

ANEXO D

Modelos de regresión de Poisson y sobredispersión

La forma general del modelo de Poisson es

$$\ln T_{jk} = \alpha_j + \mathbf{x}_{jk}\boldsymbol{\beta}$$

siendo j y k las diferentes categorías de exposición; T_{jk} las tasas de mortalidad reales, α_j los efectos de las variables de estratificación (en nuestro caso edad y período), y $\boldsymbol{\beta}$ (β_1, \dots, β_p) el vector de tamaño p de los coeficientes de regresión de la variable de interés (en nuestro caso la provincia).

Para el ajuste de los modelos con GLIM, hemos utilizado dos ficheros, uno con el número de casos para cada grupo de edad y período y otro con las personas-tiempo para cada edad y período calculados en el punto medio de cada intervalo. La variable PROVINCIA se declaró como lo que en GLIM se denomina FACTOR, ya que es necesario crear variables indicadoras dicotómicas y estimar la razón de tasas para cada provincia. La EDAD y PERIODO (categorizadas) se incluyeron como una única variable en el modelo.

El modelo que ajustamos con GLIM tiene la forma

$$\ln E(d_{jk}) = \ln(n_{jk}) + \alpha_j + \mathbf{x}_{jk}\boldsymbol{\beta}$$

La transformación logarítmica de las tasas lo convierte en una función lineal. El logaritmo de las personas-año se declara como OFFSET forzando que su coeficiente sea 1 por ser una constante.

ANNEX D

Poisson regression models and overdispersion

The general form of the Poisson model is

$$\ln T_{jk} = \alpha_j + \mathbf{x}_{jk}\boldsymbol{\beta}$$

where j and k are the different exposure categories, T_{jk} the real mortality rates, α_j the effects of the stratification variables (in our case age and period), and $\boldsymbol{\beta}$ (β_1, \dots, β_p) the p -dimensional vectors of the regression coefficients of the variable of primary interest (in our case the province).

To fit the models with GLIM, two data records were used, one containing the age- and period-specific case counts, and the other the person-years for each age and period calculated at the midpoint of each interval. The PROVINCE variable was declared as what is known as a FACTOR in GLIM parlance, since it is necessary to create dummy variables and estimate the rate ratios for each province. AGE and PERIOD (categorized) were included as a single variable in the model.

The model fitted with GLIM assumes the form

$$\ln E(d_{jk}) = \ln(n_{jk}) + \alpha_j + \mathbf{x}_{jk}\boldsymbol{\beta}$$

Logarithmic transformation of the rates converts these into a linear function. The log persons-years denominator is declared an OFFSET, forcibly making its coefficient 1 (it being a constant).

El programa proporciona una evaluación de la bondad de ajuste del modelo de Poisson a través de la «deviance». La «deviance» en los modelos de Poisson es un log-likelihood ratio test que compara el modelo ajustado con el modelo saturado (con un parámetro para cada observación). Otra medida de bondad de ajuste es el «log-likelihood ratio test» que compara la distribución observada con la determinada por el modelo (denominado en las tablas χ^2). Los grados de libertad son el número de estratos menos el número de parámetros en el modelo. Si la deviance o el χ^2 exceden los grados de libertad se dice que el modelo es inadecuado pudiendo tratarse de un problema de sobredispersión.

Dado el tamaño relativamente grande de los estratos, la utilización de los errores estándar derivados directamente de los modelos de Poisson, sugeriría la existencia de diferencias estadísticamente significativas en (prácticamente) todas las comparaciones realizadas. Sin embargo, en la mayoría de los casos, el valor de la deviance del modelo de Poisson es muy superior al de los grados de libertad, lo cual implica una gran varianza residual no tenida en cuenta por el modelo. Por lo tanto es necesario «corregir» los errores estándar obtenidos en el ajuste incorporando esa variación residual que las tasas reales manifiestan.

El término sobredispersión indica que la varianza de la variable dependiente excede la varianza nominal, en nuestro caso la varianza en Poisson, es decir, $\text{var}(d) > E(d)$. Este fenómeno es muy común y lo excepcional es que no esté presente. La sobredispersión suele ocurrir porque no se incluyen en el análisis una o más variables explicativas importantes, o cuando las tasas en los distintos estratos estudiados no son independientes entre sí.

On the basis of the «deviance», the program gives an evaluation of the Poisson model's goodness of fit. In Poisson models, deviance is a log-likelihood ratio test that compares the fitted model to the saturated model (with a parameter for each observation). Another measure of goodness of fit is the log-likelihood ratio test which contrasts the observed and fitted values (denominated χ^2 in the tables). The degrees of freedom equal the number of strata minus the number of parameters in the model. Where deviance or χ^2 exceeds its degrees of freedom, the fit is said to be inadequate, with overdispersion possibly being the underlying problem.

Given the relatively large size of the strata, the use of standard errors directly derived from the Poisson models would suggest the existence of statistically significant differences in (practically) all comparisons run. In the majority of cases however, the deviance value of the Poisson model is far higher than the degrees of freedom, implying a high residual variance which the model failed to take into account. Hence, it becomes necessary to «correct» the standard errors obtained in the fit by incorporating the residual variation manifested in the real rates.

The term «overdispersion» indicates that the variance of the dependent variable exceeds nominal variance, in our case the Poisson-specified variance, viz., $\text{var}(d) > E(d)$. This phenomenon is very common; indeed, its absence rather than its presence is the exception. Overdispersion tends to occur because one or more important explanatory variables have not been included in the analysis, or when the rates studied in the different strata under review are not independent of each other.

Para valorar la existencia de sobredispersión se ha seguido un criterio uniforme. Se adoptó como umbral una ji-cuadrado que superase más de un 10% al valor de los grados de libertad. Si bien éste no es el único criterio diagnóstico de sobredispersión, su aplicación nos hace situarnos en la posición más conservadora. El efecto de la corrección es mínimo en los estimadores puntuales.

El método utilizado está basado en una generalización (18) del procedimiento II de Breslow para modelos de Poisson (17). Se ajusta una distribución de Poisson en la que la varianza es $\mu_i + \sigma_i^2 \mu_i^2$ en vez del usual μ_i . Esto representa ajustar por quasi-verosimilitud un modelo equivalente a una binomial negativa (gamma-Poisson).

En la tabla D-1 mostramos las deviancias, la χ^2 para los modelos en los que evaluamos la tendencia temporal provincial (incluyendo el término de interacción) y si se ha utilizado la corrección por sobredispersión en esa causa y sexo. Se aplicó la corrección en una gran parte de las causas estudiadas, siendo la sobredispersión superior en las causas más frecuentes.

Nivel de referencia en los estimadores del efecto provinciales. La promediación ponderada a la unidad de los estimadores del efecto provinciales se calculó de la siguiente manera

$$\log(\text{RR}_p) = \beta_p - \frac{\sum_{p=1}^{52} \beta_p w_p}{\sum_{p=1}^{52} w_p}$$

siendo w_p la población total de la provincia p. El nivel de referencia siguiendo este método queda determinado por la media geométrica de los estimadores. Este procedimiento es equivalente a utilizar la forma de codificación denominada «desviación de la media» (deviation from means coding) que incluyen algunos paquetes de análisis estadístico (21).

To assess the extent of overdispersion a uniform criterion was employed. The threshold adopted was, a chi-square that exceeded the degrees of freedom by over 10%. While not the only diagnostic criterion for overdispersion, its application allows one to take up the most conservative stance. Correction has a minimal effect on the point estimates.

The method employed is based on a generalization (18) of Breslow's procedure II for Poisson models (17). A Poisson distribution is fitted in which the variance is $\mu_i + \sigma_i^2 \mu_i^2$ instead of the usual μ_i . This amounts to fitting a model equivalent to a binomial negative (gamma-Poisson) by means of quasi-likelihood methods.

Shown in Table D-1 are: the deviances; the χ^2 for the models in which provincial time trends (including the interaction term) were evaluated; and an indication as to whether correction was made for overdispersion in the respective cause and sex. Correction was applied to a large proportion of causes studied, with overdispersion proving higher in the most frequent causes.

Reference level in provincial effect estimates. Weighted provincial effect estimates averaged to unity, were calculated as follows:

$$\log(\text{RR}_p) = \beta_p - \frac{\sum_{p=1}^{52} \beta_p w_p}{\sum_{p=1}^{52} w_p}$$

w_p being the total population of province p. With this method, the reference level is determined by the geometric mean of the estimates. This procedure is equivalent to using deviation from means coding, included in some statistical analysis packages (21).

Tabla D-1. Bondad de ajuste de los modelos de regresión Poisson (sin corrección), indicando en los que se ha corregido la extra-dispersión (grados de libertad 1662, igual para todas las causas).

Tabla D-1. Goodness of fit of the Poisson regression models (uncorrected), indicating those corrected for overdispersion (1662 degrees of freedom, equal for all causes).

CAUSA/CAUSE	Hombres / Men			Mujeres / Women		
	Deviance	χ^2	Correc. ^a	Deviance	χ^2	Correcc.
C. BUCAL Y FARINGE / BUCCAL CAV. & PHARYNX	7547	8681	1	1913	1651	-
ESOFAGO / ESOPHAGUS	6707	7862	1	1757	1470	-
ESTOMAGO / STOMACH	8644	10040	1	4011	4346	1
COLON / COLON	4221	4678	1	3601	3929	1
RECTO / RECTUM	2753	3074	1	2621	2823	1
VESICULA / GALL-BLADDER	1816	1831	-	2475	2643	1
PANCREAS / PANCREAS	4351	5059	1	2777	3012	1
PERITONEO / PERITONEUM	1788	1727	-	1925	1846	-
FOSAS NASALES / NASAL FOSSAE	1354	1029	-	973	616	-
LARINGE / LARYNX	9391	11115	1	1171	851	-
PULMON / LUNG	36516	41974	1	3194	3597	1
PLEURA / PLEURA	1379	1219	-	1253	967	-
HUESOS / BONES	1870	1966	-	1689	1568	-
T. CONJUNTIVO / CONNECTIVE TISSUE	1662	1522	-	1622	1399	-
MELANOMA MALIGNO / MELANOMA	1836	1786	-	1631	1533	-
PIEL / SKIN	1917	1714	-	7524	1229	-
MAMA / BREAST	206	147	-	14413	17849	1
UTERO / UTERUS				5418	6365	1
OVARIO / OVARY				4725	5151	1
PROSTATA / PROSTATE	6378	6376	1			
TESTICULO / TESTIS	1613	1271	-			
VEJIGA / BLADDER	5186	5713	1	1819	1708	-
RINON / KIDNEY	2828	3079	1	1896	1833	-
ENCEFALO / BRAIN	4630	5018	1	3633	3920	1
OTROS TUMORES DE S. NERVIOSO / OTHER TUMOURS OF NERVOUS SYS.	1264	996	-	1259	886	-
TIROIDES / THYROID	1455	1182	-	1434	1287	-
TUMORES MAL DEFINIDOS / ILL-DEFINED TUMOURS	6436	7049	1	3877	4332	1
LINFOMA NO HODGKIN / NON-HODGKIN'S LYMPHOMA	2208	2328	1	2037	2119	1
LINFOMA DE HODGKIN / HODGKIN'S DISEASE	1710	1816	-	1770	1589	-
MIELOMA MULTIPLE / MULTIPLE MYELOMA	2237	2292	1	2079	2129	1
LEUCEMIAS / LEUKEMIAS	2256	2294	1	2019	2099	1
LEUCEMIA LINFATICA CRONICA / CHRONIC LYMPHATIC LEUKEMIA	1426	1273	-	1270	1018	-
DIABETES / DIABETES	5155	5352	1	7580	7581	1
E. ISQUEMICA CORAZON / ISCHEMIC HEART DISEASE	27418	31287	1	10169	10377	1
E. CEREBROVASCULAR / CEREBROVASCULAR DISEASE	19367	19192	1	22510	21203	1
ATEROSCLEROSIS / ARTERIOSCLEROSIS	7656	4645	1	25348	4668	1
NEUMONIA / PNEUMONIA	7413	6650	1	11385	6517	1
EPOC / COPD	9632	9812	1	4323	3775	1
CIRROSIS HEPATICA / HEPATIC CIRRHOSIS	23430	27142	1	7034	7877	1
E. RENAL / RENAL DISEASE	3788	3511	1	3427	3045	1
ACCIDENTES TRAFICO / TRAFFIC ACCIDENTS	6235	5903	1	2693	2748	1
CAIDAS ACCIDENTALES / ACCIDENTAL FALLS	3585	3715	1	4986	3047	1
SUICIDIO / SUICIDE	2344	2459	1	1967	2078	-
TUMORES MALIGNOS / MALIGNANT TUMOURS	73733	80283	1	31056	35233	1
E. CARDIOVASCULARES / CARDIOVASCULAR DISEASES	54986	55822	1	54301	53024	1
E. A. RESPIRATORIO / DISEASES OF RESPIRATORY SYS.	14725	14738	1	13112	11420	1
E. A. DIGESTIVO / DISEASES OF DIGESTIVE SYS.	19337	21944	1	7132	7366	1
ANOMALIAS CONGENITAS / CONGENITAL ANOMALIES	1642	1648	-	1534	1508	-
CAUSAS EXTERNAS / EXTERNAL CAUSES	10629	10230	1	6041	6076	1
CAUSAS MAL DEFINIDAS / ILL-DEFINED CAUSES	10380	9066	1	24038	10365	1
TODAS LAS CAUSAS / ALL CAUSES	100009	100517	1	112775	114409	1

^a Corrección por sobredispersión («1» sí «-» no). /^a Correction for overdispersion («1» yes «-» no)