Modelos Lineales con R: Unidad 4

Andrés García Medina
Facultad de Ciencias, Universidad Autónoma de Baja California
https://sites.google.com/view/andresgm/home

Enero-Junio, 2024

1. Propiedades generales de los GLM
2. Modelos log-lineales
3. Sobredispersión

Distribución de muestras grandes para $\boldsymbol{\beta}$

Los resultados distribucionales de GLM no son exactos, en su lugar están basados en aproximaciones de muestras grandes, y hacen uso de propiedades generales de los MLE.

De las propiedades generales de los MLE, tenemos en el límite de muestras grandes:

$$\hat{\boldsymbol{\beta}} \sim N(\beta, I^{-1}), \tag{1}$$

donde $I = \left[\mathbb{E}(-\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T})\right] \approx \mathbf{X^T U X}$ es la *matriz de información* de los parámetros del modelo, siendo $\mathbf{X}$ la la matriz de diseño y $\mathbf{U}$ los pesos óptimos a los cuales converge el algoritmo IRLS, con lo que se obtiene

$$\hat{\boldsymbol{\beta}} \sim N(\beta, (\mathbf{X'UX})^{-1}), \tag{2}$$

Esta expresión nos permite encontrar intervalos de confianza a través del parámetro de escalamiento ϕ.

Ver Jamboard: Distribuciones de muestras grandes para beta

Prueba de la razón de la verosimilitud

Suponga que tenemos dos modelos, M_0 y M_1, y considere probar

$$H_0 : \mathbf{g}(\mu) = \mathbf{X_0}\beta_\mathbf{0} \tag{3}$$
$$H_1 : \mathbf{g}(\mu) = \mathbf{X_1}\beta_\mathbf{1} \tag{4}$$
$$\tag{5}$$

donde μ es el valor esperado de la respuesta $\mathbf{Y}$, cuyos elementos son v.a. independientes provenientes de la misma familia de distribuciones exponenciales y donde $\mathbf{X_0} \in \mathbf{X_1}$

Si tenemos una observación $\mathbf{y}$ del vector respuesta, es posible aplicar una prueba de la razón de verosimilitud generalizada.

Ver Jamboard: GLRT

Prueba de la razón de la verosimilitud

Sea $l(\hat{\beta}_0)$ y $l(\hat{\beta}_1)$ las verosimilitudes maximizadas de ambos modelos.

Si H_0 es verdadera entonces en el límite de muestras grandes se cumple:

$$2[l(\hat{\beta}_1) - l(\hat{\beta}_0)] \sim \chi^2_{p_1 - p_0} \tag{6}$$

donde p_i es el número de parámetros identificables $\beta_{\mathbf{i}}$ en el modelo M_i.

Prueba de la razón de la verosimilitud

Si la hipótesis nula es falsa entonces el modelo M_1 tendrá una cantidad substancialmente mayor de verosimilitud respecto al modelo M_0.

De esta manera, el doble de la diferencia entre verosimilitudes sería demasiado grande para ser consistente con la distribución χ^2

Este resultado aproximado es útil siempre y cuando la verosimilitud pueda ser calculada.

En el caso de los GLMs estimados con IRLS esto es posible sólo si el parámetro de escala ϕ es conocido.

En otras palabras, este resultado puede ser directamente utilizado con los modelos de Poisson y Binomial, pero no con el modelo distribucionales normal, gamma o gaussiano inverso.

Devianza

El concepto de *devianza* en GLM se utiliza para cuantificar la bondad de ajuste.

De manera equivalente a como la suma residual de cuadrados (RSS) se utiliza en los modelos gaussianos lineales.

La devianza de un modelo ajustado se define como

$$D = 2\{l(\hat{\beta}_{max}) - l(\hat{\beta})\}\phi \quad (7)$$

$$= \sum_{i=1}^{n} 2w_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)] \quad (8)$$

donde $l(\hat{\beta}_{max})$ es la logverosimilitud maximizada del modelo saturado

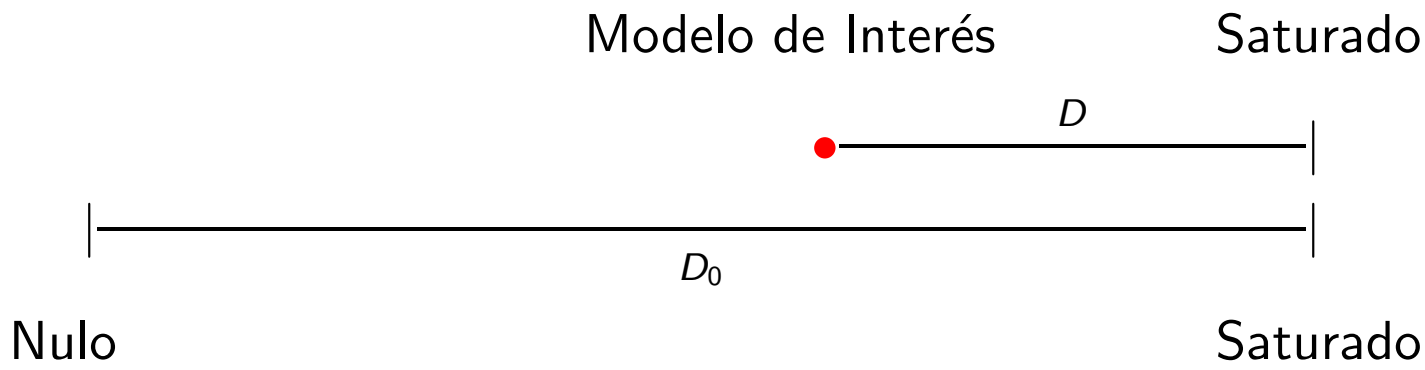
Por su parte, $l(\hat{\beta})$ es la logverosimilitud maximizada del modelo de interés.

Nota: un modelo saturado se considera aquel que contine un parámetro por cada observación, en este caso n parámetros, y se evalúa con $\hat{\mu} = \mathbf{y}$

La devianza nos ayuda a medir "distancias" entre modelos

Modelo saturado, modelo nulo

- Los modelos nulo y saturado son dos modelos extremos.
- En el nulo, el predictor lineal, $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$, es de la forma $\eta_i = \beta_0$ y, básicamente, lo que suponemos es que $y_1, \cdots, y_n$ son i.i.d. $f(y; \theta)$, donde θ es un sólo parámetro.
- Por otro lado, en el modelo saturado tenemos un número maximal de parámetros (puede haber hasta el máximo número: n), que, por supuesto, será el modelo que mejor ajuste a los datos.
- El modelo de interés es un modelo intermedio:



devianza

La *devianza escalada* se define como:

$$D^* = D/\phi \tag{9}$$

la cual no depende del parámetro de escalamiento.

Para el caso de las distribuciones binomial y de poisson, donde $\phi = 1$, la devianza y devianza escalada coiniciden.

Es fácil ver que el LRT se puede formular como

$$D_0^* - D_1^* \sim \chi^2_{p_1 - p_0} \tag{10}$$

donde D_0^* y D_1^* son las versiones escaladas de M_0 y M_1, respectivamente.

Esto esta expresión es útil sólo si es parámetro de escalamento ϕ es conocido de tal manera que se pueda calcular D^*

Comparación de modelos con ϕ desconocida

Bajo H_0 se tiene el resultado aproximado

$$D_0^* - D_1^* \sim \chi_{p_1-p_0}, \quad D_1^* \sim \chi^2_{n-p} \tag{11}$$

además, si $D_0^* - D_1^*$ y D_1^* se consideran asintóticamente independientes implica que

$$F = \frac{(D_0^* - D_1^1)/(p_1 - p_0)}{D_1^*/(n - p_1)} \sim F_{p_1-p_0, n-p_1}, \tag{12}$$

en el límite de muestras grandes (lo cual es exacto en el modelo lineal ordinario).

La propiedad útil del estadístico F es que puede ser calculado sin necesidad de conocer ϕ, puesto que se cancelan los términos obtiendo bajo H_0

$$F = \frac{(D_0 - D_1)/(p_1 - p_0)}{D_1/(n - p_1)} \sim F_{p_1-p_0, n-p_1}, \tag{13}$$

Limitaciones de la devianza

Dado el resultado del LRT generalizado

$$2[l(\hat{\beta}_1) - l(\hat{\beta}_0)] \sim \chi^2_{p_1 - p_0} \tag{14}$$

Se espera que, si el modelo es correcto, se comporte aproximadamente como

$$D^* \sim \chi^2_{n-p} \tag{15}$$

en el límite de muestras grandes.

Sin embargo el argumento es un poco fraudulento, dado que el LRT generalizado se obtiene en el límite cuando el número de parámetros es fijo, mientras que el tamaño de la muestra tiende a infinito.

No obstante, el modelo saturado tiene la misma cantidad de parámetros que de observaciones por lo que el supuesto distribucional de D_1^* es dudoso.

Estimación de $\hat{\phi}$ a través de la devianza

Aún así, este resultado es utilizado también para estimar ϕ cuando no se conoce a priori.

El valor esperado de una v.a. χ^2_{n-p} es $n-p$.

Aproximado $D^* = D/\phi$ a su valor esperado tenemos

$$\hat{\phi}_D = \frac{\hat{D}}{n-p} \tag{16}$$

Estimación de $\hat{\phi}$ a través del estadístico de Pearson

Por otro lado, un segundo estimador se basa en el estadístico de Pearson

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}, \quad V(\hat{\mu}_i) = \frac{Var(Y)}{\phi} \tag{17}$$

vemos que X^2/ϕ es la suma de v.a. al cuadrados con media cero y varianza uno con $n - p$ grados de libertad.

De esta manera, si el modelo es adecuado podemos aproximar $X^2/\phi \sim \chi^2_{n-p}$

Ahora, si igualamos respecto a su valor esperado, obtenemos

$$\hat{\phi} = \frac{\hat{X}^2}{n - p} \tag{18}$$

Residuales

Similarmente, existen varias definiciones de residuales en GLM relacionadas a cada una de las estimación del parámetro de dispersión.

El primero son los residuales de Pearson

$$\hat{\epsilon}_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \tag{19}$$

el cual debe tener una media aproximada de cero y una varianza de ϕ si el modelo es correcto.

Estos residuales no deberían de mostrar ninguna tendencia en la media o varianza cuando se grafican respecto a los valores ajustados o las covariables.

Sin embargo, en la práctica estos valores pueden ser asimétricos alrededor de cero por lo que difieren de los residuales asociados a los modelos ordinarios.

Residuales

Debido a esto se prefiere los residuales de la devianza.
La devianza tiene el mismo rol en los GLM que la suma residuales de cuadrados (RSS) en los modelos lineales ordinarios.

De hecho, la devianza se reduce a RSS en el modelo lineal ordinario.

La relación exacta lo podemos establecer a través de la siguiente definición

$$\hat{\epsilon}_i^d = sign(y_i - \hat{\mu}_i)\sqrt{d_i} \tag{20}$$

donde d_i es la devianza que contribuye la observación i-ésima de tal manera que $D = \sum_i d_i$.

Se puede ver que la suma de cuadrados de $\hat{\epsilon}_i^d$ nos da la devianza

Se espera que estos residuales se comporten como $N(0, 1)$ cuando la aproximación $D^* \sim \chi^2_{n-p}$ es razonable.

Cuasi-verosimilitud

Para algunos datos, una distribución de la exponencial no será apropiada.

Por ejemplo, supongamos que tenemos datos de conteo (como para una respuesta de Poisson), pero la varianza de los datos no es igual a la media (que Poisson supone que es el caso).

Para ajustar los GLM a tales datos, se necesita un enfoque más flexible.

La idea de la estimación de cuasi-verosimilitud es utilizar una función diferente (que la función de verosimilitud) para obtener las estimaciones de los parámetros.

A diferencia de la estimación de probabilidad (que requiere especificar la distribución de la respuesta), la estimación de cuasi-verosimilitud solo requiere especificar las funciones de media y varianza.

Funciones de enlace canonicas

- $\theta_i = \mathbf{X_i}\beta$
- $\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} w_i \left(y_i \frac{\partial \theta_i}{\partial \beta_j} - \mu_i \frac{\partial \theta_i}{\partial \beta_j} \right) = 0$, (si $a_i(\phi) = \phi / w_i$)
- Si el eblace canonica es usado entonces $\frac{\partial \theta_i}{\partial \theta_j} = X_{ij}$
- Este sistema ecuaciones se reduce entonces a $\mathbf{X^T y} - \mathbf{X^T}\mu = 0$
- Esto implica que $\sum_i y_i = \sum_i \hat{\mu}_i$
- Además los residuales suman cero.

Ejemplo: Modelos binomiales y ataques al corazón

- Se cree que el nivel de la enzima en sangre creatinina quinasa (CK) es relevante para el diagnóstico temprano de ataques cardíacos.
- El conjunto de datos a continuación proporciona los niveles de CK y los resultados del ataque cardíaco (es decir, recuentos) para n = 360 pacientes de un estudio de Smith (1967).
- Tenga en cuenta que ck es el nivel de CK, ha es el número de pacientes que sufrieron un ataque al corazón y ok es el número de pacientes que no sufrieron un ataque al corazón.
- Queremos construir un modelo que explique la proporcion de los pacientes que sufrieron un ataque al corazon dado el nivel de CK

Ver Jupyter: Modelos binomiales y ataques al corazón

Modelos log-lineales

- Estos modelos aplican regresión Poisson a tablas de contingencia en lugar de conteos
- La idea es ganar intuicion del modelo a través de un ejemplo guiado.
- La siguiente tabla clasifica una muestra aleatoria de mujeres y hombres según su creencia en la vida después de la muerte

	Creyente	No creyente
Mujer	435	147
Hombre	375	134

- Los datos (reportados en Agresti, 1996) provienen de la Encuesta Social General de Estados Unidos (1991), y la categoría de "no creyentes" incluye a los "indecisos".
- Pregunta de investigacion: ¿Existen diferencias entre hombres y mujeres en sus creencias?

Modelos log-lineales

- Podemos abordar esta cuestión utilizando el análisis de la devianza para comparar el ajuste de dos modelos competitivos de estos datos
- Uno en el que la creencia se modela como independiente del género y un segundo en el que hay cierta interacción entre la creencia y el género.
- Consideremos primero el modelo de independencia.
- Si y_i es una observación de los recuentos en una de las celdas de la tabla, entonces podríamos modelar el número esperado de recuentos como

$$\mu_i = \mathbb{E}(Y_i) = n\gamma_k\alpha_j \tag{21}$$

  si y_i es el dato para el genero k y creencia j
- donde n es el número total de personas encuestadas, α_1 la proporción de creyentes, α_2 la proporción de no creyentes y γ_1 y γ_2 las proporciones de mujeres y hombres, respectivamente

Modelos log-lineales

- Tomando logaritmos de este modelo

$$\eta_i = log(\mu_i) = log(n) + log(\gamma_k) + log(\alpha_j) \tag{22}$$

- Si definimos $\tilde{n} = log(n)$, $\tilde{\gamma}_k = log(\gamma_k)$, $\tilde{\alpha}_k = log(\alpha_k)$, podemos escribir

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \hat{\eta} \\ \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} \tag{23}$$

- ¿El modelo es identificable?

Modelos log-lineales

- Removiendo $\tilde{\gamma}_1$ y $\tilde{\alpha}_1$ resuelve el problema de identificabilidad

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \hat{\eta} \\ \hat{\gamma}_2 \\ \hat{\alpha}_2 \end{pmatrix} \tag{24}$$

- Tenemos que la creencia y el genero son variables del tipo factor con dos niveles cada una.
- Si los recuentos de la tabla de contingencia ocurrieran de forma independiente y aleatoria, entonces la distribución a utilizar sería Poisson.
- De hecho, incluso cuando el número total de sujetos en la tabla, o incluso algunos otros totales marginales, son fijos, se puede demostrar que la probabilidad correcta puede escribirse como un producto de las p.d.f. de Poisson, condicionadas a las distintas cantidades fijas.
- **Ver Jupyter**: Modelos loglineal

Sobredispersión

- Los modelos Binomial y Poisson imponen una relación entre la media y la varianza, e.g. si $E(Y) = \mu$ entonces $\text{Var}(Y) = (1 - \mu/n)\mu$ para la Binomial, y $\text{Var}(Y) = \mu$ para el caso Poisson.
- Sin embargo, en las aplicaciones, probablemente la sobredispersión es la norma más que la excepción.
- Sobredispersión es el caso en el que la varianza observada sobrepasa la varianza teórica postulada por el modelo.
- Una posible causa: Heterogeneidad en la población, por ejemplo, si tenemos Y_i's afectadas por alguna covariable latente binaria, Z_i, tal que

$$Y_i|Z_i = 0 \sim \mathcal{P}(\lambda_0), \qquad Y_i|Z_i = 1 \sim \mathcal{P}(\lambda_1)$$

en este caso ocurre sobredispersión.

Sobredispersión

- Supongamos observaciones Y_i, pero no estamos conscientes de la existencia de Z_i

$$\begin{aligned} E(Y) =& E_X[E(Y|X)] \\ \Rightarrow E(Y_i) =& E_{Z_i}E(Y_i|Z_i) \\ =& E(Y_i|Z_i=0)P(Z_i=0) + E(Y_i|Z_i=1)P(Z_i=1) \\ =& \lambda_0(1-\pi) + \lambda_1\pi \equiv \mu \end{aligned}$$

- Por otro lado (ley de la varianza total):

$$\begin{aligned} Var(Y) =& E_X(Var(Y|X)) + Var_X(E(Y|X)) \\ \Rightarrow \text{Var}(Y_i) =& E_{Z_i}\text{Var}(Y_i|Z_i) + \text{Var}_{Z_i}E(Y_i|Z_i) \\ =& E[\lambda_1 Z_i + \lambda_0(1-Z_i)] + \text{Var}[\lambda_1 Z_i + \lambda_0(1-Z_i)] \\ =& \lambda_1\pi + \lambda_0(1-\pi) + \lambda_1^2\pi(1-\pi) + \lambda_0^2\pi(1-\pi) \\ & - 2\lambda_0\lambda_1\pi(1-\pi) = \mu + (\lambda_0-\lambda_1)^2\pi(1-\pi) > \mu \end{aligned}$$

Sobredispersión

- El desconocimiento de la existencia de factores no observables que afectan a la variable Y_i nos llevaría probablemente a postular un modelo Poisson cuya varianza sobrepasa la implicada por el modelo.
- En modelos lineales generalizados tenemos,

$$E(Y) = \mu = b'(\theta)$$
$$\text{Var}(Y) = a(\phi)b''(\theta) = V(\mu)$$

- En el caso Poisson, por ejemplo, $a(\phi) = 1$ y $V(\mu) = \mu$, pero, con frecuencia, $E(Y) > \text{Var}(Y)$.
- En estos casos, se modela a la varianza como $\text{Var}(Y) = \phi V(\mu)$ donde ϕ captura la sobredispersión.
- Sin embargo, no podemos hacer máxima verosimilitud, pero podemos usar **cuasiverosimilitud**, la cual comparte propiedades asintóticas similares que la verosimilitud.

Ejemplo: Polio

- Los datos de la tabla siguiente fueron tomados de
  - Zeger, S.L. (1988). A regression model for time series of counts *Biometrika*, Vol.75, No.4, pp. 621-629.

- La pregunta de interés es si estos dan evidencia de un decrecimiento en los casos de polio en el tiempo.

Table 2. *Monthly number of U.S. cases of poliomyelitis for* 1970 *to* 1983

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1970	0	1	0	0	1	3	9	2	3	5	3	5
1971	2	2	0	1	0	1	3	3	2	1	1	5
1972	0	3	1	0	1	4	0	0	1	6	14	1
1973	1	0	0	1	1	1	1	0	1	0	1	0
1974	1	0	1	0	1	0	1	0	1	0	0	2
1975	0	1	0	1	0	0	1	2	0	0	1	2
1976	0	3	1	1	0	2	0	4	0	2	1	1
1977	1	1	0	1	1	0	2	1	3	1	2	4
1978	0	0	0	1	0	1	0	2	2	4	2	3
1979	3	0	0	2	7	8	2	4	1	1	2	4
1980	0	1	1	1	3	0	0	0	0	1	0	1
1981	1	0	0	0	0	0	1	2	0	2	0	0
1982	0	1	0	1	0	1	0	2	0	0	1	2
1983	0	1	0	0	0	1	2	1	0	1	3	6

Reported to the U.S. Centers for Disease Control and published in Morbidity and Mortality Weekly Report Annual Summary (1970-1983).

Ver Jupyter: Modelos Poisson y polio