	246	142	154
bajo	57	168	108	149

Tabla 4.32: Salud mental y nivel socioeconómico

5

Otros Modelos Lineales Generalizados

5.1. Introducción

En este tema vamos a presentar distintos modelos lineales generalizados aplicables a datos de muy diversa índole. Trataremos la regresión de Poisson para conteos que han de ser predichos en función de otras variables. Presentaremos algunos modelos para respuestas que evolucionan a lo largo del tiempo. Por último, estudiaremos cómo tratar la censura en datos de supervivencia, desde la perspectiva de los GLM.

5.2. Regresión de Poisson

Son frecuentes los datos discretos que provienen de conteos de sucesos que se producen por azar con cierta frecuencia y son modelizables en términos de tasas de incidencia que dependen de ciertas variables explicativas.

consultas	quejas	residente	sexo	ingresos	horas
2014	2	si	M	263.03	1287.25
3091	3	no	H	334.94	1588
879	1	si	H	206.42	705.25
1780	1	no	H	226.32	1005.50
3646	11	no	H	288.91	1667.25
2690	1	no	H	275.94	1517.75
1864	2	si	H	295.71	967
2782	6	no	H	224.91	1609.25
3071	9	no	M	249.32	1747.75
1502	3	si	H	269	906.25
2438	2	no	M	225.61	1787.75
2278	2	no	H	212.43	1480.50
2458	5	no	H	211.05	1733.50
2269	2	no	M	213.23	1847.25
2431	7	no	H	257.30	1433
3010	2	si	H	326.49	1520
2234	5	si	H	290.53	1404.75
2906	4	no	H	268.73	1608.50
2043	2	si	H	231.61	1220
3022	7	no	H	241.04	1917.25
2123	5	no	M	238.65	1506.25
1029	1	si	M	287.73	589
3003	3	si	M	280.52	1552.75
2178	2	no	H	237.31	1518
2504	1	si	M	218.70	1793.75
2211	1	no	M	250.01	1548
2338	6	si	H	251.54	1446
3060	2	si	H	270.52	1858.25
2302	1	no	H	247.31	1486.25
1486	1	si	M	277.78	933.75
1863	1	si	H	259.68	1168.25
1661	0	no	H	260.92	877.25
2008	2	no	H	240.22	1387.25
2138	2	no	H	217.49	1312
2556	5	no	H	250.31	1551.50
1451	3	si	M	229.43	973.75
3328	3	si	H	313.48	1638.25
2927	8	no	H	293.47	1668.25
2701	8	no	H	275.40	1652.75
2046	1	si	H	289.56	1029.75
2548	2	si	H	305.67	1127
2592	1	no	H	252.35	1547.25
2741	1	si	M	276.86	1499.25
3763	10	si	H	308.84	1747.50

Tabla 5.1: Quejas en el servicio de urgencias de un hospital

Ejemplo 5.1 (Servicio de Urgencias) En la Tabla 5.1 están recogidos los datos de un estudio sobre 44 médicos que trabajaban en el servicio de urgencias de un hospital grande (Le, 1998, pag.113). El objetivo del estudio era determinar qué variables estaban relacionadas con el número de quejas que recibían los médicos a lo largo de un año. Para ello se controló, para cada médico, el número de consultas que realizó a lo largo del año, y se tuvieron en cuenta otras posibles variables explicativas:

- ‘ingresos’ = salario anual (en dólares por hora)
- ‘horas’ = carga de trabajo (en horas) en el servicio de urgencias,
- ‘sexo’: H=hombre, M =mujer
- ‘residente’, que contiene información sobre si un médico recibió o no entrenamiento como residente en servicios de urgencias (sí/no).

Los elementos básicos para plantear una regresión de Poisson son: una variable dependiente Y basada en conteos, para la que se asume una distribución de Poisson, y unas variables explicativas $\mathbf{x} = (x_1, \dots, x_p)$ que determinan las condiciones específicas para la observación.

Si disponemos de n observaciones independientes de Y , $y_1, \dots, y_n$, cada variable Y_i va a ser predicha en términos de $\mathbf{x}_i = (x_{1i}, \dots, x_{pi})$, para $i = 1, \dots, n$, según un modelo de **regresión de Poisson**, de la forma:

$$\begin{aligned} E(Y_i) &= S_i \lambda(\mathbf{x}_i) \\ &= S_i \exp(\mathbf{x}_i' \beta), \end{aligned} \quad (5.1)$$

donde $\lambda(\mathbf{x}_i)$ es el **riesgo** o tasa de incidencia de los sucesos que contabilizamos por unidad de tiempo. Este modelo relaciona exponencialmente el riesgo con las covariables observadas, o lo que es lo mismo, asume que la respuesta media está ligada con el predictor lineal a través del link logaritmo:

$$\log[E(Y_i)] = \log(S_i) + \mathbf{x}_i' \beta = \log(S_i) + \sum_{j=1}^p x_{ji} \beta_j, \quad i = 1, \dots, n.$$

La variable $\log(S)$ actúa como *offset*, esto es, influye en la respuesta media directamente: un cambio de una unidad en $\log(S)$ provoca un cambio de una unidad en el logaritmo de la respuesta media $\log[E(Y)]$. Sólo se estiman pues, los parámetros β asociados a las covariables $\mathbf{x}$.

La verosimilitud queda pues de la forma:

$$f(\mathbf{y}; \beta) \propto \prod_{i=1}^n [S_i \lambda(\mathbf{x}_i)]^{y_i} \exp\{-S_i \lambda(\mathbf{x}_i)\},$$

y la log-verosimilitud como:

$$l(\mathbf{y}; \beta) \propto \sum_{i=1}^n y_i \log(S_i) + y_i \mathbf{x}_i' \beta - S_i \exp(\mathbf{x}_i' \beta).$$

Las estimaciones de $\beta_1, \dots, \beta_p$ se obtienen por máxima-verosimilitud mediante el algoritmo de mínimos cuadrados ponderados iterados.

Los parámetros β_j se interpretan como un **riesgo relativo** asociado a incrementos en la covariable correspondiente x_j . Efectivamente, dada la relación entre el riesgo λ y las covariables $\mathbf{x}$,

$$\log[\lambda(\mathbf{x})] = \mathbf{x}' \beta = \sum_{j=1}^p x_j \beta_j,$$

si suponemos fijos todos los valores de las covariables $x_2, \dots, x_p$, y consideramos para x_1 un valor concreto $x_1 = x$, y un incremento de una unidad, $x_1 = x + 1$. Tenemos entonces:

$$\begin{aligned} \log[\lambda(x_1 = x, x_2, \dots, x_p)] &= \beta_1 x + \sum_{j=2}^p x_j \beta_j, \\ \log[\lambda(x_1 = x + 1, x_2, \dots, x_p)] &= \beta_1 (x + 1) + \sum_{j=2}^p x_j \beta_j. \end{aligned}$$

Tomando diferencias y después aplicando la función exponencial tenemos:

$$\frac{\lambda(x_1 = x + 1, x_2, \dots, x_p)}{\lambda(x_1 = x, x_2, \dots, x_p)} = \exp\{\beta_1\}.$$

Es decir, $\exp\{\beta_j\}$ representa el **riesgo relativo** sobre la tasa de incidencia de los sucesos, asociado a un incremento de una unidad en la covariable x_j .

La interpretación del riesgo se hace más clara al utilizar la aproximación de la distribución binomial a la Poisson cuando sus parámetros $n \rightarrow \infty$ y $\pi \rightarrow 0$. En ese caso, una $Bin(n, \pi)$ se puede aproximar a una $Po(\theta = n\pi)$. De donde en una $Po(S\lambda)$, λ se puede interpretar como la probabilidad asociada a la ocurrencia de un suceso, y S como el tamaño de la observación (número límite de sucesos).

Ejemplo 5.2 (Continuación del Ejemplo 5.1) *Para los datos del Servicio de Urgencias asumimos que el número de quejas que recibe un médico en el servicio a lo largo de un año, $Y = \text{'quejas'}$, tiene una distribución de Poisson, con media proporcional al número de consultas Z que ha realizado en el servicio:*

$$Y \sim Po(Z\lambda),$$

y que la tasa de quejas que recibe un médico por consulta, λ , está relacionada exponencialmente con las variables explicativas 'ingresos', 'horas', 'sexo' y 'residente', denotadas conjuntamente por $\mathbf{x}$, a través de:

$$\log(\lambda) = \mathbf{x}'\beta.$$

Así, predeciremos el número de quejas Y con una verosimilitud Poisson y link logarítmico, en función de las covariables observadas y del offset $\log(Z) = \text{'lconsultas'}$, de la forma:

$$\log[E(Y)] = \log(Z) + \mathbf{x}'\beta.$$

Son cuestiones de interés si el número de quejas está relacionado con que el médico haya sido o no residente en urgencias; si tienen más quejas (en proporción) los médicos hombres que los médicos mujeres; si es mayor la proporción de quejas cuando el médico lleva más horas de servicio en el turno de guardia; si tiene algo que ver el número de quejas con los ingresos del médico.

Para contestar estas cuestiones procedemos al ajuste de un 'buen' modelo del que extraer conclusiones. Consideramos en primer lugar un modelo 'amplio' con interacciones a lo sumo de orden 2, en el que las covariables ingresos y horas afectan de forma diferente al número de quejas en función de si el médico ha sido o no residente en urgencias, y de si es hombre o mujer, esto es,

```
fit <- glm(quejas ~ residente*sexo*(ingresos+horas),
          family = poisson, offset = lvisits)
```

A pesar de ser un buen modelo (su deviance es de $D = 42,06$ con 32 g.l. y un p -valor de 0.109), no es el más parsimonioso posible. Para reducirlo a un modelo más simple excluyendo términos poco relevantes en la explicación de las quejas utilizamos un procedimiento automático de selección de modelos basados en el AIC,

```
step(fit)
```

que da lugar al modelo Poisson:

$$\text{quejas} \sim \text{residente} + \text{sexo} + \text{horas} + \text{residente} : \text{horas},$$

con un $AIC = 181,54$. Las estimaciones del modelo son:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-9.4057940	0.9938080	-9.464	< 2e-16
residentesi	2.6400635	1.1756064	2.246	0.02472
sexoMujer	-0.3333076	0.2074905	-1.606	0.10819
horas	0.0018860	0.0006118	3.083	0.00205
residentesi:horas	-0.0018110	0.0007471	-2.424	0.01536

Esto es, el número de quejas depende de:

1. si el médico ha sido o no residente (más quejas los residentes en urgencias)
2. si es hombre o mujer (menos quejas las mujeres)
3. el número de horas que trabaja en urgencias (afecta muy poco)
4. y además, el número de horas trabajadas le afecta de forma diferente si ha sido residente o no en urgencias: a los que han sido residentes no les afecta prácticamente nada el número de horas trabajadas, mientras que a los que no lo han sido les influye negativamente: más horas trabajadas, más quejas).

Por ejemplo, para un mismo número de horas trabajadas H , la tasa o riesgo asociado a recibir una queja (por consulta) un médico hombre que ha sido residente es de:

$$\exp\{-9,41 + 2,64 + (0,00189 - 0,001811) \times H\}$$

5.3. GLM's para Datos Longitudinales

5.3.1. Introducción y Ejemplos

Los **datos longitudinales** consisten en observaciones de respuestas a lo largo del tiempo que pueden modelizarse como un proceso estocástico. Difieren de otros tipos de datos en que ha de tenerse en cuenta la dependencia del pasado en las respuestas actuales.

Entre los datos longitudinales, las curvas de crecimiento tienen ciertas características específicas:

- el perfil de crecimiento será generalmente una función no lineal del tiempo, tendiendo con frecuencia a una asíntota;
- por definición, el crecimiento no es estacionario; ocasionalmente, los incrementos (o innovaciones) pueden serlo;
- la variabilidad aleatoria generalmente crece con el tamaño;
- las sucesivas respuestas son medidas sobre el mismo sujeto, de forma que generalmente no serán independientes;
- individuos distintos pueden tener distintas razones de crecimiento (perfil), tanto inherentemente como debido a efectos ambientales.

En principio ignoraremos la posibilidad de dependencia (los dos últimos puntos) y veremos cómo usar las técnicas de regresión estándar para tratar algunas de las características de estas curvas. En realidad, los modelos lineales generalizados que presentamos a continuación pueden resultar útiles para contrastar independencia de las observaciones a través de la comparación con otros modelos más sofisticados y apropiados que sí tengan en cuenta la dependencia entre observaciones consecutivas (ver Lindsey, 1997, caps. 5 y 10).

Veamos unos cuantos ejemplos del tipo de datos de crecimiento cuya modelización plantearemos en términos de modelos lineales generalizados.

Ejemplo 5.3 (Capital Doméstico Bruto) *En economía se presta mucha atención a comparar las razones de crecimiento entre diferentes países. La formación del capital doméstico bruto es un índice importante para medir el crecimiento económico de un país. En la Tabla 5.2 (Oliver, 1970, también en Lindsey, 1997, pag.71) tenemos los datos de este índice en el Reino Unido para el período comprendido entre 1948 y 1967. Su representación gráfica aparece en la Figura 5.1.*

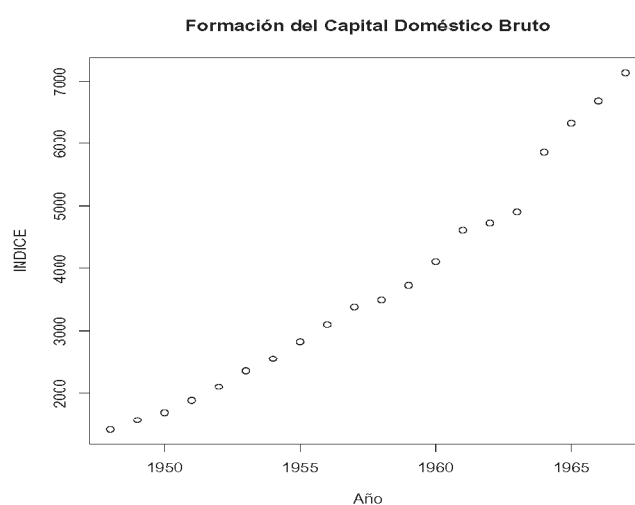


Figura 5.1: Formación del Capital Doméstico Bruto en el Reino Unido en el período comprendido entre 1948 y 1967.

1422	1577	1700	1889	2106	2359	2552	2829	3103	3381
3492	3736	4120	4619	4731	4906	5860	6331	6686	7145

Tabla 5.2: Formación del Capital Doméstico Bruto en el Reino Unido en el período comprendido entre 1948 y 1967.

Ejemplo 5.4 (Evolución del SIDA) *El SIDA tiene un notable impacto en los costes de sanidad. Es muy importante pues, poder proyectar el tamaño de la epidemia en el*

futuro de modo preciso. En la Tabla 5.3 se presenta el número de enfermos de SIDA en USA entre los años 1982 y 1990 (Hay y Wolack, 1994, también en Lindsey, 1997, pag.73). Los datos están representados en la Figura 5.2.

	Cuatrimestre			
1982	185	201	293	381
1983	536	705	769	851
1984	1148	1372	1573	1746
1985	2157	2578	2997	3107
1986	3775	4263	4692	4935
1987	5947	6409	6756	6920
1988	7560	7677	7674	7625
1989	8109	8224	7818	6935
1990	5922	329		

Tabla 5.3: Casos de SIDA diagnosticados en USA, 1982–1990.

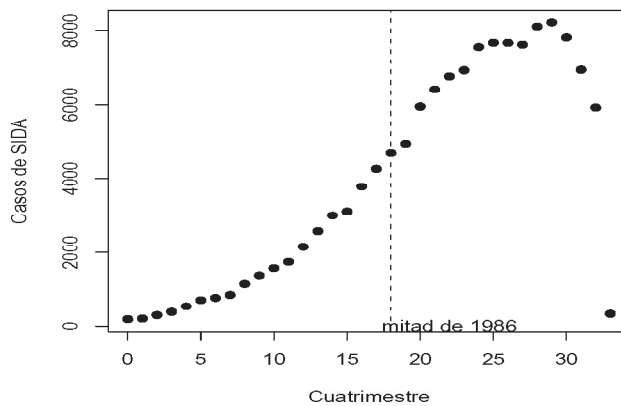


Figura 5.2: Casos de SIDA diagnosticados en USA, 1982–1990.

5.3.2. Curvas Exponenciales de Crecimiento

Generalmente, para describir el comportamiento de este tipo de datos de crecimiento se recurre a curvas exponenciales no lineales. Veamos algunas de ellas,

tratables como modelos lineales generalizados.

5.3.2.1. Respuesta Continua

Una de las curvas más simples de crecimiento de la respuesta Y a medida que transcurre el tiempo t , tiene la forma **exponencial**:

$$Y = \alpha \exp(\beta t).$$

Esta curva suele ser realista únicamente en una primera fase del crecimiento; nada puede continuar creciendo exponencialmente por siempre. Por otro lado, un valor de $\beta < 0$ puede ser un modelo razonable para un declive exponencial con el paso del tiempo.

Puesto que la relación anterior entre la respuesta y el tiempo es determinista, podemos introducir un elemento estocástico (estructura para el error) al menos de dos formas:

1. El logaritmo de la respuesta, $Z = \log(Y)$ puede tener una distribución de la familia exponencial como la normal o la gamma, dando lugar a una distribución log-normal o log-gamma, es decir, $Z \sim Normal$ o $Z \sim Gamma$. Al transformar pues la respuesta con el logaritmo, y utilizar la distribución normal o gamma para predecir $E(Z)$ y la identidad como link, el predictor lineal resulta:

$$E[\log(Y)] = \log(\alpha) + \beta t. \quad (5.2)$$

El **modelo log-normal** es muy utilizado en Econometría.

2. Otra posibilidad es utilizar la respuesta sin transformar, Y , y asumir para ella una distribución normal o gamma con un link logaritmo,

$$\log[E(Y)] = \log(\alpha) + \beta t. \quad (5.3)$$

Ambos modelos se pueden ajustar fácilmente como modelos lineales generalizados; las curvas resultantes pueden diferir bastante significativamente, tanto entre valores observados y ajustados como respecto a las predicciones que proporcionan.

En el primer modelo, la curva de ajuste atraviesa la media geométrica de los datos, mientras que en el segundo atraviesa la media aritmética. Notar que también la variabilidad modelizada por los dos modelos es bastante diferente. Por ejemplo, para el modelo (5.2) con una distribución normal para $Z = \log(Y)$, la varianza de la log-respuesta Z es constante, lo que implica que la varianza de la respuesta Y crece a medida que crece su valor. El segundo modelo (5.3) con una distribución normal para Y implica que el coeficiente de variación (sd/media) es constante. Otras distribuciones conllevarán otras relaciones sobre cómo cambia la varianza con la respuesta media cuando pasa el tiempo.

Ejemplo 5.5 (Continuación del Ejemplo 5.3) Definiendo la variable 'tiempo' de 0 en adelante,

$$\text{tiempo} < -0 : (67 - 48)$$

el ajuste de los datos, en la variable 'capital', con algunos de los modelos propuestos da los siguientes resultados:

1. *Modelo log-normal:* $E[\log(Y)] = \theta + \beta t$, $\log(Y) \sim \text{Normal}$,

$$\text{glm}(\log(\text{capital}) \sim \text{tiempo}, \text{family} = \text{gaussian})$$

$$\hat{\mu}_{\log(Y)} = 7,3110 + 0,0841 t,$$

$$\rightsquigarrow \hat{\alpha} = 1496,67, \text{ con } AIC = 254,3.$$

2. *Modelo Normal con link logaritmo:* $\log[E(Y)] = \theta + \beta t$, $Y \sim \text{Normal}$,

$$\text{glm}(\text{capital} \sim \text{tiempo}, \text{family} = \text{gaussian}(\text{link} = "log"))$$

$$\log(\hat{\mu}_y) = 7,3429 + 0,08144 t,$$

$$\hat{\alpha} = 1545,269, \text{ con } AIC = 256,86.$$

3. *Modelo Gamma con un link logaritmo (dada la forma asimétrica de la distribución de los índices):* $\log[E(Y)] = \theta + \beta t$, $Y \sim \text{Gamma}$,

$$\text{glm}(\text{capital} \sim \text{tiempo}, \text{family} = \text{Gamma}(\text{link} = "log"))$$

$$\log(\hat{\mu}_y) = 7,3117 + 0,0841 t,$$

$$\rightsquigarrow \hat{\alpha} = 1497,760, \text{ con } AIC = 254,31.$$

Las pequeñas diferencias en la estimación de los parámetros del predictor lineal se traducen en diferencias a veces importantes en la predicción de la respuesta. Así por ejemplo, la predicción que dan estos tres modelos para el año 1968 son, respectivamente, de 8046.72, 7877.558 y 8051.06. El valor finalmente observado ese año fue de 7734, lo que indica que el modelo normal con link logaritmo fue el mejor predictor para ese año, a pesar de que su AIC quedó por encima de los AIC de los otros modelos propuestos.

Los tres ajustes están representados en la Figura 5.5, y si bien ajustan los datos de forma muy similar, observamos que pueden dar resultados muy distintos entre sí en un rango de predicción diferente al utilizado en el ajuste.

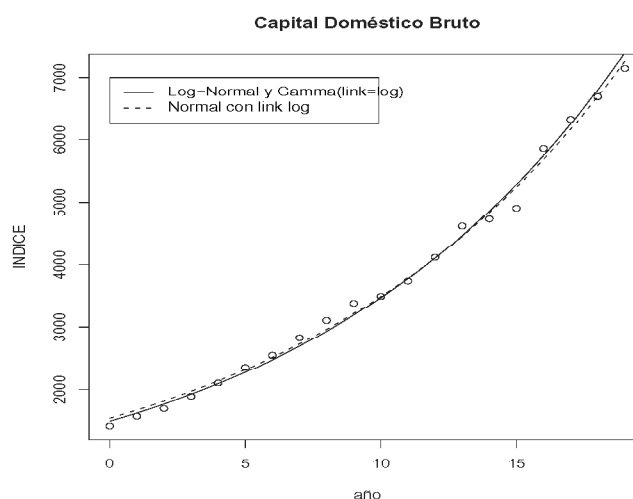


Figura 5.3: Ajuste de los datos de Capital Doméstico Bruto en Reino Unido.

5.3.2.2. Respuesta Discreta: conteos

Cuando tenemos datos de frecuencias y es razonable asumir una distribución de Poisson para la respuesta, el link canónico (logaritmo) permite ajustar un modelo de **crecimiento exponencial** de la forma:

$$E(Y) = \alpha e^{\beta t}.$$

Ejemplo 5.6 (Continuación del Ejemplo 5.4) Ajustamos una regresión de Poisson sobre los casos de SIDA entre 1982 hasta mitad de 1986: los 18 primeros datos, que componen la zona de crecimiento donde se aprecia una clara tendencia exponencial. Definimos la variable temporal 'tiempo' con valores de 0 en adelante, y la variable 'sida' con los 18 datos de sida mencionados:

```
tiempo <- 0 : (length(sida) - 1)
glm(sida ~ tiempo, family = poisson)
```

proporciona:

$$\log[\hat{\mu}] = 5,6174 + 0,1658 \text{ tiempo} ,$$

con $\mu = E(Y)$ y un $AIC = 421,07$. El modelo no es realista y ajusta mal los datos restantes: el crecimiento sería continuamente exponencial, y eso no es así. En la Figura 5.4 está representado el ajuste.

5.3.3. Curva Logística de Crecimiento

Una posibilidad para modelizar los casos en que el crecimiento no es continuamente exponencial, es utilizar la **curva logística** simétrica con forma de S, que tiene la forma determinista:

$$Y = K \frac{\alpha e^{\beta t}}{1 + \alpha e^{\beta t}},$$

donde K es el máximo valor asintótico de la respuesta Y .

Es posible ajustar un modelo lineal generalizado que respete esta tendencia de crecimiento. El modelo que resulta viene dado por:

$$\log \left[\frac{E(Y)}{K - E(Y)} \right] = \theta + \beta t, \quad (5.4)$$

con $\theta = \log(\alpha)$. Es decir, si K fuera conocido, considerando una estructura de datos binarios (éxitos= Y , fracasos= $K - Y$) y utilizando el link logit, se podrían obtener las estimaciones de los parámetros α y β . Estos modelos **no** permiten que un valor de la respuesta Y llegue a ser mayor que K , ni siquiera por azar.

Cuando la asíntota K es desconocida, se puede encontrar ajustando el modelo para varios valores de $K > \max(Y_i)$, siguiendo algún patrón de búsqueda, y utilizando aquel valor para el cual la verosimilitud del modelo ajustado es máxima. Para encontrar el K de máxima verosimilitud es necesario evaluar la verosimilitud completa ya que K está contenido en el número combinatorio que no se incluye en la verosimilitud usual.

Si $\mathbf{y} = (y_1, \dots, y_n)$ son los valores observados de la variable respuesta en los instantes de tiempo $t_1, \dots, t_n$, asumimos $Y_i \stackrel{\text{ind}}{\sim} Bi(K, \pi_i)$, para $i = 1, \dots, n$, que responden al modelo logístico de crecimiento para un valor concreto de K , entonces la verosimilitud tiene la forma:

$$f(\mathbf{y}|K, \hat{\pi}) = \prod_{i=1}^n \binom{K}{y_i} \pi_i^{y_i} (1 - \pi_i)^{K-y_i}, \quad (5.5)$$

$$\text{con } \pi_i = \frac{\exp\{\log(\alpha) + \beta t_i\}}{1 + \exp\{\log(\alpha) + \beta t_i\}}.$$

El procedimiento iterativo para localizar el valor máximo-verosímil de K es sencillo. Inicialmente se considera un grid de puntos para K , en cada punto se ajusta el modelo logístico (5.4) y se evalúa la verosimilitud (5.5) en el ajuste. Se elige finalmente el K en el grid que da mayor verosimilitud.

Una desventaja importante de la curva logística es que es simétrica, de forma que la curva inferior será simétrica a la superior, y ese fenómeno no ocurre en muchos fenómenos biológicos.

Ejemplo 5.7 (Continuación del Ejemplo 5.4) *Utilizando los datos de sida hasta mitad de 1986 (los mismos utilizados en el ajuste de la curva exponencial), conseguimos un ajuste con máxima verosimilitud para $K = 7880$.*

La forma de realizar el ajuste en $\mathcal{R}$ para un K fijo es:

```
sida.bin <- cbind(sida, rep(K, length(sida)) - sida)
glm(sida.bin ~ tiempo, family = binomial(link = "logit"))
```


El modelo ajustado resulta para $t = \text{tiempo}$:

$$\hat{\eta}_t = \log[\hat{\mu}/(K - \hat{\mu})] = \hat{\theta} + \hat{\beta}t = -3,58 + 0,021 t,$$

con un $AIC = 260,28$, de donde la curva de predicción viene dada por:

$$E(Y) = 7880 \frac{\exp(-3,58 + 0,021 t)}{1 + \exp(-3,58 + 0,021 t)}.$$

El ajuste de la curva está representado en la Figura 5.4. Como apreciamos, capta bastante bien la tendencia futura de los datos, a pesar de estar hecho el ajuste sólo con los datos de tendencia exponencial.

Este modelo sigue siendo poco razonable debido a que en el ajuste sólo se ha utilizado una parte de los datos (justo aquellos en los que la tendencia era exclusivamente exponencial). Al utilizar todos los datos, incluso los que dan información sobre la localización de la asíntota, el ajuste ha de mejorar, especialmente para la predicción por encima del rango observado. Si ya ha habido estabilización del crecimiento, iniciada por algunos datos disponibles, la predicción será mejor.

5.3.4. Curva de Crecimiento de Gompertz

Otra curva de crecimiento con asíntota que se utiliza con frecuencia es la **curva de Gompertz**, que tiene la forma:

$$Y = K(1 - e^{-\alpha e^{\beta t}}),$$

donde K es el máximo valor desconocido de una asíntota. De nuevo podemos obtener una estructura lineal utilizando un link complementario log-log:

$$\log \left[-\log \left(\frac{K - Y}{K} \right) \right] = \theta + \beta t, \quad (5.6)$$

con $\theta = \log(\alpha)$.

Para el ajuste se puede utilizar el mismo procedimiento iterativo que en la curva logística (la verosimilitud es la misma: la binomial). En realidad, los modelos (5.4) y (5.6) sólo difieren en el link que utilizan.

Ejemplo 5.8 (Continuación del Ejemplo 5.4) De nuevo, utilizando los datos hasta mitad de 1986 y probando el ajuste de datos binarios con el link cloglog para varios valores de K , conseguimos máxima verosimilitud para $K = 5820$. El modelo ajustado resulta:

$$\boxed{\text{glm}(\text{sida.bin} \sim \text{tiempo}, \text{family} = \text{binomial}(\text{link} = \text{"cloglog"}))}$$

$$\hat{\eta}_t = \text{cloglog}[E(Y_{K=5820})] = -3,23 + 0,21 t,$$

con un $AIC = 287,01$.

El ajuste de la curva está representado en la Figura 5.4. Todas las curvas, aun siendo muy diferentes entre sí, ajustan razonablemente bien las observaciones durante el período de tiempo utilizado para el ajuste (tendencia exponencial). Vemos que es difícil o imposible predecir bien la parte superior de la curva con sólo la información de la parte inferior. De hecho, la predicción fuera del rango de valores observados varía mucho en función del modelo ajustado. Hay otros modelos más complejos que proporcionan ajustes basados en información adicional sobre los retardos de los datos y hacen uso de todos los datos (ver Lindsey, 1997, caps. 5 y 10).

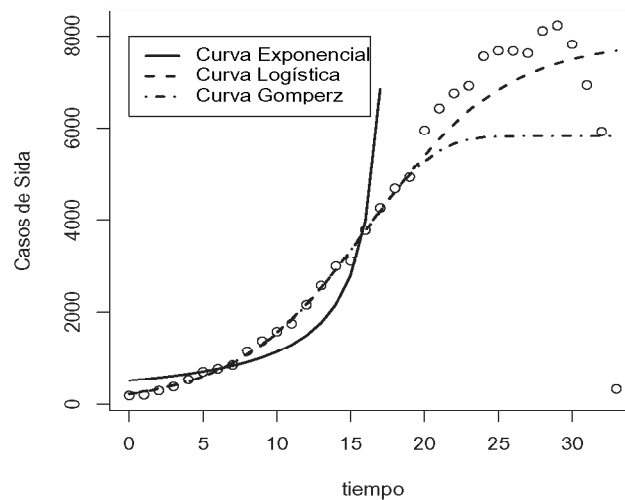


Figura 5.4: Ajuste de curvas de crecimiento para los Casos de Sida en USA.

5.4. GLM's para Datos de Supervivencia Censurados

5.4.1. Conceptos Generales: Distribuciones asimétricas y Censura

Una **duración** es el tiempo que transcurre hasta que ocurre algún suceso (muerte, fallo, remisión de una enfermedad,). La respuesta de una duración es una variable aleatoria no-negativa.

Cuando se toman datos sobre **tiempos de supervivencia** en un conjunto de sujetos, se termina la observación cuando

- acontece el suceso (el sujeto muere, la máquina falla, la enfermedad remite,),
- o termina el tiempo de observación fijado en el estudio.

Usualmente la distribución de duraciones no será simétrica; de hecho, estos tiempos son siempre datos positivos y la probabilidad de supervivencia disminuye con el paso del tiempo. Una distribución normal no será entonces adecuada para modelizar este tipo de datos. Distribuciones apropiadas dentro de la familia exponencial son la log-normal, la gamma y la Gausiana inversa.

Ejemplo 5.9 (Remisión de leucemia) En la Tabla 5.4 se presentan los tiempos de supervivencia en semanas a leucemia aguda bajo un tratamiento y un placebo (Gehan, 1965, también en Lindsey, pag. 112). Los datos en la tabla corresponden a un estudio realizado en un período de observación de un año; el límite superior para un tiempo observado es de 35 semanas. Las observaciones censuradas son aquellas que tienen un asterisco.

tratamiento	6,6,6,6*,7,9*,10,10*,11*,13,16,17*,19*,20*,22,23,25*,32*,32*,34*,35*
placebo	1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23

Tabla 5.4: Tiempos de supervivencia (en semanas) a leucemia aguda bajo dos tratamientos.

La Figura 5.5 contiene un histograma de los tiempos de supervivencia. Como se puede apreciar, su distribución es bastante asimétrica, hecho bastante frecuente en datos de supervivencia.

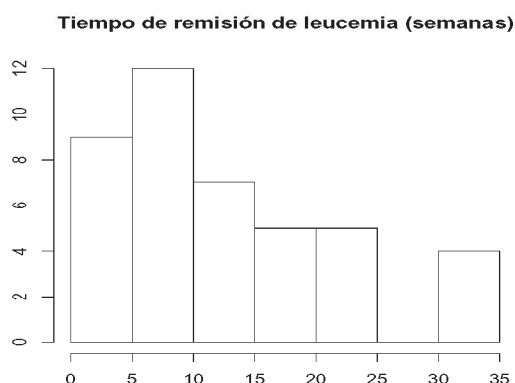


Figura 5.5: Tiempos de supervivencia (en semanas) a leucemia aguda bajo dos tratamientos.

Dado que los individuos van a ser observados a lo largo del tiempo hasta que ocurra el suceso (por ejemplo, que muera), y dado que el tiempo es limitado y costoso, no será posible a veces seguir a todos los individuos hasta el final. Se tendrán entonces **datos censurados**. Las observaciones censuradas son observaciones incompletas, pero sin embargo aún contienen información importante. Se sabe que al final del período de observación, un individuo con dato censurado es un individuo que no ha muerto, es decir, un individuo cuyo su tiempo de vida supera al tiempo de vida observado para él.

La censura puede ocurrir por varias razones. Por ejemplo, el protocolo de un estudio puede especificar observar durante un periodo fijo de tiempo, o los casos individuales pueden desaparecer del estudio por alguna razón. Las formas más comunes de censura son:

Censura por tiempo: si se deja de observar a los individuos una vez que ha transcurrido un intervalo fijo de tiempo. En consecuencia, no se sabe si actualmente siguen con vida o no. Es frecuente en estudios médicos.

Censura por fallo: si a lo largo del período de estudio se “pierden” casos (pacientes que dejan de asistir a las revisiones, etc.), a pesar de que el estudio prosiga hasta conseguir información completa sobre un número fijo de casos. Es común en tests industriales, donde para modelizar el proceso de fallos se observa, en cada máquina, el tiempo transcurrido hasta que tienen un fallo; el objetivo en estos

tests es predecir cada cuanto tiempo acontece un fallo.

Sin embargo, los casos pueden desaparecer por razones no relacionadas con el estudio o más allá del control del investigador. Esto puede estar o no ligado a la respuesta o a las variables explicativas (por ejemplo, testando un fármaco, un paciente decide suspender el tratamiento debido a los efectos secundarios que le ha provocado). Si están relacionados, la forma en que la censura ocurre no se puede ignorar en el modelo.

Por lo tanto, en casos complejos en los que la censura no es aleatoria, es necesario construir un modelo para ella, a pesar de que generalmente hay muy poca información disponible en los datos sobre el modelo a utilizar. Cuando la censura sólo depende de la información ya disponible en el punto de censura, a veces es factible la modelización a través de algún modelo de regresión con las variables explicativas disponibles. Veremos cómo modelizar algún tipo de datos censurados mediante modelos lineales generalizados. Incluso con causas ignorables de censura, el análisis es demasiado complicado porque no podemos utilizar simplemente la función de densidad tal y como es. Introducimos a continuación las funciones de probabilidad básicas en el ajuste de datos de supervivencia; los modelos que proponemos vienen en términos de estas funciones de probabilidad.

5.4.2. Funciones de probabilidad para Tiempos de Supervivencia

Si la función de densidad de probabilidad para T = tiempo de supervivencia, es $f(t)$ y la función de probabilidad acumulada es $F(t)$, entonces la **función de supervivencia** $S(t)$ se define como la probabilidad de sobrevivir más de t , esto es,

$$S(t) = Pr(T > t) = \int_t^{\infty} f(t)dt = 1 - F(t) \quad (5.7)$$

y la **función hazard** $h(t)$, que mide el riesgo instantáneo (de morir), como:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d \log S(t)}{dt}. \quad (5.8)$$

De hecho, $h(t)dt$ representa la **intensidad del proceso de muertes**, o lo que es lo mismo, la probabilidad de morir en un intervalo pequeño de tiempo dt dado que el individuo ha sobrevivido hasta el instante t .

Una distribución para tiempos de supervivencia ha de tener una función hazard con *buenas propiedades*; por ejemplo, es de esperar que la función hazard no decrezca con t , esto es, a más tiempo transcurrido, mayor riesgo de muerte.

Puesto que $S(0) = 1$, tenemos que:

$$S(t) = \exp \left\{ - \int_0^t h(u) du \right\}, \quad (5.9)$$

de donde surge la definición de la **función hazard acumulada** $H(t)$,

$$H(t) = \int_0^t h(u) du = -\log S(t), \quad (5.10)$$

y la reexpresión de la función de densidad para t , en términos de la función hazard:

$$f(t) = h(t) \exp \left\{ - \int_0^t h(u) du \right\} \quad (5.11)$$

La **esperanza de vida** se calcula pues, como:

$$\mu = \int_0^\infty u f(u) du = \int_0^\infty S(u) du,$$

igualdad que resulta del hecho de que $-f(t) = S'(t)$.

La función hazard depende, en general, tanto del tiempo como de las covariables controladas en el experimento, algunas de las cuales son también dependientes del tiempo. De cómo esté relacionada la función hazard con las covariables y éstas con el tiempo surgen distintos modelos, algunos de los cuales podrán ser formulados como modelos lineales generalizados.

5.4.3. Modelización de la Censura

Definamos una variable aleatoria I^c para distinguir observaciones completas de datos censurados:

$$I_i^c = \begin{cases} 1 & \text{si la observación } i \text{ se observó completamente} \\ 0 & \text{si la observación } i \text{ está censurada.} \end{cases}$$

Si asumimos que el tiempo t se observa de modo continuo, la verosimilitud para una muestra de n observaciones independientes de tiempos de supervivencia será proporcional a:

$$\prod_{i=1}^n f(t_i)^{I_i^c} S(t_i)^{1-I_i^c}. \quad (5.12)$$

5.4.3.1. Modelos Hazard Proporcionales

En general, la función hazard depende del tiempo t y de las covariables $\mathbf{x}$ controladas en el experimento. El **modelo hazard proporcional** separa ambas componentes de la forma:

$$h(t; \mathbf{x}) = h_0(t) \exp(\mathbf{x}'\beta), \quad (5.13)$$

donde $h_0(t)$ es la **función hazard base** y describe el riesgo para individuos con $\mathbf{x} = 0$; $\exp(\mathbf{x}'\beta)$ es el **riesgo relativo**, y representa la reducción o incremento del riesgo asociada con el conjunto de características $\mathbf{x}$, que es similar para todas las duraciones t . La exponencialidad está justificada por la necesidad de positividad y la hipótesis de un efecto multiplicativo de las covariables en los hazards. Este modelo implica que el cociente de hazards de dos individuos depende de la diferencia entre sus predictores lineales $\eta = \mathbf{x}'\beta$ en cualquier instante t .

Distintas elecciones de la función hazard base dan lugar a distintas hipótesis sobre la distribución de los tiempos de supervivencia:

- $h_0(t) = 1$ constante $\rightsquigarrow$ distrib. exponencial
- $h_0(t) = \alpha t^{\alpha-1}$ $\rightsquigarrow$ distrib. Weibull
- $h_0(t)$ no está especificado a la izquierda, de forma que hay que usar un factor para el tiempo $\rightsquigarrow$ modelo de Cox semiparamétrico.

Tomando logaritmos en (5.13), se consigue un modelo aditivo:

$$\log h(t; \mathbf{x}) = \log h_0(t) + \mathbf{x}'\beta. \quad (5.14)$$

Estos modelos hazard proporcionales se pueden ajustar de un modo sencillo como modelos lineales generalizados basados en la equivalencia con la representación de Poisson, tal y como se comenta a continuación.

5.4.3.2. Representación de Poisson

Aitkin y Clayton (1980) demostraron una útil relación entre los modelos hazard proporcionales (5.13) y la distribución de Poisson. Tal relación permite ajustar estos modelos como modelos lineales generalizados.

Si $h_0(\cdot)$ es el hazard base, para un modelo hazard proporcional se tiene, utilizando (5.9) y (5.8) respectivamente,

$$S(t_i) = \exp\{-H_0(t_i)e^{x_i'\beta}\} \quad (5.15)$$

$$f(t_i) = h(t_i)S(t_i) = h_0(t_i) \exp\{x_i'\beta - H_0(t_i)e^{x_i'\beta}\}, \quad (5.16)$$

donde

$$H_0(t) = \int_0^t h_0(u)du,$$

es el **hazard base acumulado**.

Entonces, de la ecuación (5.12), que contiene la información en la muestra sobre la censura y los tiempos de supervivencia, la función de verosimilitud queda:

$$\begin{aligned} L(\beta) &= \prod_i f(t_i)^{I_i^c} S(t_i)^{1-I_i^c} \\ &= \prod_i [h_0(t_i) \exp\{x_i'\beta\}]^{I_i^c} \exp\{-H_0(t_i)e^{x_i'\beta}\} \\ &= \prod_i \mu_i^{I_i^c} e^{-\mu_i} \left[\frac{h_0(t_i)}{H_0(t_i)} \right]^{I_i^c}, \end{aligned} \quad (5.17)$$

con

$$\mu_i = H_0(t_i) \exp(x_i'\beta). \quad (5.18)$$

Si la variable I_i^c se interpreta como el número acumulado de sucesos (muertes, remisión, ...) en las observaciones completas, transcurrido un tiempo t_i para el individuo i (que será cero o uno), se puede asumir una distribución de Poisson para ella. Si la función de verosimilitud conjunta a la que dan lugar $I_1^c, \dots, I_n^c$ viene expresada por la ecuación anterior (5.17), se tendrá que $\mu_i = E(I_i^c)$.

A partir (5.18), tomando logaritmos, se tiene que el modelo (5.13) será equivalente al modelo lineal generalizado:

$$\log(\mu_i) = \log[H_0(t_i)] + \mathbf{x}_i'\beta, \quad (5.19)$$

con distribuciones de Poisson independientes para las variables de censura I_i^c , de medias μ_i que dependen de t y de las covariables a través de (5.18).

Para ajustar un modelo (5.19) se asumirán valores fijos de los parámetros en $H_0(t)$, con lo que $\log[H_0(t_i)]$ será incorporado al predictor lineal como un *offset*, esto es, un término con coeficiente conocido e igual a 1. Para estimar el modelo es preciso iterar sobre todos los parámetros desconocidos en $H_0(t)$.

A continuación se proponen varias alternativas a asumir sobre la forma de la función hazard base, esto es, sobre la distribución de los tiempos de supervivencia, y los modelos lineales generalizados a que dan lugar.

5.4.3.3. Distribución Exponencial

Si asumimos:

$$h_0(t) = 1 \rightsquigarrow H_0(t) = t,$$

los únicos parámetros a estimar en el modelo hazard proporcional (5.13) son los de relación con las covariables $\mathbf{x}$, β , y la distribución para los tiempos de supervivencia es la exponencial, con función de densidad:

$$f(t_i) = \exp(x_i' \beta - t_i e^{x_i' \beta}).$$

El tiempo esperado de supervivencia para el individuo i viene dado por:

$$E(T_i) = \exp(-x_i' \beta). \quad (5.20)$$

Si asumimos una distribución de Poisson para la variable de censura, $I_i^c \sim Po(\mu_i)$, el modelo hazard proporcional con $h_0(t) = 1$ es equivalente al modelo log-lineal dado por:

$$\log(\mu_i) = \log(t_i) + x_i' \beta,$$

con $\log(t_i)$ como *offset*.

Ejemplo 5.10 (Continuación del Ejemplo 5.9) *Introducidos los tiempos de supervivencia en la variable 'tiempo' (todos los tratamientos y a continuación todos los controles), construimos la variable de censura*

`censura < -c(1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, rep(1, 21))`

junto con un factor 'grupo' que identifica los tratamientos de los controles. Ajustamos un modelo exponencial para predecir la censura y a través de ella los tiempos de supervivencia:

$$\text{fit} <- \text{glm}(\text{censura} \sim \text{grupo}, \text{family} = \text{poisson}, \text{offset} = \log(\text{tiempo}))$$

La deviance de este modelo es de $D = 38,017$ con 40 g.l., y el p -valor asociado a la bondad del ajuste resulta de $p = 0,56$: el modelo es bueno.

El modelo ajustado resulta:

$$f(t_i) = \exp\{-3,69 + 1,53I_i^T - t_i e^{-3,69+1,53I_i^T}\},$$

con $I_i^T = 1$ si la observación es del grupo tratamiento y 0 si es del placebo.

El valor esperado de los tiempos de supervivencia dado por (5.20), se calcula con la instrucción

$$\text{tiempo}/\text{fitted}(\text{fit})$$

El tiempo medio de supervivencia a la leucemia aguda en el grupo tratamiento es de 39,9 semanas, mientras que en el grupo control es tan sólo de 8,67. El tratamiento proporciona mayor esperanza de vida. La diferencia entre tratamientos y controles en términos del logaritmo del riesgo es de 1,53.

5.4.3.4. Distribución Weibull

La distribución Weibull para los tiempos de supervivencia surge del modelo hazard proporcional (5.13) y la función hazard base:

$$h_0(t) = \alpha t^{\alpha-1} \rightsquigarrow H_0(t) = t^\alpha.$$

La función de densidad para el tiempo de supervivencia viene dada entonces por:

$$f(t_i) = \alpha t_i^{\alpha-1} \exp\{\mathbf{x}_i' \beta - t_i e^{\mathbf{x}_i' \beta}\},$$

con $E(T_i) = \Gamma(1 + 1/\alpha) \exp\{-\mathbf{x}_i' \beta / \alpha\}$.

Considerando la representación de Poisson para las variables de censura I_i^c , el modelo log-lineal equivalente viene dado por:

$$\log(\mu_i) = \alpha \log(t_i) + \mathbf{x}_i' \beta, \quad (5.21)$$

para un valor fijo de α , estimable por máxima-verosimilitud. Así, $\alpha \log(t_i)$ actúa como *offset*.

Por otro lado, la log-verosimilitud completa es:

$$\log(L) = \sum_i I_i \log(\alpha) + \sum_i [I_i \log(\mu_i) - \mu_i].$$

Si estimamos α por máxima-verosimilitud, igualaremos su función *score* a cero,

$$\frac{\partial \log(L)}{\partial \alpha} = \frac{\sum_i I_i}{\alpha} + \sum_i (I_i - \mu_i) \log(t_i),$$

de donde obtenemos:

$$\hat{\alpha} = \frac{\sum_i I_i}{\sum_i (\hat{\mu}_i - I_i) \log(t_i)}, \quad (5.22)$$

que se puede utilizar para actualizar el valor de α en iteraciones sucesivas hasta conseguir finalmente el α máximo-verosímil y el ajuste definitivo. De hecho, en el procedimiento de estimación se puede empezar con $\alpha = 1$, que representa la distribución exponencial; una vez obtenido el ajuste del modelo (5.21) con dicho valor de α , se estima β y se actualiza el valor de α con (5.22); a continuación se vuelve de nuevo al ajuste del modelo log-lineal, iterando hasta lograr la convergencia.

5.4.3.5. Distribución del valor extremo

Asumiendo una función hazard de la forma:

$$h_0(t) = \alpha e^{\alpha t}, \rightsquigarrow H_0(t) = e^{\alpha t},$$

se obtiene la distribución del valor extremo para los tiempos de supervivencia, con función de densidad:

$$f(t) = \alpha e^{\alpha t} \exp\{\mathbf{x}_i' \beta - e^{\alpha t + \mathbf{x}_i' \beta}\}.$$

La transformación $u = \exp(t)$ convierte la distribución del valor extremo en la distribución de Weibull. Por tanto, para estimar β bajo esta distribución bastará con reemplazar t por u en el procedimiento de estimación para la Weibull.

5.5. Ejercicios

1. McCullag y Nelder (1989, pag.206) analizan datos proporcionados por *Lloyd's Register of Shipping*, concernientes a un tipo de daño causado por las olas en un componente de ciertos buques cargueros. En la Tabla 5.5 se presenta un subconjunto de estos datos, restringidos al período de operación 1975-79. Los datos consisten en los meses de servicio, el número de incidentes, el período de construcción y el tipo de carguero.

Trata el número de incidentes como aproximadamente Poisson, con media igual al producto del tiempo de servicio y una razón de daño λ que depende del tipo de carguero y del año de construcción.

¿Puedes simplificar el modelo eliminando una de las variables? ¿Y tratando el año de construcción como un efecto lineal? ¿Puedes sugerir alguna otra simplificación? ¿Existe relación entre el número de incidentes y el año de construcción del carguero?

Tipo Barco	Año Construc.	Meses Servicio	Accidentes con daños
A	1960-64	63	0
A	1968-70	1095	4
A	1970-74	3353	18
A	1975-79	2244	11
B	1960-64	17176	29
B	1968-70	20370	53
B	1970-74	13099	44
B	1975-79	7117	18
C	1960-64	552	1
C	1968-70	676	1
C	1970-74	1948	28
C	1975-79	274	1
D	1960-64	105	0
D	1968-70	192	0
D	1970-74	1208	11
D	1975-79	2051	4
E	1968-70	437	7
E	1970-74	2161	12
E	1975-79	542	1

Tabla 5.5: Número de incidentes con daños en buques cargueros, meses de servicio y año de construcción.

2. En la Tabla 5.6 tienes los datos correspondientes a un pequeño experimento en el que 7 de un total de 16 ratones fueron aleatoriamente seleccionados para recibir un nuevo tratamiento médico. Los 9 ratones restantes fueron asignados a un grupo control en el que no se administró ningún tipo de tratamiento. El objetivo del tratamiento era prolongar el tiempo de supervivencia después de una operación quirúrgica. La tabla muestra los tiempos de supervivencia (en días) tras la operación para los 16 ratones.

El tratamiento, ¿prolongaba la vida de los ratones tras la operación?

tratamiento	94, 38, 23, 197, 99, 16, 141
control	52, 104, 146, 10, 51, 30, 40, 27, 46

Tabla 5.6: Tiempos de supervivencia (en días) tras una operación.

3. Los datos de la Tabla 5.7 (Cox y Snell, 1981, Ex.U, también en Dobson, 1995, pag.47) son: '*tiem.sup*', que es el tiempo hasta la muerte (en semanas) desde la diagnosis de leucemia, y '*log.cel.cont*', igual al $\log_{10}$ del número inicial de células blancas en la sangre. Se dispone de datos de diecisiete pacientes que sufrían de leucemia.

¿Cuáles son tus conclusiones sobre la utilidad de saber el número de células blancas en la sangre para predecir el tiempo de supervivencia? ¿Cuál es la probabilidad de sobrevivir de un paciente con 5000 células blancas en la sangre?

tiem.sup	log.cel.cont	tiem.sup	log.cel.cont
65	3.36	4	4.23
156	2.88	39	3.73
100	3.63	143	3.85
134	3.41	56	3.97
16	3.78	26	4.51
108	4.02	22	4.54
121	4.00	1	5.00
5	4.72	1	5.00
65	5.00		

Tabla 5.7: Tiempos de supervivencia en semanas para 17 pacientes de leucemia

judia	0.745,1.047,1.695,2.428,3.664,4.022,5.447,6.993,8.221,8.829, 10.08,12.971,14.808,17.341, 19.705,22.597,24.537,25.869, 27.816,29.595,30.451,30.817,32.472,32.999,33.555,34.682,35.041, 35.356, 35.919,36.058,36.454,36.849,37.2,37.2,37.2
-------	---

Tabla 5.8: Crecimiento de una hoja de judía a lo largo de 36 días.

4. Se midió, durante 36 días consecutivos, el área (no se conocen las unidades) de una hoja de la planta de la judía con el fin de describir su crecimiento (Scallan, 1985, también en Lindsey, 1997, pag. 82). Las mediciones comenzaron el día 17, y se presentan en la Tabla 5.8 Ajusta una curva de crecimiento apropiada a estos datos.
5. En la Tabla 5.9 tenemos las alturas (cm) de dos girasoles medidas a intervalos de una semana (Sandland y McGilchrist, 1979, también en Lindsey, 1997, pag. 82). Ajusta una curva de crecimiento para describir estos datos. ¿Hay diferencia entre las dos plantas, debidas a diferencias en el tamaño inicial o a la velocidad de crecimiento?

girasol.1	4,4.5,7,10,16,23,32,47,77,130,148,153,154,154
girasol.2	2,3,5,9,13,19,31,51,78,91,93,96,91,93

Tabla 5.9: Alturas de dos girasoles medidas cada semana.

6. Disponemos de los pesos (g) de dos pikas afganas preñadas, a lo largo de 14 periodos de tiempo igualmente espaciados entre sí desde la concepción al parto (Sandland y McGilchrist, 1979, también en Lindsey, 1997, pag. 82). Ajusta una curva de crecimiento para describir estos datos, en la Tabla 5.10. ¿Puedes detectar alguna diferencia entre los dos animales?

pika.1	251,254,267,267,274,286,298,295,307,318,341,342,367,370
pika.2	258,263,269,266,282,289,295,308,338,350,359,382,390,400

Tabla 5.10: Pesos (g) de dos pikas afganas preñadas

7. Un total de 90 pacientes que sufrían de cancer gástrico fueron asignados aleatoriamente a dos grupos. Un grupo fue tratado con quimioterapia y radiación, mientras que el otro sólo recibió quimioterapia. En la Tabla 5.11 (Gamerman, 1991, también en Lindsey, 1997, pag.118) se presentan los tiempos de

supervivencia (en días) de los pacientes en el estudio (*c* significa que el dato está censurado: el paciente no había muerto cuando se hizo el análisis de los datos).

¿Existe evidencia para concluir que la radiación prolonga el tiempo de supervivencia?

Quimioterapia + Radioterapia				Quimioterapia			
17	167	315	1174c	1	356	524	977
42	170	401	1214	63	358	535	1245
44	183	445	1232c	105	380	562	1271
48	185	464	1366	125	383	569	1420
60	193	484	1455c	182	383	675	1460c
72	195	528	1585c	216	388	676	1516c
74	197	542	1622c	250	394	748	1551
95	208	567	1626c	262	408	778	1690c
103	234	577	1736c	301	460	786	1694
108	235	580		301	489	797	
122	254	795		342	499	955	
144	307	855		354	523	968	

Tabla 5.11: Estudio sobre la efectividad de la radiación para prolongar la vida de pacientes con cáncer gástrico

8. Cincuenta patos negros hembras, *Anas rubripes*, en dos lugares diferentes en New Jersey (USA) fueron capturados por el *U.S. Fish and Wildlife Service* a lo largo de un período de cuatro semanas, desde el 8 de noviembre hasta el 14 de diciembre de 1983. Los patos fueron marcados con pequeños aparatos emisores y liberados al final del año. De estos patos, 31 habían nacido durante ese mismo año (age 0) y 19 en el año anterior (age 1). El peso (g.) y la longitud de las alas (mm.) de cada pato fue anotada. Usualmente, estos datos se usan para calcular un índice de condición: el cociente del peso por la longitud de las alas. El estatus de cada pájaro fue anotado cada día hasta el 15 de febrero de 1984, utilizando diferentes antenas dispersadas. Los tiempos de supervivencia aparecen en la Tabla 5.12 (Pollock et al, 1989, también en Lindsey, 1997, pag.119). La *c* identifica datos censurados.

¿Qué variables influyen sobre el tiempo de supervivencia?

Edad	Peso	Ala	Tiempo	Edad	Peso	Ala	Tiempo
1	1160	277	2	0	1040	255	44
0	1140	266	6c	0	1130	268	49c
1	1260	280	6c	1	1320	285	54c
0	1160	264	7	0	1180	259	56c
1	1080	267	13	0	1070	267	56c
0	1120	262	14c	1	1260	269	57c
1	1140	277	16c	0	1270	276	57c
1	1200	283	16	0	1080	260	57c
1	1100	264	17c	1	1110	270	58c
1	1420	270	17	0	1150	271	63c
1	1120	272	20c	0	1030	265	63c
1	1110	271	21	0	1160	275	63c
0	1070	268	22	0	1180	263	63c
0	940	252	26	0	1050	271	63c
0	1240	271	26	1	1280	281	63c
0	1120	265	27	0	1050	275	63c
1	1340	275	28c	0	1160	266	63c
0	1010	272	29	0	1150	263	63c
0	1040	270	32	1	1270	270	63c
1	1250	276	32c	1	1370	275	63c
0	1200	276	34	1	1220	265	63c
0	1280	270	34	0	1220	268	63c
0	1250	272	37	0	1140	262	63c
0	1090	275	40	0	1140	270	63c
1	1050	275	41	0	1120	274	63c

Tabla 5.12: Tiempos de supervivencia de 50 patos negros hembras en New Jersey

6

Modelos de Cuasi–Verosimilitud

6.1. Introducción y Ejemplos

A menudo no hay información sobre el mecanismo aleatorio que ha dado lugar a los datos observados, esto es, sobre la distribución de probabilidad con la que modelizar la respuesta. Sin embargo, la mayoría de las veces, es posible especificar el rango de los posibles valores de la variable respuesta, así como su tipo (si es continua o discreta). Además, la experiencia en estudios previos suele permitir especificar, de una forma cualitativa al menos, algo más de información sobre ciertas características de los datos como por ejemplo:

- cómo viene afectada la respuesta media (o mediana) por estímulos externos o tratamientos;
- cómo cambia la variabilidad de la respuesta con la respuesta promedio;
- si las observaciones son estadísticamente independientes o no;

- si la distribución de la respuesta bajo tratamientos fijos es asimétrica a la izquierda, a la derecha, o es simétrica.

En las situaciones en que hay información insuficiente para construir y trabajar con la verosimilitud de los datos, pero hay alguna información disponible sobre las características de los datos, es posible trabajar y resolver el análisis a través de la **estimación cuasi-verosímil**.

Ejemplo 6.1 (Continuación del Ejemplo 3.7. Raíces de Gilligan) Recordemos los datos de la Tabla 3.6, en los que se disponía del número de raíces infectadas de un total que recibía cierto inóculo a una determinada concentración. Se trataba de predecir la probabilidad de infección en función del tipo de suelo (factor 'suelo') y del logaritmo en base 2 de la densidad de inóculo, al que habíamos llamado 'ldosis'.

Se había ajustado en primer lugar un modelo maximal polinómico de orden 4 que daba indicios de sobredispersión. De hecho, su deviance era de $D = 40,06$ con 20 g.l., de donde se estimaba un valor de ϕ aproximadamente de 2,

$$\hat{\phi} = 40,05/20 \approx 2.$$

Así pues, si bien en un principio la verosimilitud binomial parecía la lógica a utilizar para modelizar estos datos, el hecho de que el parámetro de escala ϕ sea distinto de 1, evidencia de alguna forma lo inadecuado de utilizarla para la estimación. Sin embargo, se podría pensar en que si, efectivamente el número esperado de raíces infectadas es $E(Y) = n\pi$, con π la probabilidad de infección, y n el número de raíces tratadas de la misma forma, entonces $Var(Y) = \phi n\pi(1 - \pi)$, con $\phi \approx 2$.

En un modelo de cuasi-verosimilitud las únicas hipótesis necesarias sobre el vector de respuestas son:

1. $\mu(\beta; X)$ es el vector de medias de la respuesta, $E(Y) = \mu(\beta; X)$, relacionado linealmente con un conjunto de variables explicativas X , a través de una función link $g(\mu) = \eta = X\beta$;
2. ϕ es un parámetro de escala que puede ser desconocido;
3. $V(\mu)$ es la función de varianza que relaciona la varianza covarianza de las respuestas con las respuestas medias, $Var(Y) = \phi V(\mu)$;

Con esta información es posible construir una cuasi-verosimilitud de la que obtener estimaciones de los parámetros β en el predictor lineal, y en consecuencia de las respuestas medias esperadas.

6.2. Cuasi-verosimilitud y estimación

Si se asume que las observaciones $Y_1, Y_2, \dots, Y_n$ son independientes y $Var(Y)$ representa la matriz de varianzas-covarianzas, se tiene que la función de varianza $V(\mu)$ será una matriz diagonal,

$$V(\mu) = \text{diag}\{V_1(\mu), \dots, V_n(\mu)\}.$$

Por otro lado, es difícil imaginar un mecanismo físico que produzca observaciones independientes cuyas funciones de varianza $V_i(\mu)$ dependan de otras respuestas diferentes a la i -ésima variable Y_i . Esto es, en general tendremos que $V_i(\mu) = V_i(\mu_i)$ y así la hipótesis de independencia asumida dará lugar a una estructura de varianzas de la forma:

$$V(\mu) = \text{diag}\{V_1(\mu_1), \dots, V_n(\mu_n)\}.$$

Además, en la mayoría de las aplicaciones, las funciones $V_1(), \dots, V_n()$ se suelen asumir idénticas, aunque no es una condición necesaria.

Una vez clara la estructura de varianzas, pasamos a construir los elementos básicos para la estimación de la respuesta media con un modelo de cuasi-verosimilitud.

Recordemos la forma que tomaba la función score, fuera cual fuera la verosimilitud asumida para la respuesta dentro de la familia exponencial, con $a_i(\phi) = \phi$:

$$\begin{aligned} U_j = \frac{\partial l(\theta; \mathbf{y})}{\partial \beta_j} &= \sum_{i=1}^n \frac{(y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)} \\ &- \sum_{i=1}^n \frac{(y_i - \mu_i)}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j}, \end{aligned}$$

definida únicamente por la media μ_i , la varianza $Var(Y_i)$, el parámetro de dispersión ϕ y la relación entre la media y las covariables $g(\mu_i) = X\beta$.

Análogamente al caso verosímil, en el caso en que no se dispone de una especificación de la verosimilitud, se construye el vector de **cuasi-scores** U_q de forma que su componente j -ésima viene dada por:

$$U_{qj} = \sum_{i=1}^n \frac{(y_i - \mu_i)}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j}, \quad (6.1)$$

Expresado U_q en forma matricial tenemos:

$$U_q = \frac{1}{\phi} D' W^{-1} (\mathbf{y} - \mu), \quad (6.2)$$

con $W = \text{diag}\{V_1(\mu_1), \dots, V_n(\mu_n)\}$ y $D_{ij} = \frac{\partial \mu_i}{\partial \beta_j}$.

U_q se comporta como el vector de scores ordinario, en el sentido de que:

- $E(U_q) = 0$
- $E(U_q U_q') = \text{Var}(U_q) = \frac{1}{\phi} D' W^{-1} D$
- $E\left(\frac{\partial U_q}{\partial \beta}\right) = -\frac{1}{\phi} D' W^{-1} D = -E(U_q U_q')$.

El **estimador cuasi-verosímil** de β , $\hat{\beta}$, se obtiene como la solución del sistema

$$D' W^{-1} (\mathbf{y} - \mu(\beta)) = 0, \quad (6.3)$$

es decir,

$$\sum_{i=1}^n \frac{\partial \mu_i}{\partial \beta_j} \frac{y_i - \mu_i}{V_i(\mu_i)} = 0, \quad \forall j = 1, \dots, p.$$

Si $\hat{\beta}$ es solución de (6.3), entonces asintóticamente $\hat{\beta}$ es un estimador consistente de β .

La resolución de (6.3) se puede obtener por el Método Scoring de Fisher,

$$\hat{\beta}^{(r)} = \hat{\beta}^{(r-1)} + (D' W^{-1} D)^{-1} D' W^{-1} (\mathbf{y} - \mu)|_{\hat{\beta}^{(r-1)}}. \quad (6.4)$$

Así, si la distribución de Y pertenece efectivamente a la familia exponencial, con $E(Y) = \mu$ y $\text{Var}(Y) \propto V(\mu)$, entonces la solución de (6.3) para β es la misma que

la que se obtiene resolviendo las ecuaciones de máximo verosimilitud usuales para la distribución en cuestión. Sin embargo, los errores estándar de las estimaciones han de ser ajustados por la constante de proporcionalidad ϕ ,

$$\text{Var}(\hat{\beta}) = \phi(D'W^{-1}D)^{-1}. \quad (6.5)$$

Si ϕ es desconocido, será reemplazado por una estimación adecuada, como por ejemplo la deviance media o la estimación corregida de los grados de libertad,

$$\hat{\phi} = \frac{D}{n-p} \quad (6.6)$$

$$\tilde{\phi} = \frac{1}{n-p} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}. \quad (6.7)$$

La función de **log cuasi-verosimilitud** $Q(\beta, \mathbf{y})$, se define como:

$$Q(\beta; \mathbf{y}) = \sum_{i=1}^n \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V_i(t)} dt. \quad (6.8)$$

Esta función se comporta como la log-verosimilitud usual, en el sentido de que su derivada respecto de β_j da la función cuasi-score,

$$\frac{\partial Q(\beta; \mathbf{y})}{\partial \beta} = U_q.$$

Por analogía, se puede definir la **cuasi-deviance** con

$$D_q(\mathbf{y}; \mu) = -2\phi Q(\beta; \mathbf{y}). \quad (6.9)$$

Ejemplo 6.2 Supongamos $Y_1, \dots, Y_n$ v.a. independientes, con $E(Y_i) = \mu_i$ y $\text{Var}(Y_i) = \mu_i$, de donde $\phi = 1$ y $V_i(\mu_i) = \mu_i$. La función de cuasi-deviance re-

sulta:

$$\begin{aligned}
 D_q(\mathbf{y}; \mu) &= -2 \sum_{i=1}^n \int_{y_i}^{\mu_i} \frac{(y_i - t)}{t} dt = -2 \sum_{i=1}^n [y_i \log(t) - t]_{y_i}^{\mu_i} \\
 &= -2 \sum_{i=1}^n (y_i \log(\mu_i) - \mu_i - y_i \log(y_i) + y_i) \\
 &= 2 \sum_{i=1}^n y_i \log\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i),
 \end{aligned}$$

que es la deviance para variables Poisson. De hecho, $\phi = 1$ implica que la relación media-varianza define unívocamente la distribución para la variable respuesta.

Una de las aplicaciones de la función de cuasi-deviance será la del cálculo de estadísticos de cociente de verosimilitudes en la comparación de modelos anidados.

Sea M_1 un modelo con p_1 parámetros, anidado en M_2 , con p_2 parámetros. El **estadístico del cociente de cuasi-verosimilitud** para resolver el contraste de $H_0 : M_1$ versus $H_1 : M_2$ cuando ϕ es conocido, viene dado por

$$\Lambda_q = -2[D_q(\hat{\beta}_1; \mathbf{y}) - D_q(\hat{\beta}_2; \mathbf{y})]. \quad (6.10)$$

Cuando M_1 es un modelo adecuado, Λ_q tiene aproximadamente una distribución $\chi_{p_2-p_1}^2$.

Ejemplo 6.3 Sean $Y_1, \dots, Y_n$ v.a. incorreladas, con $E(Y_i) = \mu_i$ y $Var(Y_i) = \phi \mu_i^2$. Supongamos el link logaritmo para ligar la media con el predictor lineal, de la forma:

$$\log(\mu_i) = \beta_1 + \beta_2(x_i - \bar{\mathbf{x}}), \quad i = 1, \dots, n.$$

Entonces, las ecuaciones de estimación cuasi-verosímil resultan del sistema:

$$\sum_{i=1}^n \frac{1}{\phi} \frac{\partial \mu_i}{\partial \beta_j} \frac{y_i - \mu_i}{\mu_i^2} = 0,$$

que da lugar a:

$$\begin{aligned}\frac{\partial U_q}{\partial \beta_1} = 0 &\rightsquigarrow \sum_{i=1}^n \frac{y_i - \mu_i}{\mu_i} = 0, \\ \frac{\partial U_q}{\partial \beta_2} = 0 &\rightsquigarrow \sum_{i=1}^n \frac{(y_i - \mu_i)(x_i - \bar{\mathbf{x}})}{\mu_i} = 0,\end{aligned}$$

con $\mu_i = \exp\{\beta_1 + \beta_2(x_i - \bar{\mathbf{x}})\}$.

La log cuasi-verosimilitud $Q(\beta; \mathbf{y})$ es:

$$Q(\beta; \mathbf{y}) = -2 \sum_{i=1}^n \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi t^2} dt = \frac{2}{\phi} \sum_{i=1}^n \left[\frac{y_i}{\mu_i} + \log \left(\frac{\mu_i}{y_i} \right) + 1 \right].$$

Aparte de utilizar la cuasi-verosimilitud para decidir si eliminar o no algunos parámetros de un modelo, se puede acudir también a los estadísticos de Wald y a los score. Para calcular estos tests será necesario utilizar la matriz de información de Fisher.

La matriz de información de Fisher para el modelo de cuasi-verosimilitud viene dada por:

$$\mathcal{I}_q = \frac{1}{\phi} D' W^{-1} D, \quad (6.11)$$

evaluada en estimación cuasi-verosímil de β , $\hat{\beta}$, obtenida de (6.4).

Si además ϕ es desconocido, se utiliza alguna de sus estimaciones según (6.6) o (6.7).

Resultado 6.2.1 Cuando $\text{Var}(Y_i) = \phi$, tenemos que $V(\mu_i) = 1$, y la estimación cuasi-verosímil se reduce a la estimación de mínimos cuadrados ordinarios:

$$D' W^{-1} (\mathbf{y} - \mu) = 0 \rightsquigarrow \sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta_j} \right) (y_i - \mu_i) = 0,$$

con $D_{ij} = \partial \mu_j / \partial \beta_i$ y $W = I$, la matriz identidad.

Resultado 6.2.2 Cuando $Var(Y_i) = a_i\phi$, con a_i constantes conocidas, la estimación cuasi-verosímil es una estimación de mínimos cuadrados ordinarios:

$$\sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta_j} \right) \frac{y_i - \mu_i}{a_i} = 0,$$

6.3. Sobredispersión: Modelización y Análisis

La sobredispersión es un problema muy común en datos que consisten en conteos o frecuencias de sucesos. Asumida una determinada distribución para los datos, se presenta sobredispersión cuando la variabilidad de éstos es mayor que la esperada bajo el modelo asumido. Cómo modelizar sobredispersión es entonces un problema añadido. Los modelos de cuasi-verosimilitud permiten en esas ocasiones, prescindir una verosimilitud concreta para las respuestas y modelizar en función del predictor lineal y de la forma asumida para representar la varianza en función de la media.

Si bien hay muchas formas posibles de representar la sobredispersión de unos datos, modelos habituales son:

- en v.a. Poisson, $Var(Y_i) = \phi\mu_i$
- en v.a. Binomiales, $Var(Y_i) = \phi n_i \pi_i (1 - \pi_i)$
- en v.a. Exponenciales, $Var(Y_i) = \phi\mu_i^2$.

Para estos modelos en los que la función de varianza difiere únicamente por un factor de escala de la función de varianza en el modelo de verosimilitud correspondiente (Poisson, binomial o exponencial), las estimaciones de los parámetros que se obtienen son iguales. Es decir, las estimaciones para el modelo de cuasi-verosimilitud son idénticas a las que se derivan utilizando la verosimilitud completa. Sin embargo, los errores estándar de las primeras se calculan multiplicando los errores estándar de los EMV por $\phi^{1/2}$ (generalmente por una estimación suya si ϕ es desconocido). Es decir,

$$s.e._q(\hat{\beta}) = \hat{\phi}^{1/2} s.e.(\hat{\beta})$$

Cuando la sobredispersión es ligera/media, Cox (1983) demostró que la estimación por cuasi verosimilitud era bastante eficiente para la estimación de un parámetro de momentos en una distribución de la familia exponencial.

Los modelos de cuasi-verosimilitud son adecuados para modelizar situaciones de sobredispersión. Cuando unos datos son modelizables según algún modelo lineal generalizado (se ha identificado la tendencia lineal) pero manifiestan claramente sobredispersión, es decir, desviación de alguna verosimilitud concreta respecto a la variabilidad, bastará con determinar una forma adecuada para asimilar el exceso de variabilidad (función de varianza) y realizar el ajuste por cuasi-verosimilitud. Veamos a continuación algunos ejemplos para datos binomiales y Poisson.

Ejemplo 6.4 (Continuación del Ejemplo 6.1. Sobredispersión en datos binomiales)

Si se obvia la sobredispersión detectada para los datos de las raíces infectadas de Gilligan, y se ajusta un modelo binomial con link cloglog con interacción entre las dos variables explicativas 'ldosis' y 'S', obtenemos un ajuste razonablemente bueno, comparado con el modelo maximal polinómico.

```
glm(yres ~ suelo * ldosis, family = binomial(link = "cloglog"))
```

El p-valor asociado al test de comparación entre este modelo y el maximal resultaba de $p = 0,82$. Las estimaciones de los parámetros según este modelo binomial eran:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.99717	0.15279	-6.527	6.73e-11
sueloS	-0.35735	0.24361	-1.467	0.14241
ldosis	0.61471	0.06583	9.338	< 2e - 16
sueloS.ldosis	0.32212	0.12308	2.617	0.00887

En la salida del ajuste se hace notar la hipótesis distribucional de datos binomiales:

```
(Dispersion Parameter for Binomial family taken to be 1 )
Null Deviance: 340.9804 on 29 degrees of freedom
Residual Deviance: 42.95783 on 26 degrees of freedom
AIC: 121.49.
```

Si tenemos en cuenta los indicios de sobredispersión que mencionamos ($\phi \neq 1$), podemos ajustar un modelo cuasi-verosímil asumiendo que la varianza responde a la forma:

$$\text{Var}(Y_i) = \phi n_i \pi_i (1 - \pi_i),$$

con ϕ desconocido (mayor que 1). Los resultados son similares para las estimaciones de los parámetros y deviances, y las diferencias estriban sólo en los errores estándar de las estimaciones de los parámetros:

```
glm(yres ~ suelo * ldosis, family = quasi(link = "cloglog",
                                           variance = "mu(1 - mu)"))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.99717	0.18830	-5.296	1.54e-05
sueloS	-0.35735	0.30023	-1.190	0.2447
ldosis	0.61471	0.08113	7.577	4.82e-08
sueloS.ldosis	0.32212	0.15169	2.124	0.0434

(Dispersion parameter for quasibinomial family taken to be 1.518884)

Null deviance: 340.980 on 29 degrees of freedom

Residual deviance: 42.958 on 26 degrees of freedom.

Como se puede comprobar, las estimaciones de los errores estándar en el modelo cuasi son igual a las del modelo binomial multiplicadas por la estimación de ϕ obtenida, $\phi = 1,518884$.

Otra forma de ajustar el modelo propuesto es mediante un análisis de verosimilitud completo asumiendo otra estructura para la sobredispersión, dado por la distribución beta-binomial:

$$Y_i|P_i \sim Bi(n_i, P_i)$$

$$P_i \sim Beta(), \text{ con } E(P_i) = \pi_i, \text{ Var}(P_i) = \phi\pi_i(1 - \pi_i).$$

Resultado 6.3.1 (Distribución Binomial Negativa para modelizar sobre-dispersión en datos Poisson)

Hablamos de datos Poisson sobredispersos cuando datos basados en conteos presentan una varianza mayor que la media. La utilización de la distribución binomial negativa en la modelización de datos Poisson con sobredispersión tiene una larga historia (ver Lawless, 1987). Esta distribución puede surgir de un modelo jerárquico con dos niveles. Supongamos en primer lugar que

$$Y|\lambda \sim Po(\mu\lambda), \quad (6.12)$$

de donde $E(Y|\lambda) = Var(Y|\lambda) = \mu\lambda$. Si además asumimos que λ es a su vez una variable aleatoria con distribución Gamma, media 1 y varianza $1/\theta$,

$$\lambda \sim Ga(\theta, \theta), \quad (6.13)$$

tendremos que la distribución marginal de Y es una binomial negativa, cuya función de densidad, media y varianza vienen dadas por:

$$E(Y) = \mu, \quad Var(Y) = \mu + \mu^2/\theta,$$

$$f(y; \theta, \mu) = \frac{\Gamma(\theta + y)}{\Gamma(\theta)y!} \frac{\mu^y \theta^\theta}{(\mu + \theta)^{\theta+y}}.$$

Si θ es conocida, esta distribución tiene la forma general que caracteriza a las distribuciones de la familia exponencial.

Ejemplo 6.5 (Sobredispersión en datos Poisson. Taras en rollos de fábrica) *Los datos de la Tabla 6.1 (Bissell, 1972, también en Garthwaite et al, 1995, pag. 271) corresponden al número de taras encontradas en rollos de fábrica de varias longitudes. Parece que la distribución de las taras podría ser bien aproximada por una distribución de Poisson. Hinde (1982) utilizó el logaritmo de la longitud ('long.log') como una variable explicativa, y es lo que haremos nosotros aquí. Los datos se presentan a continuación. La longitud (long) está expresada en metros.*

Planteamos a continuación diversos ajustes:

taras	6	4	17	9	14	8	5	7	7	7	
long	551	651	832	375	715	868	271	630	4911	372	
taras	6	8	28	4	10	4	8	9	23	9	6
long	645	441	895	458	642	492	543	842	905	542	522
taras	1	9	4	9	14	17	10	7	3	9	2
long	122	657	170	738	371	735	749	495	716	952	417

Tabla 6.1: Taras en rollos de fábrica

1. Un primer ajuste con el modelo Poisson,

```
glm(taras ~ long.log, family = poisson)
```

da como estimaciones:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.1730	1.1352	-3.676	0.000237
log(long)	0.9969	0.1759	5.668	1.45e-008

con una deviance de $D = 64,537$ con 30 g.l., e indicios de sobredispersión, ya patentes en la Figura 6.1, en la que se aprecia que la variabilidad aumenta a medida que lo hace la respuesta media. De hecho,

$$\hat{\phi} = 64,54/30 \approx 2,15.$$

2. Se plantea a continuación un modelo cuasi similar al anterior salvo por el parámetro de dispersión $\phi \neq 1$, con el que se pretende corregir la sobredispersión.

```
glm(taras ~ long.log, family = quasi(link = "log", variance = "mu"))
```

que da como estimaciones:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.1730	1.7093	-2.441	0.020749
log(long)	0.9969	0.2649	3.764	0.000727

con idéntica deviance y grados de libertad, diferente del primer ajuste en los errores estándar de las estimaciones.

3. Un último ajuste está basado en asumir una distribución binomial negativa para la variable que contiene el número de taras. Utilizando una función de $\mathcal{R}$ que ajusta este modelo, denominada 'neg.bin' escrita al final del Tema, y utilizando una estimación del parámetro θ empleando el método de los momentos,

$$\text{Var}(Y) = \mu + \mu^2/\theta \rightsquigarrow \theta = \frac{\mu^2}{\text{Var}(Y) - \mu} \cong 3,1613$$

obtenemos:

<i>glm(taras ~ log(long), family = neg.bin(3,1613))</i>				
	<i>Estimate</i>	<i>Std. Error</i>	<i>t value</i>	<i>Pr(> t)</i>
<i>(Intercept)</i>	-3.6946	1.4083	-2.623	0.013552
<i>log(long)</i>	0.9220	0.2213	4.166	0.000242

con una deviance $D = 16,663$ para una chi-cuadrado con 30 g.l.

Este último modelo da estimaciones algo diferentes a los anteriores. En la Figura 6.1 están representados los datos junto con los tres ajustes realizados. El modelo Poisson y el cuasi Poisson dan estimaciones similares; el modelo binomial negativo difiere levemente de los anteriores en los valores más extremos (por la izquierda y por la derecha).

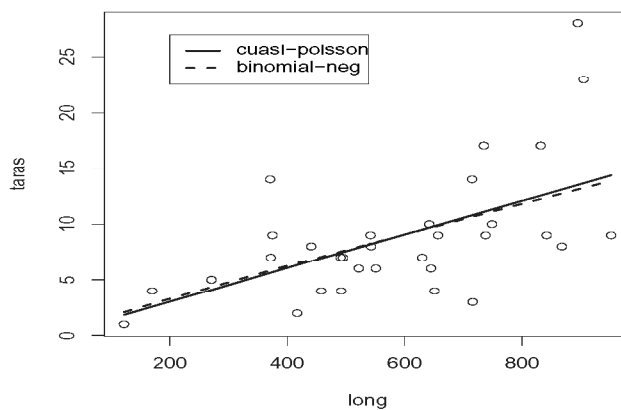


Figura 6.1: Ajuste de los datos sobre taras en rollos de fábrica de distintas longitudes.

Apéndice: Función 'bin.neg' en $\mathcal{R}$ (Se encuentra disponible en las librerías de $\mathcal{R}$)

```

neg.bin <- function(theta = stop("theta must be given")){
  .Theta <- -theta
  stats <- -make.link("log")
  variance <- -function(mu)mu + mu2/.Theta
  validmu <- -function(mu)all(mu > 0)
  dev.resids <- -function(y, mu, wt)2 * wt * (y * log(pmax(1,
y)/mu) - (y + .Theta) * log((y + .Theta)/(mu + .Theta)))
  aic <- -function(y, n, mu, wt, dev){
    term <- -(y + .Theta) * log((y + .Theta)/(mu + .Theta)) -
y * log(mu) + lgamma(y + 1) - .Theta * log(.Theta) +
lgamma(.Theta) - lgamma(.Theta + y)
    2 * sum(term * wt)
  }
  initialize <- -expression({
    if(any(y < 0))stop(paste("Negative values not allowed for",
"the Poisson family"))
    n <- -rep(1, nobs)
    mustart <- -y + (y == 0)/6
  })
  structure(list(family = "Negative Binomial", link = "log",
linkfun = stats$linkfun, linkinv = stats$linkinv,
variance = variance, dev.resids = dev.resids,
aic = aic, mu.eta = stats$mu.eta, initialize = initialize,
validmu = validmu, valideta = stats$valideta),
class = "family")
}

```

6.4. Ejercicios

1. En la Tabla 5.1 están los datos completos de un estudio sobre 44 médicos que trabajaban en el servicio de urgencias de un hospital grande. Estos datos ya fueron analizados con una regresión de Poisson en el Tema anterior (Ejemplo 5.1). El objetivo del estudio era determinar qué variables estaban relacionadas con el número de quejas que recibían los médicos a lo largo de un año.

Considera la variable respuesta Y_{ij} = número de quejas que recibe el médico j -ésimo del grupo i ($i = 1$ los que han sido residentes, $i = 2$ los que no han sido residentes en servicio de urgencias), con $j = 1, \dots, J_i$. Asume que Y_{ij} tiene una distribución en la familia exponencial, con media $E(Y_{ij}) = \mu_{ij} = \lambda_{ij} \cdot Z_{ij}$, con Z_{ij} el número de consultas que realiza el médico, y que su varianza es $Var(Y_{ij}) = \phi \mu_{ij}$, con ϕ desconocido. Asume también que la relación entre la tasa de quejas λ_{ij} y las variables explicativas *residente* y *horas* viene dada por:

$$\log(\lambda_{ij}) = \theta + \alpha_i + \beta_i T_{ij},$$

siendo $\alpha_2 - \alpha_1$ la diferencia entre el efecto de los residentes y el de los no residentes respecto al número de quejas recibidas, y T_{ij} el número de horas en el servicio de urgencias del médico en cuestión.

- a) Identifica la función link en el modelo propuesto. ¿Qué papel desempeña la variable *consultas* Z_{ij} ?
 - b) Ajusta un modelo de cuasi-verosimilitud (*fit*) que respete todas las hipótesis propuestas en este apartado, y por supuesto sea significativo. Fórmalo con las estimaciones obtenidas. Justifica la bondad del ajuste obtenido.
 - c) Obtén asimismo, una estimación de ϕ . ¿Cuál es el significado de este parámetro?
 - d) En función del ajuste que has obtenido, da conclusiones prácticas que pueda entender cualquier directivo de hospital preocupado por la cuestión del rendimiento de los médicos del servicio de urgencias. ¿Qué consejos le darías para la contratación de personal?
 - e) Para completar la versión teórica del apartado anterior, construye los cuasi-scores, la cuasi-verosimilitud y la cuasi-deviance. Obtén las ecuaciones a resolver para obtener los estimadores máximo-cuasi-verosímiles.
2. Los datos en la Tabla 6.2, (Wedderburn, 1974, también en McCullagh y Nelder, 1989, pag.) conciernen a la incidencia de *Rhynchosporium secalis* (manchas en

las hojas) en 10 variedades de cebada cultivadas en 9 localizaciones distintas en 1965. La respuesta, que es la proporción del área de la hoja afectada por las manchas, una proporción continua en el intervalo $[0, 1]$.

Para un primer análisis, considera los datos como observaciones pseudo-binomiales, asumiendo inicialmente que la varianza es $\phi\mu(1 - \mu)$ y un link logit. Obtén el ajuste y concluye sobre las diferencias con la variedad 1. Representa el ajuste para verificar si esa modelización de la varianza es adecuada (residuos versus predicción lineal).

Prueba en segundo lugar con una función de varianza alternativa, $\mu^2(1 - \mu)^2$ para intentar modelizar el efecto que observaste en el gráfico anterior. Deriva la función de cuasi-verosimilitud para este modelo. Representa el ajuste y verifica si este modelo describe mejor los datos observados.

loc	Variedad										media
	1	2	3	4	5	6	7	8	9	10	
1	0.05	0.0	0.0	0.1	0.25	0.05	0.5	1.3	1.5	1.5	0.52
2	0.0	0.05	0.05	0.3	0.75	0.3	3.0	7.5	1.0	12.7	2.56
3	1.25	1.25	2.5	16.6	2.5	2.5	0.0	20.0	37.5	26.25	11.03
4	2.5	0.5	0.01	3.0	2.5	0.01	25.0	55.0	5.0	40.0	13.35
5	5.5	1.0	6.0	1.1	2.5	8.0	16.5	29.5	20.0	43.5	13.36
6	1.0	5.0	5.0	5.0	5.0	5.0	10.0	5.0	50.0	75.0	16.6
7	5.0	0.1	5.0	5.0	50.0	10.0	50.0	25.0	50.0	75.0	27.51
8	5.0	10.0	5.0	5.0	25.0	75.0	50.0	75.0	75.0	75.0	40.0
9	17.5	25.0	42.5	50.0	37.5	95.0	62.5	95.0	95.0	95.0	61.5
media	4.2	4.77	7.34	9.57	14.0	21.76	24.17	34.81	37.22	49.33	20.72

Tabla 6.2: Incidencia de RS sobre el crecimiento de cebada

Índice de tablas

1.1. Distribuciones usuales. Relación entre la media y la varianza.	3
1.2. Datos de un servicio de urgencias hospitalario.	4
1.3. Algunos modelos y links usuales en ajustes de GLM's.	7
1.4. Datos de mortalidad de cucarachas.	8
1.5. Melanoma maligno: frecuencias por tipo de tumor y localización. . .	8
1.6. Satisfacción en el trabajo e ingresos.	9
1.7. Datos sobre el crecimiento de plantas.	9
1.8. Tiempos de supervivencia (en semanas) para pacientes de leucemia y conteo inicial de células blancas en la sangre (en escala $\log_{10}$). . . .	10
1.9. Formación de capital doméstico bruto en Gran Bretaña durante 1948– 1966.	10
1.10. Número de taras versus la longitud de los rollos.	11
2.1. Datos Poisson.	32

2.2. Datos sobre el crecimiento de plantas.	42
2.3. Peso al nacer y edad gestacional para bebés varones y hembras . . .	58
2.4. Consumo de azúcar como refinado o para manufacturas de alimentos. .	61
2.5. Producción de hierbas y legumbres versus la cantidad de fosfatos in- cluidos en el abono	68
2.6. Obediencia a un régimen alto en carbohidratos de 20 diabéticos varones dependientes de la insulina	69
2.7. Efecto del peso en el nivel de colesterol	70
2.8. Efectividad de la Cholestyramine para reducir el nivel de colesterol. Obediencia de los pacientes a las prescripción de medicamento . . .	71
3.1. Datos de Kifosis	75
3.2. Datos de mortalidad de cucarachas	76
3.3. Datos sobre toxicidad con polillas del tabaco.	76
3.4. Proporción de cucarachas muertas observadas y estimadas por los modelos logit, probit y cloglog	94
3.5. Residuos deviance en los modelos logit, probit y cloglog. Cucarachas. .	94
3.6. Datos de raíces infectadas de Gilligan.	100
3.7. Pruebas de insecticidas	105
3.8. Efecto de la radiación en la supervivencia de las células	106
3.9. Uso de drogas psicotrópicas en Londres	107
3.10. Incidencia del cáncer de esófago versus la edad y el consumo de alcohol	107
3.11. Supervivencia en pacientes de cáncer.	109
3.12. Cáncer de pulmón en los astilleros de Georgia	109

3.13. Cáncer cervical y sensibilidad del test citológico.	110
3.14. Hábitos tabaquísticos y pancreatitis.	110
3.15. Problemas de sueño en los veteranos de Vietnam	110
3.16. Muerte coronaria y alcohol.	111
3.17. Muerte coronaria y uso de contraceptivos.	111
3.18. Consumo de café y cáncer de ovarios.	112
3.19. Estudios y tabaco en lugares públicos.	112
3.20. Cinturón de seguridad y daños en accidentes de tráfico.	113
3.21. Edad de la madre y niños prematuros.	113
3.22. Uso de condón y embarazos.	113
3.23. Uso de servicios médicos y raza.	114
3.24. Sensibilidad del diagnóstico de tuberculosis con rayos X.	114
3.25. Cáncer de mama, edad, reproducción y menopausia.	114
3.26. Cáncer en el endometrio, edad y estrógenos.	115
4.1. Melanoma maligno: frecuencias para tipo de tumor y localización. . .	118
4.2. Ensayo clínico sobre la vacuna de la gripe.	119
4.3. Relación entre las úlceras gástrica - duodenal y el consumo de aspirina.	119
4.4. Datos sobre el Melanoma. Resumen de la Tabla 4.1.	130
4.5. Aspiraciones universitarias, estatus social e incentivos familiares . .	136
4.6. Usos contraceptivos y edades en mujeres casadas	137
4.7. Ideología y afiliación políticas	149

4.8. Satisfacción en el trabajo e ingresos	149
4.9. Cambio de residencia en la Commonwealth	155
4.10. Estimaciones del cambio de residencia en la Commonwealth con el modelo mover-stayer.	157
4.11. Evolución en el voto de Suecia entre 1968 y 1970.	158
4.12. Predicción de la evolución en el voto de Suecia entre 1968 y 1970 con un modelo de casi-simetría	161
4.13. Tabla de contingencia $I \times 2$	162
4.14. Encuesta danesa sobre edad y estatus marital.	163
4.15. Incidencia de Torus Mandibularis en tres poblaciones de esquimales. Presencia (+) y ausencia (−).	166
4.16. Supervivencia en pacientes de cáncer.	167
4.17. Datos sobre contraceptivos.	168
4.18. Efecto de la edad de la madre y sus hábitos con el tabaco sobre el período gestacional y la mortalidad perinatal.	169
4.19. Número de huevos que no se comieron los escarabajos castaneum de un total de 50 huevos de cada especie en tres replicaciones distintas del experimento.	171
4.20. Trabajadores en Houston clasificados por estatus de fumador, edad y resultados de un test respiratorio	172
4.21. Satisfacción en el hogar y condiciones de la vivienda	173
4.22. Elecciones británicas en 1966. Comparación con los resultados de 1964.	173
4.23. Estudio de movilidad social intergeneracional.	174
4.24. Tests sobre los ojos para visión de lejos sin ayuda de lentes	174

4.25. Pertenencia a bandas y actitud en ellas	175
4.26. Efecto de la contaminación en los niños de dos ciudades de USA . .	176
4.27. Actitud ante la despenalización del aborto	177
4.28. Salarios de profesores de colegios	178
4.29. Hábitos de tabaco en estudiantes y en sus padres.	178
4.30. Uso de drogas psicotrópicas en Londres.	179
4.31. Relaciones sexuales en adolescentes	179
4.32. Salud mental y nivel socioeconómico	179
5.1. Quejas en el servicio de urgencias de un hospital	182
5.2. Formación del Capital Doméstico Bruto en el Reino Unido en el período comprendido entre 1948 y 1967.	188
5.3. Casos de SIDA diagnosticados en USA, 1982–1990.	189
5.4. Tiempos de supervivencia (en semanas) a leucemia aguda bajo dos tratamientos.	197
5.5. Número de incidentes con daños en buques cargueros, meses de ser- vicio y año de construcción.	206
5.6. Tiempos de supervivencia (en días) tras una operación.	207
5.7. Tiempos de supervivencia en semanas para 17 pacientes de leucemia	207
5.8. Crecimiento de una hoja de judía a lo largo de 36 días.	208
5.9. Alturas de dos girasoles medidas cada semana.	208
5.10. Pesos (g) de dos pikas afganas preñadas	208
5.11. Estudio sobre la efectividad de la radiación para prolongar la vida de pacientes con cáncer gástrico	209

5.12. Tiempos de supervivencia de 50 patos negros hembras en New Jersey	210
6.1. Taras en rollos de fábrica	222
6.2. Incidencia de RS sobre el crecimiento de cebada	226

Bibliografía

- Agresti, A. (1990) *Categorical Data Analysis*. New York: John Wiley.
- Aitkin, M. and Clayton, D. (1980). The fitting of exponential, Weibull, and extreme value distributions to complex censored survival data using GLIM. *Journal of the Royal Statistical Society* **C29**, 1256-163.
- Barndorff-Nielsen (1978). *Information and Exponential Families in Statistical Theory*. Chichester: Wiley.
- Barnhart, H.X. and Sampson, A.R. (1995). Multiple population models for multivariate random length data –with applications in clinical trials. *Biometrics*, **51**, 195-204.
- Begg, C.B. and McNeil, B. (1988). Assessment of radiologic tests: Control of bias and other design considerations. *Radiology*, **167**, 565-569.
- Berkowitz, G.S. (1981). An epidemiologic study of preterm delivery. *American Journal of Epidemiology*, **113**, 81-92.
- Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.
- Bissel, A.F. (1972) A negative binomial model with varying element sizes. *Biometrika* **59**, 435-441.
- Bliss, C.I.(1935) The calculation of the dosage mortality curve. *Annals of Applied Biology*, **22**, 134-167.

- Blot, W.J., Harrington, M., Toledo, A., Hoover, R., Heath, C.W., and Fraumeni, J.F. (1978). Lung cancer after employment in shipyards during World War II. *New England Journal of Medicine*, **299**, 620-624.
- Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research, 1: The Analysis of Case-Control Studies*. I.A.R.C., Lyon.
- Coleman, J.S. (1964). *Introduction to Mathematical Sociology*. New York: The Free Press of Glencoe.
- Collet, D. (1991) *Modelling Binary Data*. Chapman and Hall, London.
- Cox, D.R. (1983). Some Remarks on over dispersion. *Biometrika*, **70**, 269-274.
- Cox, D.R. & Snell, E.J. (1981) *Applied Statistics: Principles and Examples*. Chapman and Hall, London.
- Chambers, J.M. and Hastie, T.J. (1993) *Statistical Models in S*. Chapman and Hall.
- Daniel, W.W. (1987). *Biostatistics: A foundation for analysis in the Health Sciences*. New York: John Wiley and Sons.
- Dobson, A.J. (1990) *An Introduction to Generalized Linear Models*. London: Chapman and Hall.
- Duggan, J.M., Dobson, A.J., Johnson, H., and Fahey, P.P. (1986). Peptic ulcer and non-steroidal anti-inflammatory agents. *Gut*, **27**, 929-933.
- Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.
- Fienberg, S.E. (1978). *The analysis of cross-classified categorical data*. M.I.T. Press.
- Fingleton, B. (1984). *Models of Category Counts*. Cambridge: Cambridge University Press.
- Fitzmaurice, G.M. and Laird, N.M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141-151.
- Forthofer, R.N. and Lehnen, R.G. (1981) *Public Program Analysis, a New Categorical Data Approach*. Belmont, CA: Lifetime Learning Publications.

- Gamerman, D. (1991). Dynamic Bayesian models for survival data. *Journal of the Royal Statistical Society* **C40**, 63-79.
- Garthwaite, P.H., Jolliffe, I.T. and Jones, B. (1997) *Statistical Inference*. Prentice Hall.
- Gilligan, C.A.(1983) A test for randomness of infection by soilborne pathogens. *Phytopathology*, **73**, 300-303.
- Hastie, T. and Tibshirani, R.J. (1990) *Generalized Additive Models*. Chapman and Hall.
- Hay, J.W. and Wolack, F.A. (1994). A procedure for estimating the unconditional cumulative incidence curve and its variability for the human immunodeficiency virus. *Journal of the Royal Statistical Society* **C43**, 599-624.
- Hedlund, R.D. (1978). Cross-over voting in a 1976 presidential primary. *Public Opinion Quart.*, **41**, 498-514.
- Hewlett, P.S. and Plackett, R.L.(1950) Statistical aspects of the independent joint action of poisons, particularly insecticides. II. Examination of data for agreement with hypothesis. *Annals of Applied Biology*, **37**, 527-552.
- Hinde (1982), en Garthwaite et al
- Jackson, R., Scragg, R. and Beaglehole, R. (1992). Does recent alcohol consumption reduce the risk of acute myocardial infarction and coronary death in regular drinkers? *American Journal of Epidemiology*, **136**, 819-824.
- Kelsey, J.L., Livolsi, V.A., Holford, T.R., Fischer, D.B., Mostow, E.D., Schartz, P.E., O'Connor, T. and White, C. (1982). A case-control study of cancer of the endometrium. *American Journal of Epidemiology*, **116**, 333-342.
- Krzanowski, W.J. (1998) *An Introduction to Statistical Modeling*. Arnold.
- Le, C.T. (1998). *Applied Categorical Data Analysis*. Wiley.
- Li, D.K., Daling, J.R., Stergachis, A.S., Chu, J. and Weiss, N.S. (1990). Prior condom use and the risk of tubal pregnancy. *American Journal of Public Health*, **80**, 964-966.
- Lindsey, J.K. (1998) *Applying Generalized Linear Models*. Springer Verlag.
- Lloyd, C.J. (1999). *Statistical Analysis of Categorical Data*. Wiley.

- McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.
- Madsen, M. (1976). Statistical Analysis of Multiple contingency tables: Two examples. *Scandinavian Journal of Statistics*, **3**, 97-106.
- May, D. (1974) Error rates in cervical cytological screening tests. *British Journal of Cancer*, **29**, 106-113.
- Morgan, S.P. and Teachman, J.D. (1988). Logistic regression: description, examples, and comparisons. *Journal of Marriage Family*, **50**, 929-936.
- Muller, T.P. and Mayhall, J.T. (1971). Analysis of contingency table data of torus mandibularis using a loglinear model. *American Journal of Physical Anthropology*, **34**, 149-154.
- Negri, E., Vecchia, C.L., Bruzzi, P., Dardanoni, G., Decali, A., Palli, D., Parazzini, F. and Del Turco, M.R. (1988). Risk factors for breast cancer: pooled results from three Italian case-control studies. *American Journal of Epidemiology*, **128**, 1207-1215.
- Oliver, F.R. (1970) Estimating the exponential growth function by direct least squares. *Journal of the Royal Statistical Society* **C19**, 92-100.
- Pierce, D.A. and Schafer, D.W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, **81**, 977-983.
- Pollock, K.H., Winterstein, S.R., and Conroy, M.J. (1989). Estimation and analysis of survival distributions for radio-tagged animals. *Biometrics*, **45**, 99-109.
- Powsner, L. (1935). The effect of temperature on the durations of the developmental stages of *Drosophila Melanogaster*. *Physiological Zoölogy*, **8**, 474-520.
- Pregibon, D. (1981) Logistic regression diagnostics. *Annals of Statistics* **29**, 705-724.
- Roberts, G., Martyn, A.L., Dobson, A.J., and McCarthy, W.H. (1981). Tumour thickness and histological type in malignant melanoma in New South Wales, Australia, 1970-76. *Pathology*, **13**, 763-770.
- Rodríguez, G.(1998). *Notes on Generalized Linear Models*, disponible via internet. Princeton University (USA).

- Sandland, R.L. and McGilchrist, C.A. (1979). Stochastic growth curve analysis. *Biometrics*, **35**, 255-271.
- Scallan, C.V. (1985). Fitting autoregressive processes in GLIM. *GLIM Newsletter*, **9**, 17-22.
- Shapiro, S.L., Rosenberg, D., Slone, D. and Kaufman, D.W. (1979). Oral contraceptive use in relation to myocardial infarction. *Lancet*, **1**, 743-746.
- Srole, L., Langner, T.S., Michael, S.T., Kirkpatrick, P., Opler, M.K. and Rennie, T.A.C. (1978). *Mental Health in the Metropolis: The Midtown Manhattan Study*, rev.ed. NYU Press, New York.
- Strogatz, D. (1990). Use of medical care for chest pain differences between blacks and whites. *American Journal of Public Health*, **80**, 290-293.
- Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables. *Biometrika*, **40**, 105-110.
- True, W.R., Golberg, J. and Eisen, S.A. (1988). Stress symptomology among Vietnam veterans. *American Journal of Epidemiology*, **128**, 85-92.
- Upton, G.J.G. (1978) *The analysis of cross-tabulated data*. New York Wiley.
- Venables, W.N. and Ripley, B.D. (1997) *Modern Applied Statistics with Splus*, 2nd ed. Springer Verlag.
- Wedderburn, R.W.M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439-447.
- Weisberg, S. (1985). *Applied Linear Regression*, 2nd ed. John Wiley and Sons.
- Wermuth, N. (1976). Exploratory analyses of multidimensional contingency tables. *Proceedings of the 9th International Biometrics Conference*, VI, 279-295.
- Whittemore, A.S., Wu, M.L., Paffenbarger, R.S., Sarles, D.L., Kampert, J.B., Brosner, S., Jung, D.L., Ballon, S. and Hendrickson, M. (1988). Personal and environmental characteristics related to epithelial ovarian cancer. *American Journal of Epidemiology*, **128**, 1228-1240.
- Yen, S., Hsieh, C. and MacMahon, B. (1982). Consumption of alcohol and tobacco and other risk factors for pancreatitis. *American Journal of Epidemiology*, **116**, 407-414.