MAT-466: Bondad de ajuste y residuos en GLM

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Recuerde que, en general, $D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})$ no sigue una distribución asintótica $\chi^2(n-p)$. Por ejemplo,

- Poisson: Cuando $\mu_i \to \infty$ para $i = 1,\dots,n$ tenemos $D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}) \sim \chi^2(n-p)$.
- Binomial: $Y_i \sim \text{Bin}(n_i,\mu_i)$, para $i = 1,\dots,k$. Así para k fijo y $n_i \to \infty$, $\forall\, i$ tenemos $D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}) \sim \chi^2(n-p)$.
- Normal: Es bien sabido que para σ^2 fijo, tenemos $D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})/\sigma^2 \sim \chi^2(n-p)$.

En general tenemos que (Jørgensen, 1987)

$$D^*(\boldsymbol{y};\widehat{\boldsymbol{\mu}}) \sim \chi^2(n-p), \qquad \text{cuando } \phi \to \infty$$

Es decir,

- Normal: $\sum_{i=1}^n (y_i - \widehat{\mu}_i)^2/\sigma^2 \sim \chi^2(n-p)$ si $\sigma^2 \to 0$.
- Gamma: $D^*(\boldsymbol{y};\widehat{\boldsymbol{\mu}}) \sim \chi^2(n-p)$ si CV $\to 0$ ($\phi^{-1/2} \to \infty$)

Bondad de ajuste

Sabemos que, si $T \sim \chi^2(n-p)$. Entonces, $\mathsf{E}(T) = n - p$. De este modo,

Regla de trabajo:

Un valor del desvío cercano a $n - p$ puede indicar que el modelo está bien ajustado.

Otra alternativa es usar la estadística chi-cuadrado de Pearson:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

Residuos en regresión lineal

Suponga que

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

El vector de residuos es dado por:

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y},$$

con $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. En nuestra notación, tenemos $\widehat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{Y} = \widehat{\mathsf{E}}(\boldsymbol{Y})$.

En particular,

$$h_{ii} = \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i, \qquad i = 1, \dots, n,$$

y $0 \leq h_{ii} \leq 1$. Además,

$$\overline{h} = \frac{1}{n}\sum_{i=1}^n h_{ii} = \frac{1}{n}\operatorname{tr}(\boldsymbol{H}) = \frac{p}{n}.$$

Residuos en regresión lineal

Bajo el supuesto de normalidad $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$, tenemos

$$\boldsymbol{e} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H})),$$

es decir

$$\mathsf{E}(e_i) = 0, \quad \mathsf{var}(e_i) = \sigma^2(1 - h_{ii}), \quad \mathsf{Cov}(e_i, e_j) = -\sigma^2 h_{ij}.$$

De ahí que los residuos tienen varianzas diferentes y son correlacionados.

El residuo estandarizado es definido como:

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}, \qquad i = 1, \dots, n.$$

Cook y Weisberg (1982) mostraron que

$$\frac{r_i^2}{n - p} \sim \mathsf{Beta}\Big(\frac{1}{2}, \frac{n - p - 1}{2}\Big),$$

de este modo

$$\mathsf{E}(r_i) = 0, \quad \mathsf{var}(r_i) = 1, \quad \mathsf{Cov}(r_i, r_j) = \frac{-h_{ij}}{\sqrt{(1 - h_{ii})(1 - h_{jj})}}.$$

Considere el residuo studentizado:

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}, \qquad i = 1,\dots,n.$$

donde

$$s_{(i)}^2 = \frac{1}{n-p-1}\sum_{j\neq i}^{n}(y_j - \widehat{\mu}_j)^2 = s^2\Big(\frac{n-p-r_i^2}{n-p-1}\Big),$$

sigue que $t_i \sim t(n-p-1)$.

Una interpretación interesante de t_i es que corresponde al estadístico t para probar la hipótesis $H_0 : \gamma = 0$ en el modelo de salto en la media:

$$Y_j = \boldsymbol{x}_j^\top \boldsymbol{\beta} + d_j\gamma + \epsilon_j, \qquad j = 1,\dots,n,$$

donde $d_j = 1$ si $j = i$ y 0 en caso contrario.

Objetivo:

Evaluar desvios de normalidad de los residuos studentizados t_i's.

Notación:

Considere $Z_i = t_i$, para $i = 1, \dots, n$[1]

Idea:

Comparar la CDF muestral para los Z_i's contra la CDF de la N$(0, 1)$.

Asuma que los residuos Z_i están ordenados

$$Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)},$$

los $Z_{(i)}$ son los cuantiles de la CDF muestral, definida como

$$\text{Proportion}(Z \leq Z_{(i)}) = \frac{i}{n}$$

[1] La descripción es válida otras medidas de interés.

QQ-plot en regresión lineal

Asuma que los residuos Z_i están ordenados

$$Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)},$$

los $Z_{(i)}$ son los cuantiles de la CDF muestral, definida como

$$\text{Proportion}(Z \leq Z_{(i)}) = \frac{i}{n}.$$

Los cuantiles de la distribución teórica, son dados por:

$$q_i^* = \Phi^{-1}\Big(\frac{i}{n}\Big).$$

Si los errores son aproximadamente normales, se debe tener que el gráfico de los pares $(q_1^*, Z_{(1)}), \dots, (q_n^*, Z_{(n)})$ sea a recta identidad.

Se ha sugerido la siguiente aproximación para la esperanza de los estadísticos de orden desde $\mathsf{N}(0,1)$ como:

$$q_i = \Phi^{-1}\Big(\frac{i-3/8}{n+1/4}\Big),$$

de este modo, se utilizará el gráfico cuantil-cuantil (QQ-plot) de los pares $(q_i, Z_{(i)})$.

Observación:

- Podemos construir QQ-plots para diversas distribuciones.[2]
- Es difícil chequear visualmente desvios de la distribución de interés.

[2] Por ejemplo, χ^2, t de Student, Poisson, Gama, etc.

QQ-plot con envelopes en regresión lineal

Envelopes simulados son herramientas gráficas para chequear el ajuste de un modelo. Atkinson (1985) sugirió usar el siguiente procedimiento:

- Ajustar un modelo de regresión lineal, calcular residuos y estandarizar para obtener varianza unitaria.
- Generar M (≈ 1000) muestras como respuesta. Para cada muestra ajuste el mismo modelo y calcule los residuos estandarizados
- Ordenar todos los conjuntos de residuos estandarizados.
- El envelope consiste de los cuantiles 2.5% inferior y superior de los residuos estandarizados generados en cada posición.

Ejemplo (Herencia de la estatura):

Se recolectó la altura de $n = 1375$ madres en UK (bajo 65 años) y una de sus hijas adultas (sobre 18 años).

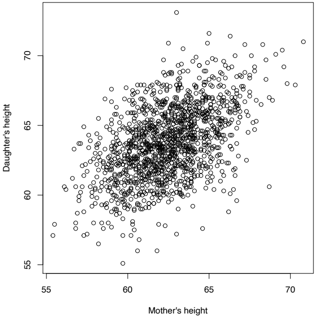
Exploramos el conjunto de datos por medio del gráfico:

```
> plot(dheight ~ mheight, data = Heights)
```

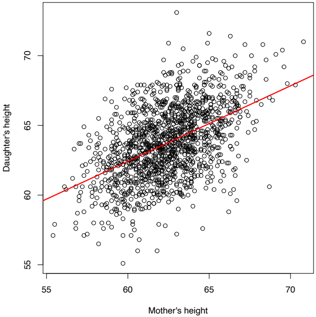
Ajuste de un modelo de regresión lineal simple

```
> fm <- lm(dheight ~ mheight, data = Heights)
> fm

Call:
lm(formula = dheight ~ mheight, data = Heights)

Coefficients:
(Intercept)      mheight
    29.9174       0.5417
```

Herencia de la estatura

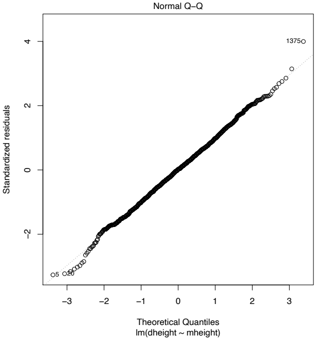


(a) datos estatura

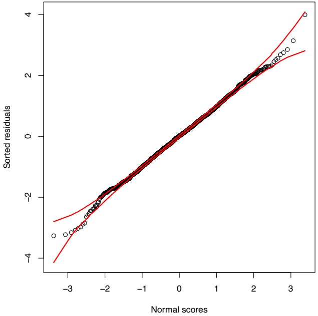


(b) recta ajustada

Herencia de la estatura



(a) datos estatura



(b) recta ajustada

Note que a la convergencia del proceso iterativo,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \widehat{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \widehat{\boldsymbol{W}} \widehat{\boldsymbol{Z}},$$

con $\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{\eta}} + \widehat{\boldsymbol{W}}^{-1/2} \widehat{\boldsymbol{V}}^{-1/2} (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})$. Una de las primeras alternativas ha sido definir

$$\boldsymbol{r}_P = \widehat{\boldsymbol{W}}^{1/2} (\widehat{\boldsymbol{Z}} - \widehat{\boldsymbol{\eta}}) = \widehat{\boldsymbol{V}}^{-1/2} (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}),$$

cuyo i-ésimo elemento es el residuo Pearson

$$r_{P_i} = \frac{y_i - \widehat{\mu}_i}{\widehat{V}_i^{1/2}}, \qquad i = 1, \dots, n,$$

donde $\widehat{V}_i = V(\widehat{\mu}_i)$.

Asumiendo que Cov$(\boldsymbol{Z}) \approx \phi^{-1}\widehat{\boldsymbol{W}}^{-1}$, tenemos Cov$(\boldsymbol{r}_P) \approx \phi^{-1}(\boldsymbol{I} - \widehat{\boldsymbol{H}})$, con

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^\top \widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}^{1/2},$$

podemos definir la versión estandarizada

$$t_{S_i} = \frac{\phi^{1/2}(y_i - \widehat{\mu}_i)}{\widehat{V}_i^{1/2}(1 - \widehat{h}_{ii})^{1/2}}, \qquad i = 1, \dots, n,$$

Considere la función

$$A(\mu) = \int_0^{\mu} \frac{\mathrm{d}t}{V^{1/3}(t)},$$

se ha sugerido el uso de la función $A(\cdot)$ para definir residuos cuya distribución puede ser más cercana de la normal. Esto lleva al residuo Anscombe

$$t_{A_i} = \frac{\phi^{1/2}\{A(y_i) - A(\widehat{\mu}_i)\}}{\widehat{V}_i^{1/2} A'(\widehat{\mu}_i)}.$$

Algunos ejemplos de $A(\cdot)$:

- Normal: μ.
- Binomial: $\int_0^{\mu} t^{-1/3}(1-t)^{-1/3}\,\mathrm{d}t$.
- Poisson: $\frac{3}{2}\mu^{2/3}$.
- Gama: $3\mu^{1/3}$.
- Normal inversa: $\log\mu$.

Residuos en GLM

Considere

$$D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} d(y_i;\widehat{\mu}_i),$$

es decir $d_i = d(y_i;\widehat{\mu}_i)$ es el componente i-ésimo del desvío. Eso lleva a definir el residuo componente de desvio

$$\begin{aligned} r_{D_i} &= \operatorname{sign}(y_i - \widehat{\mu}_i)\sqrt{d_i} \\ &= \operatorname{sign}(y_i - \mu_i)\sqrt{2}\{y_i(\widetilde{\theta}_i - \widehat{\theta}_i) + (b(\theta))\}^{1/2} \end{aligned}$$

para $i = 1, \dots, n$. Esto lleva a la versión estandarizada

$$t_{D_i} = \frac{\phi^{1/2}\operatorname{sign}(y_i - \widehat{\mu}_i)\sqrt{d_i}}{(1 - \widehat{h}_{ii})^{1/2}}, \qquad i = 1, \dots, n.$$

Observación:

El residuo componente de desvío es posiblemente el más usado en GLM.

Un cuarto tipo de residuo fue propuesto por Williams (1987), definido como:

$$t_{W_i} = \operatorname{sign}(y_i - \widehat{\mu}_i)\{(1 - \widehat{h}_{ii})t_{D_i}^2 + \widehat{h}_{ii}t_{S_i}^2\}^{1/2}, \qquad i = 1, \dots, n,$$

que puede ser interpretado como un promedio ponderado entre t_{D_i} y t_{S_i}.

Más recientemente Dunn y Smith (1996) introdujeron el residuo cuantil, definido como:

$$r_{Q_i} = \Phi^{-1}\{F(y_i; \widehat{\mu}_i, \widehat{\phi})\},$$

donde F es la CDF asociada al modelo estadístico FE(θ_i, ϕ) y Φ es la CDF de la distribución N$(0, 1)$.

Observación:

Evidentemente, salvo la aleatoriedad en $\widehat{\mu}_i$ y $\widehat{\phi}$, r_{Q_i} es exactamente normal.