Regresión Lineal Múltiple - Semana 07

Johnatan Cardona Jiménez
jcardonj@unal.edu.co
Profesor Asistente - Departamento de Estadística
Universidad Nacional de Colombia, Sede Medellín

Semestre 02-2024

Inferencias sobre la respuesta media y valores futuros

- Suponga que se desea estimar la respuesta media para los valores en las predictoras $X_1 = x_{01}, X_2 = x_{02}, \ldots, X_k = x_{0k}$.
- Sea Y_0 la respuesta desconocida en tal conjunto de valores. **Si se define el vector fila**: $\underline{\mathbf{x}}_0 = [\ 1 \quad x_{01} \quad x_{02} \quad \ldots \quad x_{0k}\ ]$, entonces se puede escribir $Y_0 = \underline{\mathbf{x}}_0\underline{\boldsymbol{\beta}} + \varepsilon_0$, **por lo tanto la respuesta media en tal punto es**:

$$\mu_{Y|\underline{\mathbf{x}}_0} = E\left[Y|\underline{\mathbf{x}}_0\right] = \underline{\mathbf{x}}_0\underline{\boldsymbol{\beta}} = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_k x_{0k}.$$

Este valor es estimado por la correspondiente respuesta o valor ajustado, $\widehat{Y}_0$, que puede escribirse como:

$$\widehat{Y}_0 = \underline{\mathbf{x}}_0\underline{\widehat{\boldsymbol{\beta}}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{01} + \widehat{\beta}_2 x_{02} + \cdots + \widehat{\beta}_k x_{0k}.$$

$\widehat{Y}_0$ **tiene las siguientes propiedades**:

- $E\left[\widehat{Y}_0\right] = E\left[\underline{\mathbf{x}}_0\widehat{\underline{\beta}}\right] = \underline{\mathbf{x}}_0\, E\left[\widehat{\underline{\beta}}\right] = \underline{\mathbf{x}}_0\underline{\beta} = E\left[Y|\underline{\mathbf{x}}_0\right]$, esto es, $\widehat{Y}_0$ es un estimador insesgado de la respuesta media $E\left[Y|\underline{\mathbf{x}}_0\right]$.
- $Var\left[\widehat{Y}_0\right] = Var\left[\underline{\mathbf{x}}_0\widehat{\underline{\beta}}\right] = \underline{\mathbf{x}}_0\, Var\left[\widehat{\underline{\beta}}\right]\underline{\mathbf{x}}_0' = \sigma^2\,\underline{\mathbf{x}}_0\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\underline{\mathbf{x}}_0'$, que es estimada por: $\widehat{Var}\left[\widehat{Y}_0\right] = \mathrm{MSE}\,\underline{\mathbf{x}}_0\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\underline{\mathbf{x}}_0'$.
- Bajo el supuesto de normalidad en los errores, $\widehat{Y}_0$ es una variable aleatoria normal, debido a que es una combinación lineal de los $\widehat{\beta}_j$'s que también son normales.

En resumen:

$$\widehat{Y}_0 \sim N\Big(E\left[Y|\underline{\mathbf{x}}_0\right]\ ,\ \sigma^2\,\underline{\mathbf{x}}_0\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\underline{\mathbf{x}}_0'\Big)$$

Luego, se tiene que el estadístico $T = \dfrac{\widehat{Y}_0 - E[Y|\underline{\mathbf{x}}_0]}{\text{ee}(\widehat{Y}_0)} \sim t_{n-p}$, con $\text{ee}(\widehat{Y}_0) = \sqrt{\widehat{Var}[\widehat{Y}_0]}$, lo cual permite demostrar lo siguiente:

- Para la respuesta media $E[Y|\underline{\mathbf{x}}_0]$ en un vector apropiado $\underline{\mathbf{x}}_0$.

Pruebas de hipótesis sobre la respuesta media para un nivel de significancia α

Juego de hipótesis	Estadístico de prueba	Criterio de rechazo
$H_0 : \mu_{Y\|\underline{\mathbf{x}}_0} = c$ $H_1 : \mu_{Y\|\underline{\mathbf{x}}_0} \neq c$ con $c \in \mathbb{R}$	$T_0 = \dfrac{\widehat{Y}_0 - c}{\text{ee}\left(\widehat{Y}_0\right)} \overset{\text{bajo } H_0}{\sim} t_{n-p}$	Se rechaza H_0 si $\lvert T_0 \rvert > t_{1-\alpha/2,n-p}$

donde $t_{1-\alpha/2,n-p}$ es el percentil $1-\alpha/2$ de la distribución t-Student con $n-p$ grados de libertad.

IC del $(1-\alpha)100\%$ para la respuesta media $E[Y|\underline{\mathbf{x}}_0]$:

Basados de nuevo en que el estadístico:

$$T = \frac{\widehat{Y}_0 - E[Y|\underline{\mathbf{x}}_0]}{\text{ee}(\widehat{Y}_0)} \sim t_{n-p}$$

lo cual implica que:

$$P\left(-t_{\alpha/2,n-p} < \frac{\widehat{Y}_0 - E[Y|\underline{\mathbf{x}}_0]}{\text{ee}(\widehat{Y}_0)} < t_{\alpha/2,n-p}\right) = 1-\alpha$$

De donde se obtiene que un IC del $(1-\alpha)100\%$ para la respuesta media: $\mu_{Y|\underline{\mathbf{x}}_0} = E[Y|\underline{\mathbf{x}}_0]$ es:

$$\widehat{y}_0 \pm t_{\alpha/2,n-p}\,\text{ee}(\widehat{Y}_0).$$

- **Considere ahora el problema de predecir un valor futuro Y_0 (no observado en la muestra) de la variable respuesta, en $X_1 = x_{01}, X_2 = x_{02}, \ldots, X_k = x_{0k}$.**
- Claramente, usando el modelo ajustado, predecimos de manera puntual tal valor por $\widehat{Y}_0$, pero sabemos que no es un estimador insesgado de Y_0, por lo que siempre se genera un error de predicción dado por: $Y_0 - \widehat{Y}_0$.
- Note que el error de predicción tiene media cero y dado que el valor futuro y su pronóstico son independientes, entonces la varianza del error de predicción $\widehat{Y}_0 - Y_0$ está dada por:

$$Var\left[Y_0 - \widehat{Y}_0\right] = Var\left[Y_0\right] + Var\left[\widehat{Y}_0\right] = \sigma^2\left[1 + \underline{\mathbf{x}}_0\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\underline{\mathbf{x}}_0\right],$$

que es estimada por: $\widehat{Var}\left[\widehat{Y}_0 - Y_0\right] = \text{MSE}\left[1 + \underline{\mathbf{x}}_0\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\underline{\mathbf{x}}_0\right]$.

Con esto podemos hallar los siguientes resultados:

- Para un valor futuro Y_0 en un vector apropiado $\underline{\mathbf{x}}_0$:

IP del $(1-\alpha)100\%$ para un valor futuro Y_0:

Basados en este caso en que el estadístico:

$$T = \frac{Y_0 - \widehat{Y}_0}{\text{ee}(Y_0 - \widehat{Y}_0)} \sim t_{n-p},$$

con $\text{ee}(Y_0 - \widehat{Y}_0) = \sqrt{\widehat{Var}[Y_0 - \widehat{Y}_0]}$, lo cual implica que:

$$P\left(-t_{\alpha/2,n-p} < \frac{Y_0 - \widehat{Y}_0}{\text{ee}(Y_0 - \widehat{Y}_0)} < t_{\alpha/2,n-p}\right) = 1-\alpha$$

De donde se obtiene que un IP del $(1-\alpha)100\%$ para un valor futuro Y_0 es:

$$\widehat{Y}_0 \ \pm \ t_{\alpha/2,n-p}\,\text{ee}(Y_0 - \widehat{Y}_0)$$

Notas:

- Los intervalos de predicción estiman los posibles valores para un valor particular de la variable respuesta (no para su media) en un vector dado $\underline{\mathbf{x}}_0$.
- Asumimos que este valor particular es un valor futuro de la variable aleatoria Y, y por tanto, no fue utilizado en la regresión.
- Si Y_0 es un valor futuro y $\widehat{Y}_0 = \underline{\mathbf{x}}_0 \underline{\widehat{\beta}}$ es su estimador, entonces estas dos variables aleatorias son estadísticamente independientes, dado que Y_0 no fue utilizado para hallar los parámetros estimados, de ahí el estadístico y los límites del intervalo de predicción.

Precaución:

Deben evitarse las extrapolaciones por fuera del rango de experimentación en el espacio de las predictoras, para lo cual no basta con evaluar si cada valor componente del vector $\underline{\mathbf{x}}_0$ se encuentra dentro del rango usado (u observado) para la correspondiente predictora, **sino que es necesario evaluar si $\underline{\mathbf{x}}_0$ pertenece a la región de observación conjunta**.

Para ello basta con verificar si:

$$h_{00} = \boldsymbol{x_0} \left(\boldsymbol{X'X}\right)^{-1} \boldsymbol{x_0'} < \max_{1 \leq i \leq n} \{h_{ii}\}$$

con h_{ii} el i-ésimo elemento de la matriz 'hat' $\underset{n \times n}{\boldsymbol{H}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Validación de los supuestos del modelo de RLM

Para la validación de supuestos se usan generalmente los residuales del modelo, los cuales sabemos que se definen así:

$$e_i = y_i - \hat{y}_i, \ \ i = 1, \ldots, n$$

Observe que, la magnitud de los residuales e_i depende de la escala de medida de la respuesta Y, lo cual no permite determinar cuando un residual es 'grande'. Para resolver este problema en lugar de usar los residuales crudos definidos arriba, se recomienda utilizar residuales escalados que transforman a los anteriores para tener media cero y varianza unitaria.

Residuales escalados

Se han definido varias versiones de residuales escalados, entre los que se destacan:

- **Residuales estandarizados:** para su definición se considera el supuesto sobre los errores, que establece que ε_i se distribuye con media cero y varianza σ^2. Por tanto, los residuales estandarizados, denotados d_i se definen como:

$$d_i = \frac{e_i}{\sqrt{\text{MSE}}}, \ i = 1, \ldots, n$$

Si el supuesto es adecuado los valores de d_i deben oscilar entre -3 y 3. Por tanto, Un d_i grande ($|d_i| > 3$) es indicio de una observación potencialmente atípica.

- **Residuales estudentizados:** para su definición se considera el hecho de que realmente los residuales e_i en general no son independientes ni tienen varianza constante como los errores ε_i. Veamos,
  Sabemos que, $\underline{\boldsymbol{e}} = (\boldsymbol{I} - \boldsymbol{H})\underline{\boldsymbol{Y}}$, donde $\boldsymbol{I} - \boldsymbol{H}$ es una matriz simétrica e idempotente. Luego,

$$\begin{aligned} E\left[\underline{\boldsymbol{e}}\right] &= E\left[(\boldsymbol{I} - \boldsymbol{H})\underline{\boldsymbol{Y}}\right] = (\boldsymbol{I} - \boldsymbol{H})E\left[\underline{\boldsymbol{Y}}\right] = (\boldsymbol{I} - \boldsymbol{H})\mathbf{X}\underline{\beta} \\ &= \mathbf{X}\underline{\beta} - \boldsymbol{H}\mathbf{X}\underline{\beta} = \mathbf{X}\underline{\beta} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\ \boldsymbol{X}'\mathbf{X}\underline{\beta} = \mathbf{0} \end{aligned}$$

$$Var\left[\underline{\boldsymbol{e}}\right] = Var\left[(\boldsymbol{I} - \boldsymbol{H})\underline{\boldsymbol{Y}}\right] = (\boldsymbol{I} - \boldsymbol{H})Var\left[\underline{\boldsymbol{Y}}\right](\boldsymbol{I} - \boldsymbol{H})' = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

  De donde: $V(e_i) = \sigma^2(1 - h_{ii})$, y $\text{cov}(e_i, e_j) = -\sigma^2 h_{ij}$ para $i \neq j$.

Por tanto, mientras que los errores ε_i tienen varianza constante σ^2 y son incorrelacionados, los residuales no necesariamente tienen la misma varianza y pueden ser correlacionados.

De esta forma, los residuales estudentizados, denotados r_i, se definen como:

$$r_i = \frac{e_i}{\sqrt{\text{MSE}(1 - h_{ii})}}, \ i = 1, \ldots, n.$$

Este residual, también debe ubicarse entre -3 y 3. Se considera atípica aquella observación con un r_i grande ($|r_i| > 3$).

NOTAS:

- Si el modelo de RLM especificado es correcto los r_i tienen varianza aproximadamente constante!! igual a 1.
- En conjuntos grandes de datos la varianza de los r_i se puede estabilizar en 1 y así no habrá mucha diferencia entre éstos y los d_i.
- Si todos los supuestos del modelo se cumplen, se espera que aproximadamente el 68% de los residuales d_i ó r_i, estén entre -1 y $+1$, aproximadamente el 95% entre -2 y $+2$ y aproximadamente 99.7% entre -3 y $+3$.

La validación de los supuestos vista en regresión lineal simple se mantiene, solo que ahora se recomienda utilizar residuales escalados (d_i ó preferiblemente r_i) en lugar de utilizar los residuales crudos e_i.

Validación de los supuestos en los errores

Recuerde que en los modelos de regresión se han impuesto las siguientes cuatro condiciones sobre el término de error:

- Los errores son variables aleatorias normales.
- Los errores tienen media cero.
- Los errores tienen varianza constante.
- Los errores son mutuamente independientes.

Recuerde que en esta asignatura asumiremos el supuesto de independencia de los errores y en virtud de que usando los residuales del modelo el supuesto de media cero siempre se cumple, entonces se define lo siguiente:

- El supuesto de normalidad puede chequearse bien sea con el gráfico de probabilidad normal de los residuales o con alguna de las pruebas estadísticas de normalidad, entre las cuales se tienen las de Shapiro Wilk, Kolmogorov Smirnov, Cramér von Mises y Anderson Darling.
- Para chequear el supuesto de varianza constante, resulta útil un gráfico de residuales versus valores ajustados de la respuesta.

Medidas Remediales

Las medidas remediales descritas en el caso de RLS también son aplicables en RLM. Con el fin de superar las deficiencias del modelo se pueden realizar transformaciones sobre la variable respuesta y/o sobre las variables predictoras.

Las transformaciones sobre la respuesta pueden ayudar en el caso de que los errores no resulten normales o la varianza no sea constante. Transformaciones sobre las variables predictoras resultan útiles cuando la superficie de respuesta es curvilínea.

Si las desviaciones respecto al supuesto de normalidad son severas, y ninguna transformación resulta útil y/o interpretable, existe otra alternativa, los llamados **modelos lineales generalizados** con los cuales se pueden modelar respuestas que no se distribuyen normales; sin embargo, tales modelos están más allá del alcance de este curso.

Identificación de observaciones extremas en el modelo de RLM

Además de la validación de supuestos en los errores de un modelo de RLM, se debe chequear la presencia de observaciones extremas, tales como:

- Observaciones atípicas (outliers)
- Puntos de balanceo
- Observaciones influenciales

Observaciones atípicas

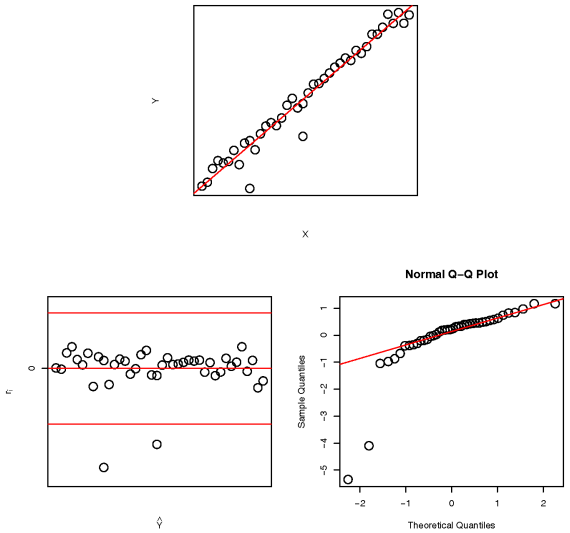
Una observación **atípica** (o **outlier**) es aquella que está separada (en su valor de la respuesta Y) del resto de las observaciones y por tanto puede afectar los resultados del ajuste del modelo de regresión.

Interesa identificarlas para luego, si es posible analizar si se tratan de observaciones malas (por errores de registro o medición) que pueden ser descartadas, o si realmente son datos correctos pero extraños que no deben ser eliminados del conjunto de datos.

Para detectar observaciones atípicas se usan los **residuales escalados** definidos anteriormente. Se considera que una observación es **atípica** cuando su residual estudentizado r_i, es tal que: $|r_i| > 3$.

Muchos **outliers** en los datos pueden causar niveles de confianza reales menores de lo esperado.

La siguiente Figura ilustra el caso de **dos observaciones atípicas**.



Puntos de balanceo

Un **punto de balanceo** es una observación en el espacio de las predictoras, alejada del resto de la muestra y que puede controlar ciertas propiedades del modelo ajustado.

Este tipo de observaciones posiblemente no afecte los coeficientes de regresión estimados pero sí las estadísticas de resumen como el R^2 y los errores estándar de los coeficientes estimados.

Los puntos de balanceo son detectados mediante el análisis de los elementos de la diagonal principal de la matriz $\boldsymbol{H}$, los h_{ii}, que proporcionan una medida estandarizada de la distancia de la i-ésima observación al centro del espacio definido por las predictoras.

Se tiene lo siguiente:

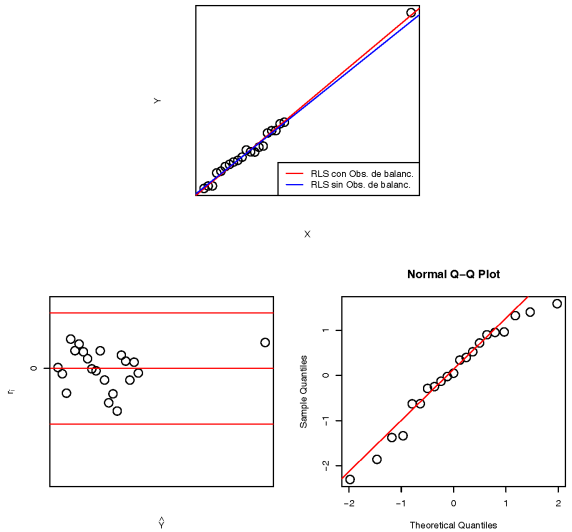
- La media de los h_{ii} es:

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{\text{traza}\,(\boldsymbol{H})}{n} = \frac{p}{n}$$

  con p el número de parámetros del modelo de RLM.

- Se asume que la observación i es un **punto de balanceo** si $h_{ii} > 2p/n$, pero si $2p/n > 1$ este criterio no funciona pues los h_{ii} siempre son menores que 1.

La próxima Figura ilustra el caso de **una observación de balanceo**.



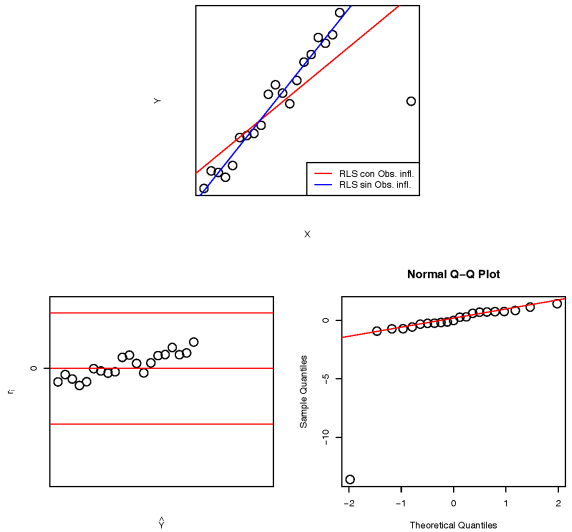
Observaciones influenciales

Una observación es **influencial** si tiene un impacto notable sobre los coeficientes de regresión ajustados, esto es, una observación influencial se dice que hala al modelo en su dirección, es decir, una observación es influencial si su exclusión del modelo causa cambios importantes en la ecuación de regresión ajustada.

Estas observaciones se caracterizan por tener un valor moderadamente inusual tanto en el espacio de las predictoras como en la respuesta.

Después de identificar las observaciones que están alejadas con respecto a los valores de Y (atípicas) y/o con respecto a sus valores en X (puntos de balanceo) evaluamos si éstas son influenciales.

La Figura siguiente ilustra el caso de **una observación influyente**.



Para la evaluación se cuenta con las siguientes medidas:

- Distancia de Cook.
- Diagnóstico DFFITS.
- Diagnóstico DFBETAS.

A continuación se presentan los diagnósticos para detectar observaciones influenciales.

1. **Distancia de Cook:** es una medida de la distancia cuadrática entre, el estimador de $\underline{\beta}$ por mínimos cuadrados basado en las n observaciones, y el estimador de $\underline{\beta}$ obtenido eliminando la i-ésima observación, así:

$$D_i = \frac{\left(\underline{\widehat{\beta}}_{(i)} - \underline{\widehat{\beta}}\right)' \boldsymbol{X}'\boldsymbol{X} \left(\underline{\widehat{\beta}}_{(i)} - \underline{\widehat{\beta}}\right)}{p\,\mathrm{MSE}} = \frac{r_i^2}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right), \quad i = 1, \ldots, n$$

donde, $\underline{\widehat{\beta}}_{(i)}$ es el vector de parámetros estimados obtenido cuando no se considera en el ajuste del modelo a la observación i.
Note que si D_i es alto, entonces la observación i tiene influencia sobre el vector de parámetros estimados $\widehat{\beta}$. Como $f_{0.5,\ p,n-p} \approx 1$ **se dice que la observación i será influencial si $D_i > 1$**

2. **Diagnóstico DFFITS:** es el número de desviaciones estándar que el valor ajustado $\widehat{y}_i$ se mueve si la observación i es omitida:

$$\text{DFFITS}_i = \frac{\widehat{Y}_i - \widehat{Y}_{(i)}}{\sqrt{\text{MSE}_{(i)}\, h_{ii}}} = \frac{e_i}{\sqrt{MSE_{(i)}\,(1-h_{ii})}} \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

donde, $\widehat{Y}_{(i)}$ es el i-ésimo valor ajustado obtenido cuando no se considera en el ajuste del modelo a la observación i y $\text{MSE}_{(i)}$ es el cuadrado medio del error obtenido cuando no se considera en el ajuste del modelo a la observación i.
Una observación será **influencial** si $|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}}$.

3. **Diagnóstico DFBETAS:** indica cuánto cambia el j-ésimo coeficiente de regresión estimado $\widehat{\beta}_j$ en unidades de desviación estándar, si se omite la i-ésima observación:

$$\text{DFBETAS}_{j(i)} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{\sqrt{\text{MSE}_{(i)}\, c_{jj}}}$$

donde c_{jj} es el j-ésimo elemento en la diagonal principal de la matriz: $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ y $\text{MSE}_{(i)}$ es el MSE de la regresión sin la observación i.
Una observación será **influencial** si $|\text{DFBETAS}_{j(i)}| > 2/\sqrt{n}$.

NOTA: Las observaciones que sean identificadas como puntos influenciales y/o de balanceo deben ser investigadas. Si las observaciones identificadas según los criterios anteriores son errores de registro en la base de datos, estas deben ser removidas del análisis. Caso contrario, se debe emplear métodos más robustos para mantener dichos registros en el conjunto de datos: **métodos estadísticos robustos que no se vean afectados por observaciones atípicas, de balanceo o influenciales.**