IIMAS | Especialización en estadística aplicada
Análisis de datos categóricos | 2024-2 | 18 de mayo de 2024

CAT - 05) Modelos lineales generalizados

Eduardo García Tapia

5.1) Introducción

El texto de referencia es Agresti, 2002 - Categorical Data Analysis

Los modelos lineales generalizads (GLM, por sus siglas en inglés) se utilizan para modelar variables de respuesta a través de múltiples variables predictoras. Este documento se enfoca en las variables de respuesta cetogóricas.

Los GLM extienden el análisis de regresión ordinaria y permiten modelar variables de respuesta que no se distribuyen normalmente.

Tres componentes especifican un GLM:

i. Componente aleatorio: Identifica a la variable de respuesta Y y su distirbución de probabilidad
ii. Componente sistemático: Especifíca a las variables explicativas utilizadas en una función de predicción lineal
iii. Función de enlace: Especifíca la función de $E(Y)$ que el modelo iguala con la función de predicción.

5.2) Componentes de un GLM

El componente aleatorio consiste de la viarable Y con observaciones independientes $(y_1, ..., y_n)$ de una distribución en la familia de distribuciones exponencial. Esta familia tiene función de densidad de probabilidad (PDF) de la forma:

$$f(y|\theta) = a(\theta)b(y)\, exp\,[y\, Q(\theta)]$$

En la expresión anterior $Q(\theta)$ es llamado el parámetro natural.

La componente sistemática de un GLM se relaciona con un vector de variables explicativas $(\eta_1, ..., \eta_n)$. Denotamos x_{ij} al valor de la variable predictora \$j ; (j=1,2,. . . ,p) para el sujeto i. Entonces:

$$\eta_i = \sum_{j=1}^{p} \beta_j x, \qquad i = 1, ..., N$$

La combinación lineal de las variables explicativas es llamado el predictor lineal. Para el coeficiente de intercepto se considera $x_{i,j} = 1$ para alguna j.

La tercera componente de un GLM es la función de enlace, que conecta la componente aleatoria con la sistemática. Suponga $\mu_i = E(Y_i)$. El modelo conecta μ_i con η_i a través de $\eta_i = g(\mu_i)$, donde g es la función de enlace, que debe es monótona y diferenciable.

La función de enlace identidad $\eta_i = \mu_i$ especifíca un modelo para la media en sí misma. Es la que se utiliza para el análisis de regresión lineal con Y_i normalmente distribuida.

La función de enlace que transforma la media con el parámetro natural es llamada la *enlace canónico*. Por lo que:

$$Q(\theta_i) = \sum_{j=1}^{p} \beta_j x_{i,j}$$

Usualmente al intercepto se le denomina como α, por lo que el modelo GLM es de la forma:

$$Q(\theta_i) = \alpha + \sum_{j=1}^{k} \beta_j x_j$$

En resumen, un GLM es un modelo lineal para predecir la media transformada de una distribución en la familia exponencial.

5.3) Devianza

Suponga un GLM para un conjunto de observaciones $(y_1, ..., y_n)$. Sea $L(\boldsymbol{\mu}; \boldsymbol{y})$ la función de log-verosimilitud expresada en términos de la media $\boldsymbol{\mu} = (\mu_1, ..., \mu_N)$. Suponga un modelo saturado, donde se tienen tantos parámetros estimados como observaciones, donde $\hat{\boldsymbol{\mu}} = \boldsymbol{y}$. O sea, que la media de cada observación es estimada a partir del valor observado, logrando la máxima log-verosimilitud posible. Este modelo no sería útil, porque no proveé reducción de datos.

Entonces, la devianza se define como:

$$D = -2\left[L(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) - L(\boldsymbol{y}; \boldsymbol{y}))\right]$$

Observe que la log-verosimilitud del modelo ajustado $(\hat{\boldsymbol{\mu}})$ siempre será menor o igual que la del modelo saturado, por lo que para un modelo dado la devianza toma un valor igual o mayor que 0.

Por otro lado, el peor modelo que se podría ajustar solo considera el intercepto (α). La *devianza nula* D_0 se define como la devianza que existe entre el modelo saturado y el modelo que solo contempla α. La devianza sirve para evaluar la bondad de ajuste del modelo. Observe que se preferirán modelos con D pequeña, lo más alejada posible de la devianza nula.

En particular, para el modelo de regresión lineal, se tiene la siguiente expresión:

$$R^2 = 1 - \frac{D}{D_0} = 1 - \frac{SS_{Res}}{SS_T}$$

Note que para cualquier GLM se puede construir una métrica R^2, pero que no tiene el mismo significado que para la regresión por mínimos cuadrados, pues no esl el porcentaje de variabilidad explicada por el modelo, ni se relaciona con ningún coeficiente de correlación entre variables.

5.3) GLM para datos binarios

Sea Y una variable binaria que toma valores $(0, 1)$. Se define $P(Y = 1) = \pi(x)$, reflejando su dependencia en los valores predictores $\boldsymbol{x} = (x_1, .., x_p)$.

En este caso es de interés modelar $\pi(x)$, pero igualarlo a un predictor lineal (η) tendría deficiencias estructurales, debido a que $\pi(x) \in (0, 1)$ y η no tiene esa restricción. Es por eso que es necesario transformar a $\pi(x)$ para que pueda ser predecido con un regresor lineal. Entonces, el modelo propuesto es con una función de enlace logit (logaritmo de momios), con lo que se obtiene un GLM.

$$log\frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

La enlace *logit* es la canónica para la distribución Binomial. La función *logit* no tiene restricciones sobre su dominio, por lo que es apropiada para modelarse con un regresor lineal. Observe que la función traslado de dominio a $\pi(x)$ para utilizar η. Existen otras funciones que logran el mismo objetivo. Por ejemplo, la función *probit* utiliza la inversa de la función de distribución acumulada de la normal $\Phi^{-1}[\pi(x_i)] = \eta_i$. La función de liguadura *cloglog* es de la forma $log(-log(\pi(x_i)) = \eta_i$

5.4) GLM para conteos

Los GLM más conocidos para modelar datos de conteo son los que asumen $Y \sim Poisson(\mu)$. La función *log* es la función de enlace canónica, mientras que $log(\mu)$ es el parámetro natural. Entonces, el modelo es de la forma:

$$log(\mu) = \eta_i = \alpha + \beta_1 x_1 + ... + \beta_k x_k$$

Un problema análogo al de heterocedasticidad en regresión lineal, para el caso de GLM Poisson es la sobredispersión. Recuerde que $Var(Y) = \mu$, por lo que si los datos muestrales $(y_i, ..., y_n)$ muestra una varianza distinta, se dice que existe sobredispersión. Esto puede suceder cuando en la muestra hay un exceso de valores 0, u otros errores de muestreo. Existen herramientaas para modelar datos con estas características (ver CAT - 06) Regressión Poisson).

5.5) Momentos y verosimilitu de GLM

Se extiende la notación de PDF en la familia exponencia, para cubrir aquellas distribuciones con dos parámetros. La distribuciones dentro de la familia de dispersión exponencial son de la forma:

$$f(y_i, \theta_1, \phi) = exp\left\{\left[y_i\theta_i - b(\theta_i)\right]/a(\phi) + c(y_i, \phi)\right\}$$

Donde ϕ es llamado el parámetro de dispersión y θ_i es el parámetro natural. Cuando ϕ es conocido, la definición se reduce a la que se dio al principio del documento (ver pçágina 133 para detalles). Cuando se trata de distribuciones de un solo parámetro como la Binomial o la Poisson, no es necesaria esta formulación.

Definiendo la función de log-verosimilitud como $K = \sum_i L_i$ y derivando con respecto a θ se puede llegar a los resultados de que:

$$\mu_i = E(Y_i) = b'(\theta_i) \quad ; \quad Var(Y_i) = b''(\theta_i)a(\phi)$$

Adicionalmente, a partir de la función de verosimilitud se define un *score statistic (U)*, la cual se distribuye asintóticamente normal $U \sim N(0, \mathcal{I})$. Por lo tanto, cuando se generaliza y $\theta = (\beta_0, ..., \beta_k)$, entonces cada uno de las componentes se distribuyen asintóticamente normal. Recuerde que para el caso de OLS, los coeficientes siguen una distribución $t_{(n-p)}$

5.6) Bondad de ajuste e inferencia

La devianza es la métrica utilizada para evaluar la bondad de ajuste de los GLM. La devianza se puede interpretar como una medida de *"carencia de ajuste"*, por lo que entre mayor sea, pero se ajusta el modelo. La devianza nula (que considera el modelo que cuenta únicamente con α) es la que tiene una devianza mayor. Al ir agregando variables y, por lo tanto, parámetros, se espera que la devianza disminuya.

Suponga un modelo M_0 que está anidado dentro de M_1. Entonces se tiene la expresión debajo. Lo que quiere decir que al ajustar una variable extra en un GLM, siempre se ajustará un modelo igual o mejor. Dicho de otra manera, modelos más simples tendrán devianzas más grandes.

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_{\mathbf{1}}) \leq D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_{\mathbf{0}})$$

Para probar la significancia de las nuevas variables ajustadas se sabe que, donde r es la diferencia de parámetros entre un modelo y otro.

$$\left[D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_{\mathbf{0}}) - D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_{\mathbf{1}})\right] \sim \chi^2_{(r)}$$

En este caso, la hipótesis nula es que M_0 no ajusta mejor que que M_1. Si se tiene un p-valor muy pequeño es que no hay evidencia suficiente para sostener esta hipótesis, por lo que el modelo reducido es mejor. Entonces, note que cuando la diferencia de devianzas es muy grande se espera que M_0 se un buen modelo.

5.7) Estimación de parámetros

Los métodos para ajustar un GLM son estrictamente numéricos, debido a que se tiene ecuaciones diferenciales parciales no lineales que no tienen una solución analítica. Los métodos más utilizados son Newton-Raphson y el de Puntuación de Fisher. Cuando se trabaja con enlaces canónicos y parámetros naturales, se hacen simplificaciones significativas, que ayudan a resolver las ecuaciones de manera más sencilla y eficiente.