Modelos no paramétricos y de regresión

Validación de supuestos

Javier Santibáñez

Facultad de Ciencias, UNAM

jsantibanez@sigma.iimas.unam.mx

Semestre 2019-1

Contenido

Supuestos de los modelos RL

Los supuestos de los modelos de regresión lineal son

1. Linealidad: $E(Y \mid \mathbf{X}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.
2. Homocedasticidad: $V(Y \mid \mathbf{X}) = \sigma^2$.
3. Observaciones no correlacionadas: $Cov(Y_j, Y_k \mid \mathbf{X}_i, \mathbf{X}_j) = 0$.
4. Independencia lineal de las variables explicativas: $rango(\mathbf{X}) = p + 1$.
5. Normalidad: $\mathbf{Y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

La validez de las inferencias realizadas depende de que los supuestos se cumplan, lo cual nunca ocurre, puesto que los modelos serán siempre aproximaciones a la realidad. Sin embargo, es posible tener una idea de qué tan *buena* es la aproximación realizada.

Tipos de residuos

- Algunas técnicas para detectar desviaciones a los supuestos del modelo RL se basan en los residuos, ya que estos pueden ser considerados como *estimadores* de los errores verdaderos.
- Por definición $\hat{\boldsymbol{\epsilon}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$ y, por el supuesto de normalidad, se sigue que $\hat{\boldsymbol{\epsilon}} \sim N_n\left(\mathbf{0}_n, \sigma^2(\mathbf{I}_n - \mathbf{H})\right)$. Mientras que $\boldsymbol{\epsilon} \sim N_n\left(\mathbf{0}_n, \sigma^2\mathbf{I}_n\right)$.
- En general, la matriz $\mathbf{I}_n - \mathbf{H}$ no es diagonal ni todos los elementos de su diagonal principal son iguales, por lo que los residuos $\hat{\epsilon}_i$, pueden tener varianzas distintas y estar correlacionados.
- Para eliminar el efecto de la matriz $\mathbf{I}_n - \mathbf{H}$ en la varianza de los residuos, se definen dos transformaciones de éstos, que sí tienen varianza constante.

Residuos estandarizados

- Los residuos estandarizados se denorarán por $R_1, \ldots, R_n$ y se definen como

$$R_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{MCO}\sqrt{1 - h_{ii}}}, \qquad i = 1, \ldots, n,$$

  donde h_{ii} es el i-ésimo elemento de la diagonal principal de $\mathbf{H}$.

- Los residuos estandarizados tienen media cero y varianza unitaria, sin embargo no son independientes.

- En `R`, los residuos estandarizados se calculan con la función `stdres` del paquete `MASS`.

Residuos studentizados

Los residuos studentizados se denotarán por $T_1, \ldots, T_n$ y se definen como

$$T_i = \frac{Y_i - \hat{Y}_i^{(i)}}{\hat{\sigma}_{MCO}^{(i)} \left(1 + \mathbf{x}_i' \left(\mathbf{X}_{(i)}' \mathbf{X}_{(i)}\right)^{-1} \mathbf{x}_i\right)^{1/2}}, \qquad i = 1, \ldots, n.$$

donde:

- $\hat{Y}_i^{(i)}$ es el valor predicho de Y_i, del modelo ajustado sin la i-ésima observación de la muestra,
- $\hat{\sigma}_{MCO}^{2(i)}$ es el estimador insesgado de σ^2, del modelo ajustado sin la i-ésima observación en la muestra,
- $\mathbf{X}_{(i)}$ es la matriz de diseño original sin la i-ésima fila, y
- $\mathbf{x}_i$ es el vector de mediciones de las variables auxiliares de la i-ésima observación en la muestra.

Residuos studentizados

- Los residuos studentizados también tienen media cero y varianza unitaria, con la diferencia que sí son independientes.
- El nombre de estos residuos se debe a que tienen distribución t_{n-p-2}.
- En `R`, los residuos studentizados se calculan con la función `studres` del paquete `MASS`.
- Los T_i también se conocen como residuos de validación cruzada o de *jackknife*, debido a que se calculan omitiendo la observación correspondiente.
- Existe una relación entre los T_i y los R_i que permite calcular los primeros sin tener que ajustar el modelo n veces:

$$T_i = R_i \left(\frac{n-p-2}{n-p-1-R_i^2} \right)^{1/2}$$

Validación de los supuestos de los modelos RL

El esquema a seguir para presentar los resultados de cada supuesto se puede plantear como respuesta a las siguientes preguntas:

1. ¿Por qué es importante el supuesto? ¿Qué problemas causa que el supuesto no se cumpla?
2. ¿Cómo detectar las desviaciones al supuesto? Gráficamente o con una prueba.
3. ¿Qué medidas se pueden aplicar para corregir las desviaciones?

Nota importante

Los procedimientos para detectar desviaciones a los supuestos y las medidas para corregirlas, pueden depender de la validez del resto de los supuestos, por lo que detectar desviaciones es mucho más complicado cuando falla más de un supuesto.

Linealidad

- El supuesto de linealidad es el más importante, puesto que es el que determina el tipo de relación que existe entre la variables respuesta Y y las variables auxiliares $X_1, \ldots, X_p$.
- El término lineal en los modelos de regresión lineal se debe a que la función que describe la media de Y es lineal en los parámetros $\beta_0, \beta_1, \ldots, \beta_p$ y no a linealidad en las variables $Y, X_1, \ldots, X_n$.
- En el caso más general, se puede plantear el primer supuesto como

$$E\left(g(Y) \,|\, \mathbf{X} = \mathbf{x}\right) = \beta_0 + \beta_1 z_1 + \ldots + \beta_q z_q + \epsilon$$

donde $z_k = h_k(x_1, \ldots, x_p)$, $k = 1, \ldots, p$, g y $h_1, \ldots, h_k$ son funciones conocidas.

Linealidad

- En consecuencia, el objetivo es describir el valor esperado de alguna transformación de Y (g), como combinación lineal de los parámetros β_k y de un conjunto de características del los individuos $(Z_1, \ldots, Z_q)$, derivadas de las mediciones de las variables auxiliares (a partir de las funciones $h_1, \ldots, h_q$).
- En el planteamiento que se usó en el curso se consideró: $g(x) = x$, $h_k(\mathbf{x}) = x_k$, para $k = 1, \ldots, p$. Por lo que las desviaciones al supuesto de normalidad se pueden considerar como una mala especificación de la función g o de las funciones $h_1, \ldots, h_k$.
- Si el supuesto de linealidad no se cumple, entonces es incorrecta la descripción de la distribución de la variable respuesta Y y en consecuencia, las conclusiones que se puedan extraer a partir del modelo son incorrectas.

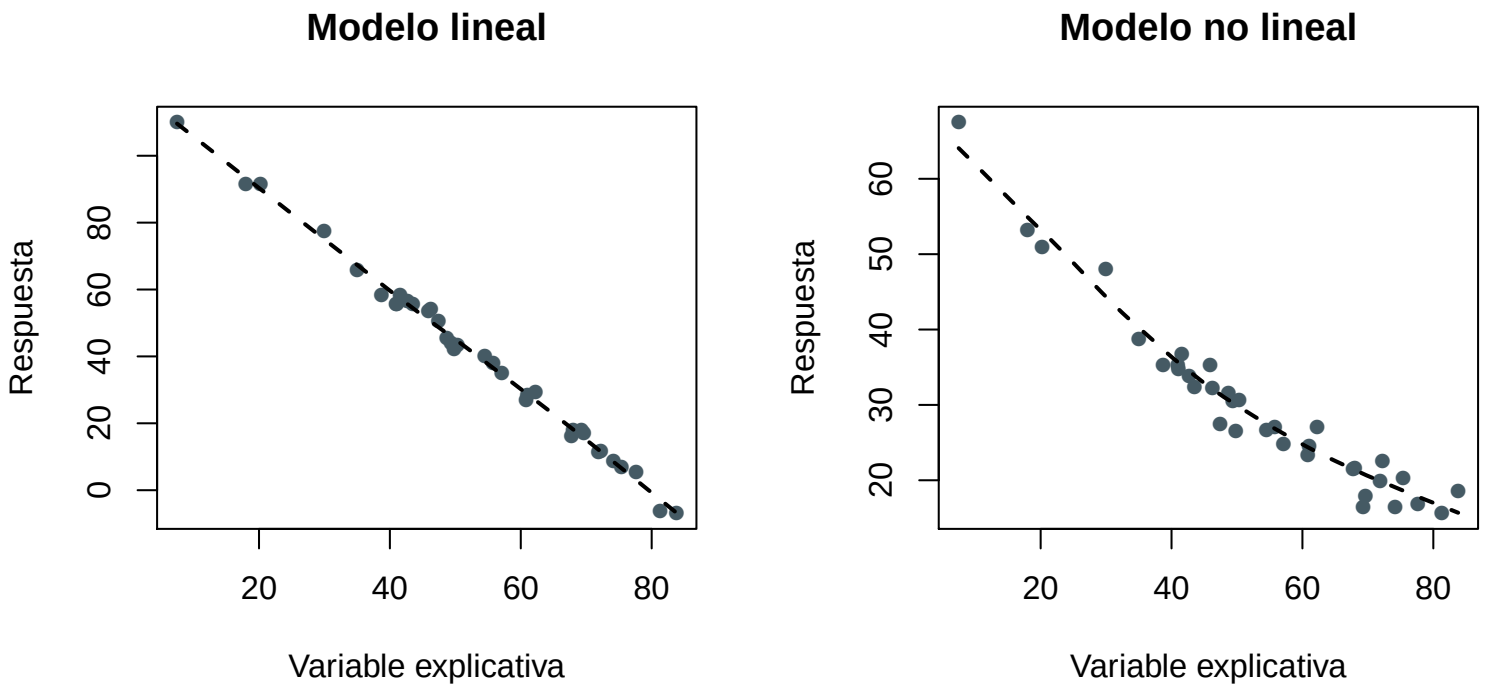
Métodos gráficos para detectar no linealidad

- Gráficas de dispersión de Y vs. X_k.
- Gráficas de dispersión de $\hat{\epsilon}$ vs. X_k.
- Gráficas de dispersión de $\hat{\epsilon}^{(k)}$ vs. X_k, donde $\hat{\epsilon}^{(k)}$ son llamados residuos parciales y son los residuos de un modelo ajustado con todas las variables explicativas excepto X_k.
- Gráficas de regresión parcial de X_k. Estas son gráficas de dispersión de $\hat{\epsilon}^{(k)}$ contra los residuos de regresar X_k sobre el resto de las variables explicativas.
- Gráficas de dispersión de $\hat{\epsilon}$ vs. $\hat{Y}$.

En cualquiera de los casos, el objetivo es identificar remanentes de una relación no lineal en los residuos del modelo. Para eliminar efectos de la escala se recomienda utilizar los residuos estandarizados.

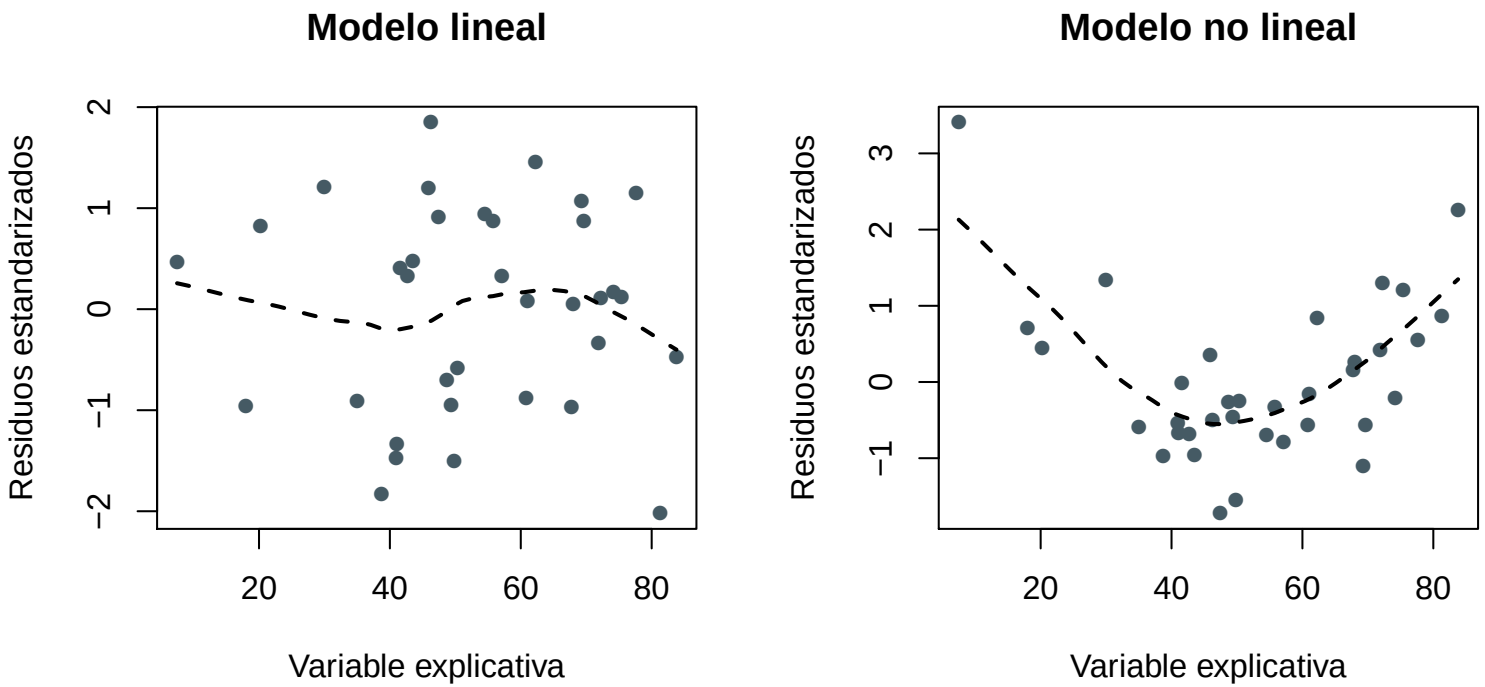
Linealidad: gráfica de dispersión de Y vs. X_k

El primer supuesto a verificar es el de la linealidad del modelo. En el caso simple, basta un gráfico de dispersión de la variable respuesta contra la variable explicativas.



Linealidad: gráfica de dispersión de R_i vs. X_k

También se pueden utilizar un gráfico de dispersión de los residuos estandarizados modelo contra las variables explicativas. Si la especificación del modelo es la correcta, no deberían haber patrones en la gráfica.



Linealidad: gráfica de dispersión de $R_i^{(k)}$ vs. X_k

Linealidad: gráfiaca de regresión parcial de $R_i^{(k)}$ vs. $X_k^{(k)}$

Validación del supuesto de linealidad

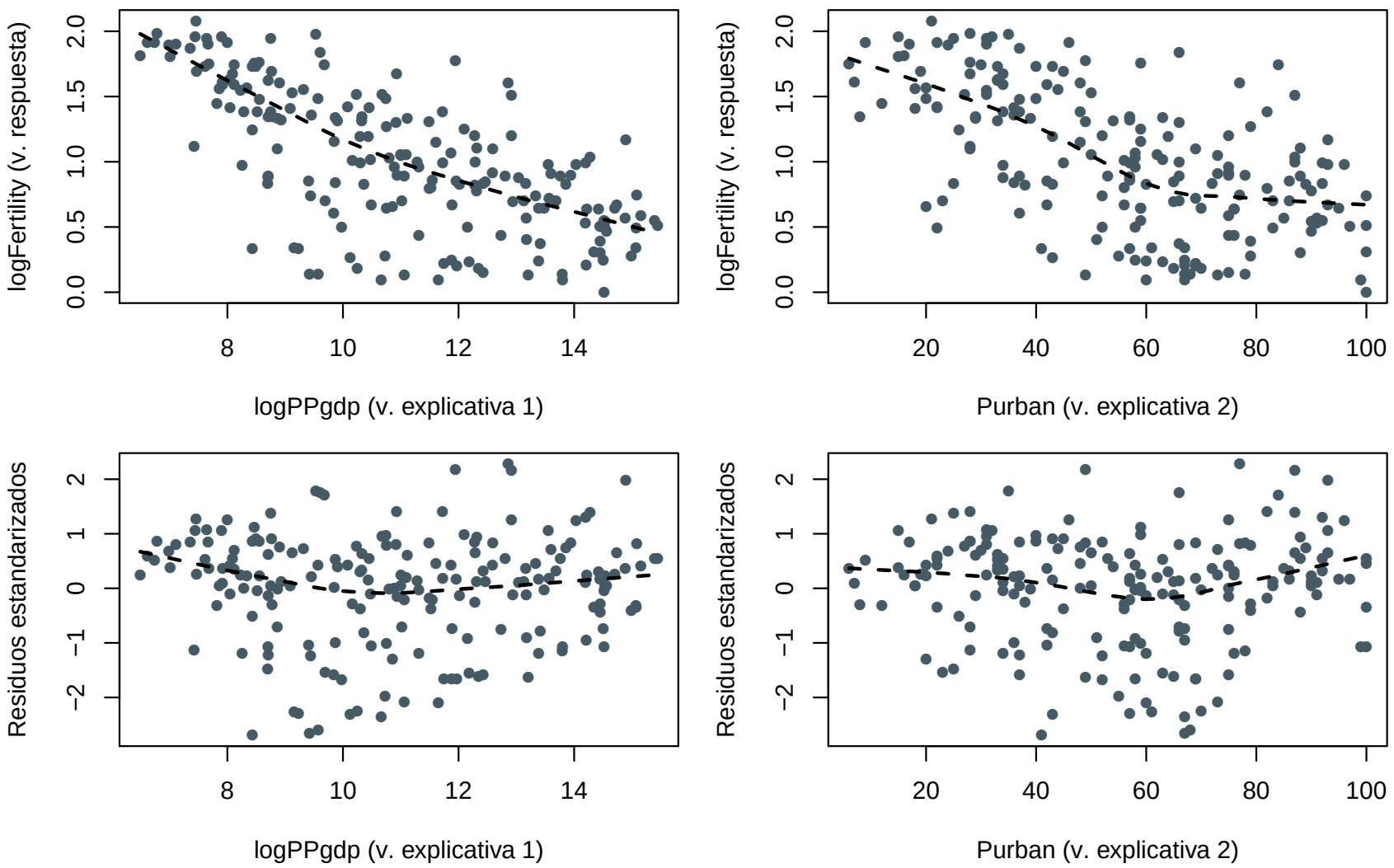
En el caso múltiple, se complica verificar la linealidad dado que se requiere graficar en dimensiones mayores y que puede haber asociaciones entre las variables explicativas.

Una solución simple es verificar las gráficas anteriores de respuesta y residuos contra cada X.

Las gráficas de respuesta vs. cada variable explicativa nos permiten explorar la relación marginal de la respuesta con cada variable explicativa, por lo que nos pueden llevar a concluir erróneamente.

Las gráficas de residuos vs. cada variable explicativa no prensentan el problema anterior, por lo que se pueden utilizar para explicar desviaciones al supuesto de linealidad.

Consideremos los datos del ejemplo de fecundidad de Naciones Unidas. Se tiene interés en ajustar el modelo


logFertility (v. respuesta)
0.0
0.5
1.0
1.5
2.0
8
10
12
14
logPPgdp (v. explicativa 1)
logFertility (v. respuesta)
0.0
0.5
1.0
1.5
2.0
20
40
60
80
100
Purban (v. explicativa 2)
Residuos estandarizados
−2
−1
0
1
2
8
10
12
14
logPPgdp (v. explicativa 1)
Residuos estandarizados
−2
−1
0
1
2
20
40
60
80
100
Purban (v. explicativa 2)

Validación del supuesto de linealidad

Una mejorar en las gráficas de residuos es considerar los residuos parciales. Supongamos que deseamos ajustar el modelo

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

Para verificar si la relación entre la respuesta Y y la variable X_k es lineal, considerando el efecto del resto de las variables explicativas, se ajusta el modelo anterior eliminando a X_k. Los residuos de este modelo parcial se llaman residuos parciales y se denotan por $e_i^{(k)}$.

Ahora se grafican los e_i^k vs. X_k. La idea es que si la relación entre Y y X_k es lineal, también se aprecie una relación lineal entre los e_i^k y X_k.

Pruebas para detectar no linealidad

Para detectar desviaciones al supuesto podemos utilizar dos pruebas:

- Falta de ajuste (*lack-of-fit*), para deterctar no linealidad en las variables explicativas y,
- No aditividad de Tuckey, para detectar no adivididad en la respuesta.

Si consideramos el modelo

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

Para probar la falta de ajuste (no linealidad) en la variable X_k se ajusta el modelo

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \eta X_k^2 + \epsilon$$

Y se contrasta la hipótesis $H_0 : \eta = 0$. Lo anterior se hace individualmente para cada variable.

Aún cuando la no linealidad sea de otro tipo, exponencial por ejemplo, el caso más sencillo X_k^2 ajusta mejor que considerar únicamente X_k.

Ejemplo: falta de ajuste

Consideremos los datos de las primeras gráficas de esta unidad, que fueron generados como

$$Y = 80 \exp(-0.02X) + \epsilon$$

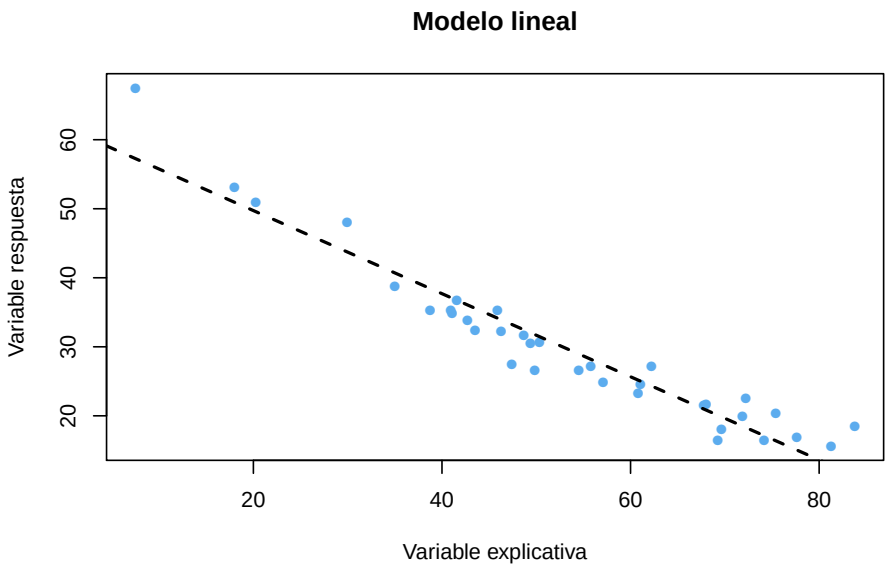
con $\epsilon \sim N(0, 4)$.

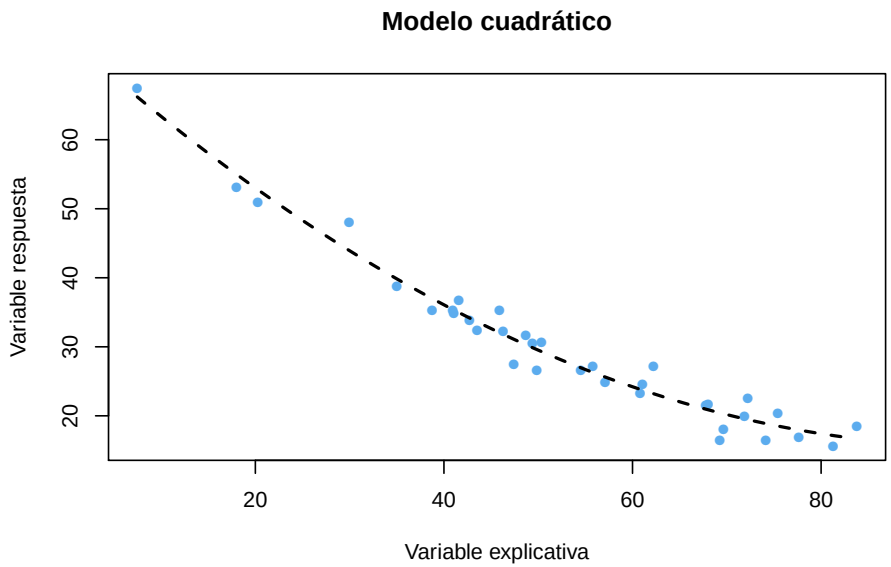
Aplicamos una prueba de falta de ajuste:

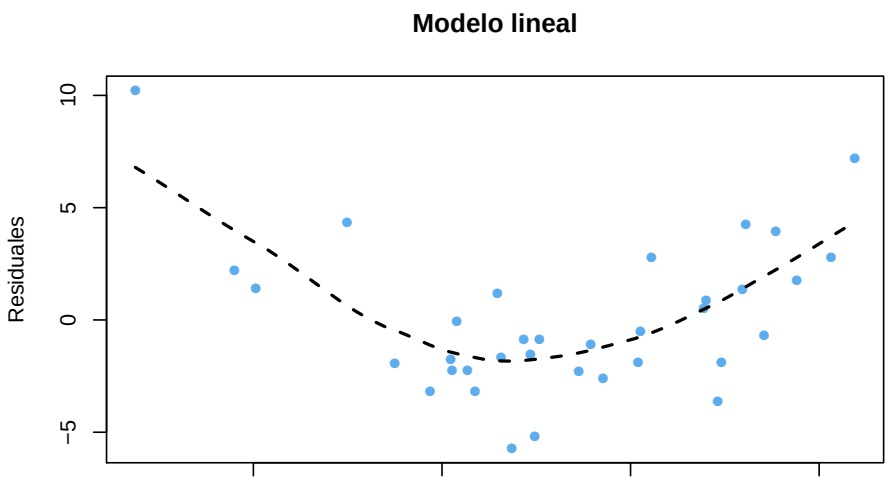
	Estimate	Std. Error	t value	Pr($>|t|$)
(Intercept)	75.0213	2.0754	36.15	0.0000
x	-1.2268	0.0857	-14.32	0.0000
x^2	0.0063	0.0008	7.48	0.0000

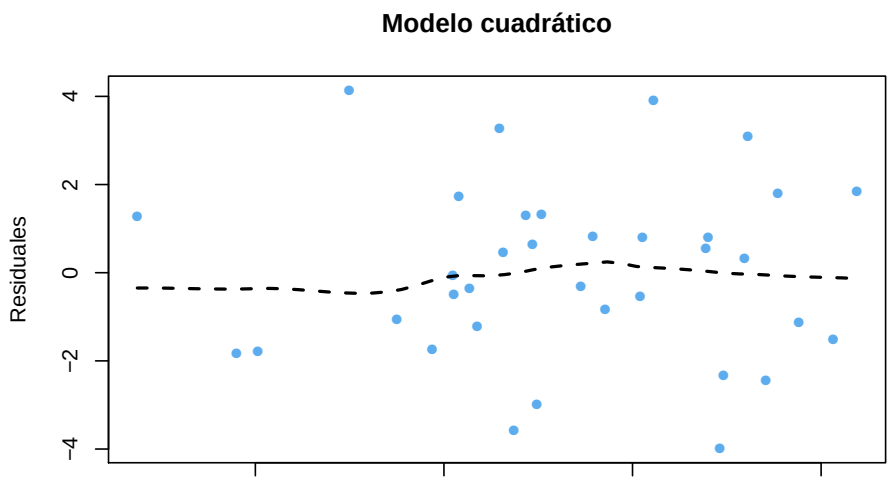
El resultado anterior indica que hay falta de linealidad en la v. explicativa x.

Ejemplo: falta de ajuste









Prueba de no aditividad de Tukey

En el caso múltiple se puede detectar no linealidad en el modelo ajustando el modelo

$$y = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_p X_p + \gamma Z + \epsilon$$

donde Z se calcula como $\hat{Y}^2/2\bar{Y}$ y $\hat{Y}$ corresponde al valor ajustado de Y bajo el modelo múltiple usual. La prueba de no aditividad de Tukey contrasta las hipótesis $H_0 : \gamma = 0$ vs. $H_1 : \gamma \neq 0$.

La idea intuitiva es que si la relación entre Y y las variables explicativas no es lineal, ésta es capturada por $\hat{Y}$.

Si se rechaza la hipótesis $H_0 : \gamma = 0$, se puede aplicar una transformación a los datos de la forma $Y^* = Y^{1-\hat{\gamma}}$, si $\hat{\gamma} \neq 1$ y $Y^* = \log Y$, si $\hat{\gamma} = 1$.

La desventaja de aplicar tal transformación es que se puede afectar el supuesto de varianza constante.

Correcciones a no linealidad

En general se pueden corregir desviaciones al supuesto de linealidad aplicando alguna de las siguientes acciones:

- Transformaciones en las variables explicativas. Generalmente se consideran transformaciones potenciales de la forma X^λ,

$$X_k^* = \begin{cases} X^\lambda, & \lambda \neq 0 \\ \log X, & \lambda = 0. \end{cases}$$

  es claro que algunas de las transformaciones solamente tienen sentido cuando $X_k > 0$. Generalmente basta con considerar $\lambda \in \{-1/2, -1, 0, 1/2\}$.
- Transformaciones en la variable respuesta. Se pueden utilizar transformaciones potenciales como en el caso de las variables explicativas. Se debe tener cuidado al aplicar transformaciones sobre Y ya que esto puede afectar la homocedasticidad.
- Transformaciones en ambas variables.
- Ajustar modelos polinomiales (más detalles en selección de variables).

Validación del supuesto de homocedasticidad

Las principales desviaciones al supuesto de homocedasticidad se pueden resumir como

$$V(Y|\mathbf{X}) = \sigma^2 g(\mathbf{X}, \boldsymbol{\gamma})$$

Es decir, la varianza de Y no es constante y (posiblemente) depende de $\mathbf{X}$.

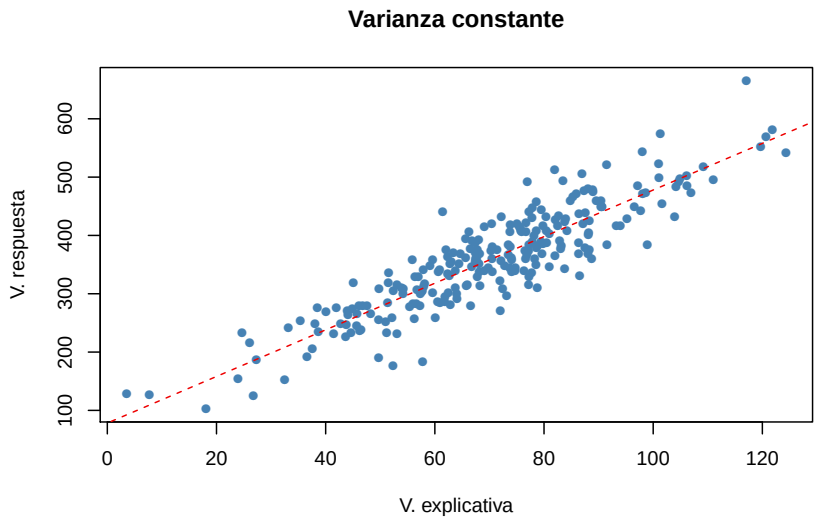
¿En qué casos no se cumple el supuesto de varianza constante?

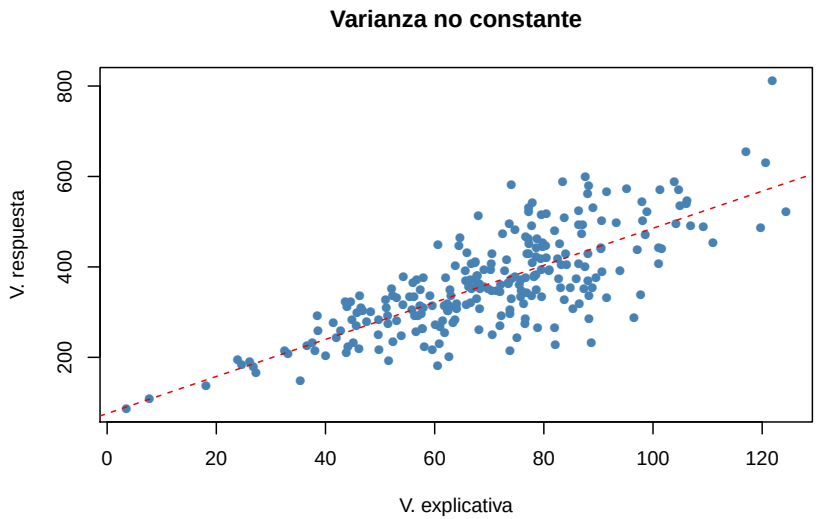
- Prácticamente en cualquier problema real.
- En la medición de magnitudes físicas, la precisión de las mediciones está relacionada con la magnitud que se desea medir.
- Cuando Y corresponde al total de m observaciones independientes con igual varianza σ^2, entonces $V(Y) = m\sigma^2$.
- Si Y es el promedio de m observaciones idenpendientes con igual varianza σ^2, entonces $V(Y) = \sigma^2/m$.

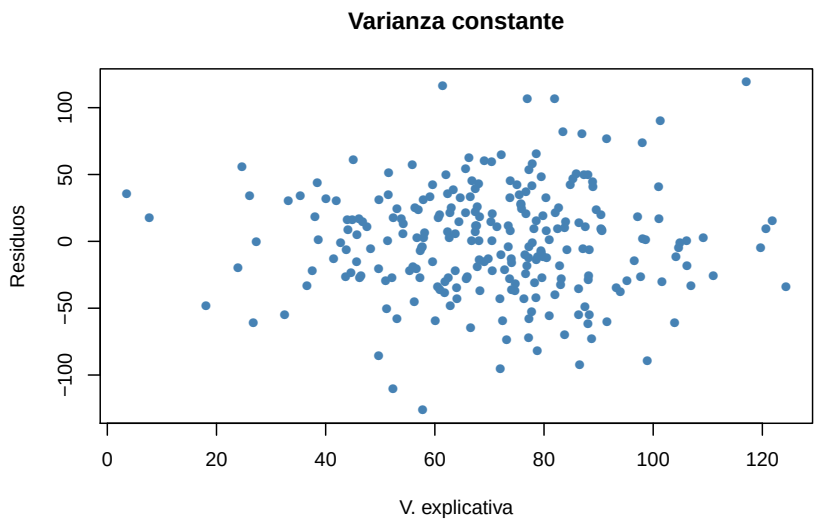
Verificación del supuesto de homocedasticidad

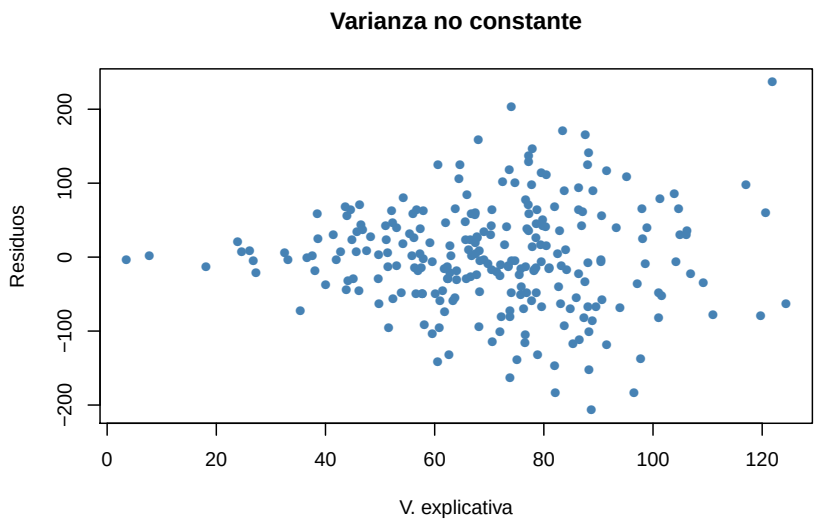
- Las desviaciones al supuesto de varianza constante se pueden detectar a partir de las gráficas de dispersión de la variable respuesta contra las explicativas o bien de los residuos contra las explicativas.
- Cuando no hay varianza constante, los residuos no se distribuyen uniformemente alrededor de la recta $y = 0$, sino que se observan patrones, el más común es el *megáfono*.
- Si la varianza no es constante, el estimador de $\boldsymbol{\beta}$ sigue siendo insesgado, el problema es realmente la estimacion de la varianza de $\hat{\boldsymbol{\beta}}$, que puede ser sobre estimada.
- Se pueden aplicar transformaciones a la variable respuesta para estabilizar la varianza, generalmente de la forma $\sqrt{Y}$, $\log Y$ o $1/Y$, aunque esto puede tener consecuencias en el supuesto de linealidad.
- Si no se toma alguna medida correctiva, se puede estimar $\boldsymbol{\beta}$ por MCO y utilizar *bootstap* para estimar su varianza. En general será mayor que la

Verificación del supuesto de homocedasticidad









Verificación del supuesto de homocedasticidad

- Algunas pruebas para detectar varianza sólo aplican cuando se tienen varias observaciones de Y para un mismo nivel de las X. En estudios experimentales esto se pueden incluir observaciones repetidas para cada nivel de las variables de diseño y así contrastar si hay varianza constante, sin embargo en estudios observacionales no se tiene control sobre las repeticiones.
- Otra alternativa es agrupar los datos para niveles similares de las variables de diseño, sin embargo, no hay una forma única de hacerlo y esto introduce subjetividad. Distintos agrupamientos pueden llevar a conclusiones diferentes.
- Una prueba sencilla para detectar varianza no constante consiste en ajustar el modelo

$$\hat{e}_i^{*2} = \gamma_0 + \gamma_1 x_{ki} + \eta.$$

donde $\hat{e}_i^*$ es el i-ésimo residuo estandarizado (dividido entre $\hat{\sigma}^2$).

Mínimos Cuadrados Ponderados

Una posible solución al problema de heterocedasticidad es ajustar el modelo por Mínimos Cuadrados Ponderados (MCP) o *Weighted Least Squares* (WLS). Si reemplazamos el supuesto de homocedasticidad por $V(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{W}$ donde $\mathbf{W}$ es una matriz diagonal de constantes conocidas, que pueden estar relacionadas con las variables explicativas. Entonces se define la suma de cuadrados de residuos ponderados como

$$Q_{\mathbf{W}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{w_i}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

donde w_i es el i-ésimo elemento de la diagonal de $\mathbf{W}$. El estimador de $\boldsymbol{\beta}$ que resulta de minimizar $Q_{\mathbf{W}}$ se conoce como estimador de MCP y se puede mostrar que

$$\hat{\boldsymbol{\beta}}_{\mathbf{W}} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y}.$$

y el estimador de la varianza es $\hat{\sigma}^2_{\mathbf{W}} = \frac{1}{n-p-1}Q_{\mathbf{W}}(\hat{\boldsymbol{\beta}}_{\mathbf{W}})$.

Verificación del supuesto de homocedasticidad

Cuando se estima por MCP, se calculan los residuos ponderados

$$\hat{\epsilon}_{\mathbf{W}} = \mathbf{W}^{-1/2}\left(\mathbf{I}_n - \right)\hat{e}_i\mathbf{W} = \frac{1}{\sqrt{w_i}}\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}\right).$$

El problema de ajustar por MCP es determinar los pesos adecuados. En el caso simple se puede considerar $w_i = g(x_i)$, para alguna función $g(\cdot)$. Por ejemplo, $w_i = x_i$ o $w_i = x_i^2$.

Existe una técnica de estimación en que se incluyen cómo parámetros del modelo los pesos. Se conoce como Mínimos Cuadrados Generalizados. Sin embargo va más allá del alcance de este curso.

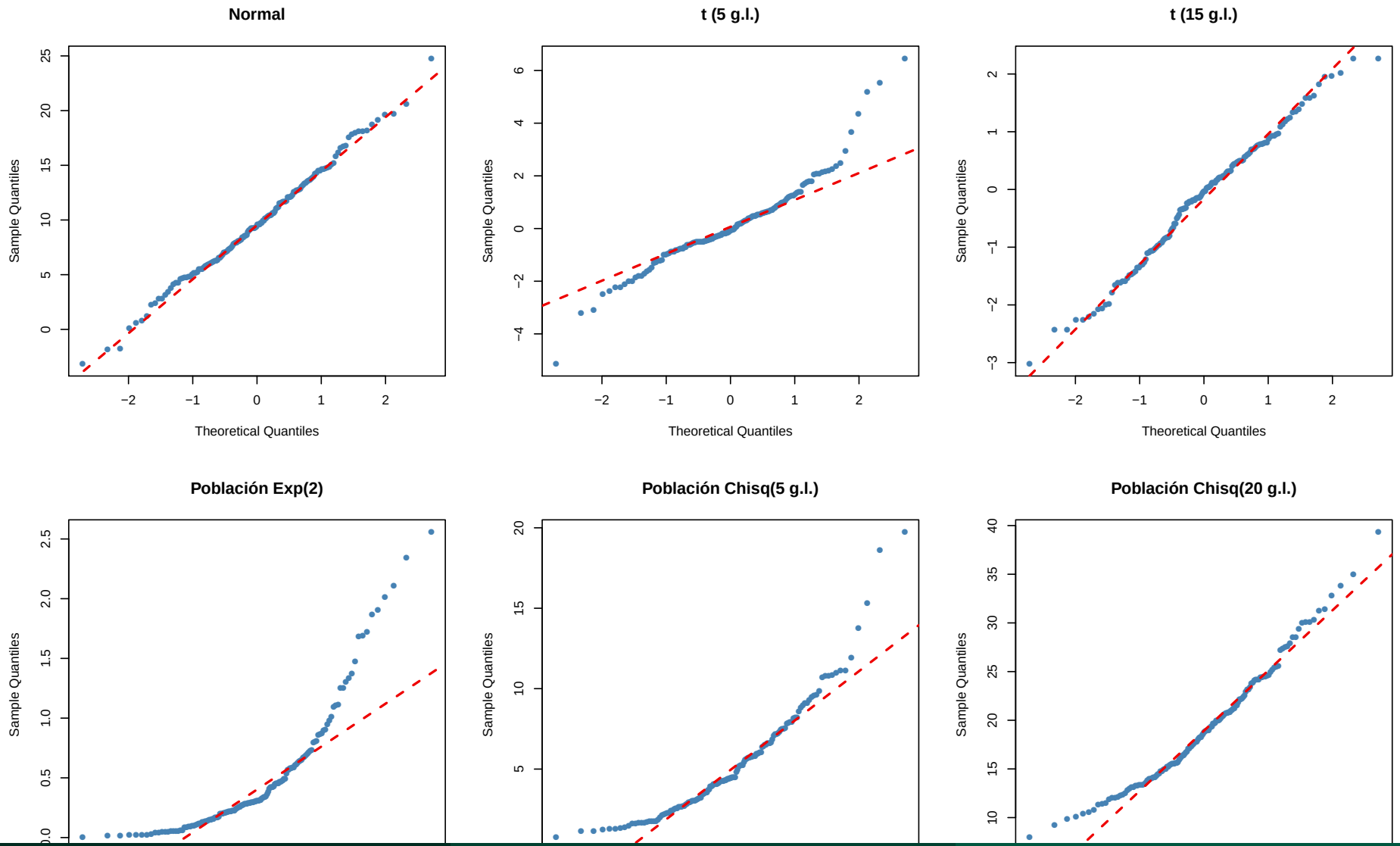
Verificación del supuesto de independencia

- Este supuesto es importante principalmente cuando los datos se obtuvieron con algún orden en el tiempo y las mediciones realizadas pueden estar influidas por mediciones previas.
- La presencia de errores correlacionados tiene efectos en la estimación de la varianza del modelo y de los estimadores. Si se detecta dependencia en los errores lo mejor es incluirla en el modelo. Por ejemplo, se puede considerar un proceso autorregresivo en los errores (AR).
- Si la estructura de dependencia de los errores es más complicada se puede utilizar otro método de estimación, como Mínimos Cuadrados Ponderados o Mínimos Cuadrados Generalizados. En el caso de los MCP, el inconveniente es proponer una matriz de pesos para ajustar el modelo.

Verificación del supuesto de normalidad

- Las desviaciones al supuesto de normalidad afectan la significancia de las pruebas y la confianza de los intervalos que hemos construido.
- Afortunadamente, cuando el tamaño de muestra es grande, los estimadores EMV se distribuyen aproximadamente normales. La *velocidad* de la convergencia es mayor cuando más *parecida* sea la distribución de los errores a la distribución normal.
- Dos formas gráficas de verificar normalidad: los histogramas y los gráficos cuantil-cuantil (*qq-plot*)de los residuos.
- Al realizar histogramas de los residuos, se espera encontrar un comportamiento normal con media 0. Si se observa asimetría o colas pesadas, puede haber problemas de normalidad.
- Al realizar la gráfica *qq-plot* se espera que los puntos caigan aproximadamente sobre una linea recta, principalmente en la parte central del gráfico. Si se observa asimetría o colas pesadas, puede haber problemas

Verificación del supuesto de normalidad



Verificación del supuesto de normalidad

- También se pueden aplicar pruebas de normalidad en los residuos. El paquete `nortest` contiene las pruebas más conocidas para normalidad: Anderson-Darling, Cramer-von Mises, Lilliefors (Kolmogorov-Smirnov), Pearson, Shapiro-Wilk. Aunque no es recomendable aplicar pruebas, puesto que formalmente, los residuos no son *normales*.
- Estas pruebas establecen como hipótesis nula la normalidad de las observaciones, de manera que *p-values* pequeños (menores que α) indican no normalidad.
- Cuando hay evidencia de no normalidad, podemos confiar en los resultados asintóticos o utilizar *bootstrap*. En el caso de los intervalos para las componentes de $\boldsymbol{\beta}$ es sencillo, para la prueba ANOVA no lo es tanto.

Multicolinealidad

- El supuesto de rango completo (por columnas) se traduce en que $\mathbf{X}'\mathbf{X}$ es invertible y así así hay una solución única a las ecuaciones normales.
- En la practica suele ocurrir que algunas columnas son *casi* combinaciones lineales de otras columnas. La teoría nos dice que al no haber una igualdad exacta, $\mathbf{X}'\mathbf{X}$ es invertible, sin embargo, puede haber problemas numéricos para calcular dicha inversa.
- Se puede detectar multicolinealidad por pares de variables a partir de la matriz de correlaciones y de los gráficos de dispersión por pares de variables. Aunque no hay criterios estadísticos para decidir si hay o no problemas de multicolinealidad.

Multicolinealidad: índice de condición κ

- Otra forma de detectar la multicolinealidad es con el *índice de condición* de la matriz de diseño $\mathbf{X}'\mathbf{X}$, el cual se define como:

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}},$$

  donde λ_{max} y λ_{min} a los valores propios máximo y mínimo de $\mathbf{X}'\mathbf{X}$, respectivamente.

- Generalmente, $\kappa < 100$ indica que no hay problemas de multicolinealidad, si κ está entre 100 y 1000 entonces hay multicolinearidad moderada y si $\kappa < 1000$, entonces hay multicolinealidad grave y podemos tener problemas numéricos.

- En `R` se puede calcular κ con la función `kappa`.

Multicolinearidad: Fractores de Inflación de la Varianza

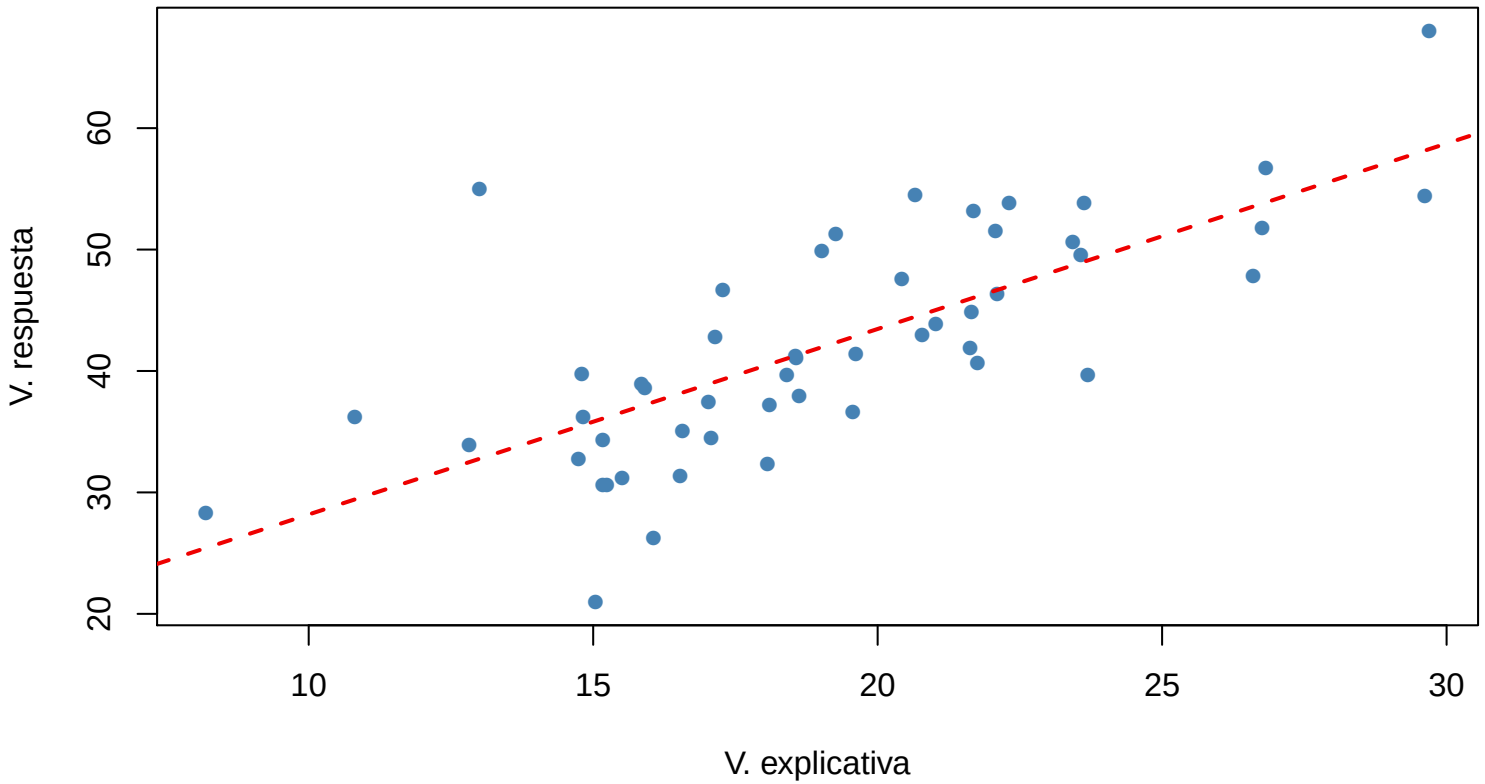
- Se puede medir la dependencia lineal de X_k a través del coeficiente R^2 de una RLM de X_k contra el resto de las variables auxiliares, R_k^2. Si el R_k^2 es cercano a 1, entonces X_k es *casi* una c.l. del resto de las X.
- En los modelo RLM con intercepto se puedes mostrar que

$$\hat{V}(\hat{\beta}_k) = \frac{\hat{\sigma}^2}{(n-1)V(X_k)} \frac{1}{1-R_k^2}.$$

- En la ecuación anterior, el segundo factor se conoce como Factor de Inflación de la Varianza (VIF) y se interpreta como el número de veces que es más grande $V(\hat{\beta}_k)$ si X_k que en el caso que X_k fuera ortogonal al resto de las X.
- En la práctica se suele tomar como punto de corte 5 o 10.

Observaciones atípicas: Definición

Outlier: Es un punto que *no se comporta como al resto*.



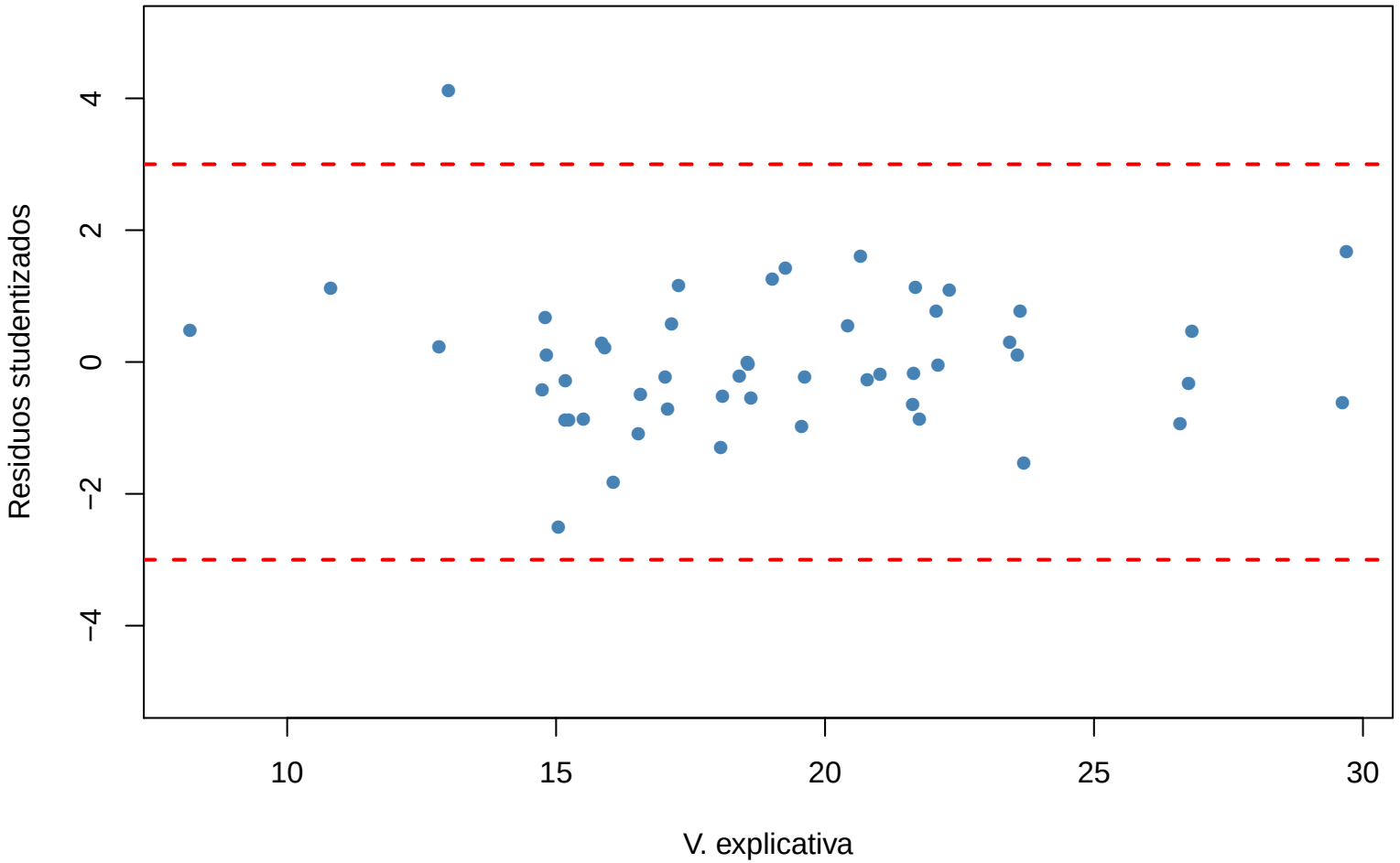
Observaciones atípicas: residuos estandarizados

- Un primer intento para encontrar este tipo de observaciones consiste en obtener los residuos *estandarizados*:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

- La idea de esto es homologar la varianza, pues teóricamente los residuales no tiene varianza constante.
- Luego se espera que estos residuales r_i se encuentren en una banda entre -3 y 3, aquellos fuera de estas bandas serán candidatos a ser analizados como outliers del modelo.

Observaciones atípicas: residuos estandarizados



Observaciones atípicas: residuos studentizados

- Para eliminar la posible influencia del dato atípico se utilizan los residuos de validación cruzada o *jackknife* o *studentizados*, que se definen como

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)} \left(1 + \mathbf{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i\right)^{1/2}}$$

  donde $\hat{y}_{(i)}$ es el valor ajustado de la i-ésima observación con el modelo ajustado quitando esa misma observación, $\hat{\sigma}^2_{(i)}$ es el estimador de σ^2 del modelo sin considerar la i-ésima observación. De la misma forma $\mathbf{X}_{(i)}$ es la matriz de diseñoo sin el i-ésimo renglón.

- Existe una forma fácil de calcular t_i a partir de r_i

$$t_i = r_i \left(\frac{n-p-1}{n-p-r_i^2}\right)^{1/2}$$

Observaciones atípicas: residuos studentizados

- Se puede probar que bajo el supuesto de que normalidad en los errores $t_i \sim t_{n-p-1}$.
- para detectar una observación atípica se suele utilizar la corrección de Bonferroni y comparar contra el cuantil α/n de una distribución t_{n-p-1}.
- Por ejemplo si $\alpha = 0.05$ entonces se dice que la i-ésima observación es atípica si

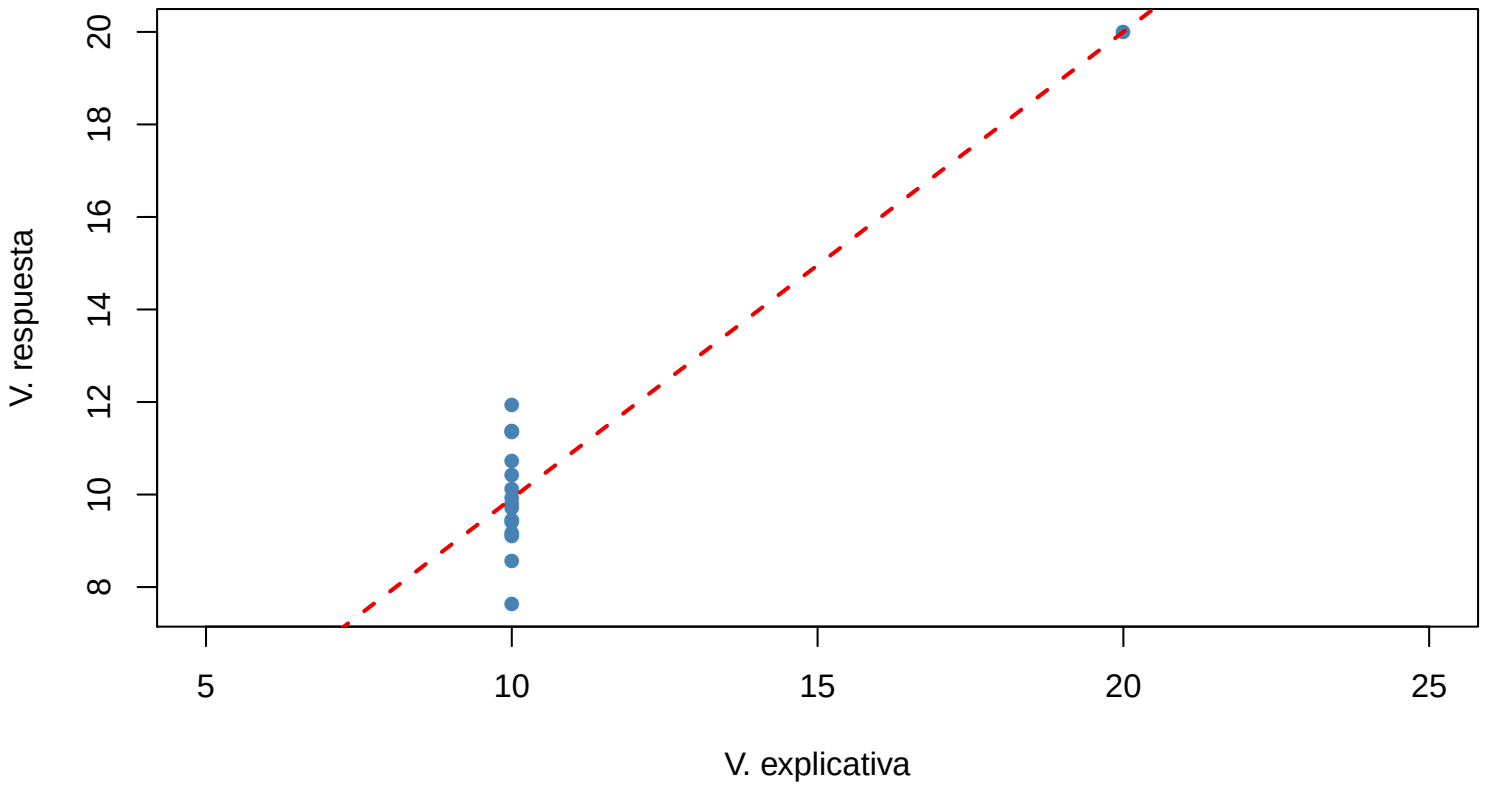
$$|\, t_i \,| > t_{n-p-1}^{(1-\alpha^*)} \qquad \text{con } \alpha^* = \frac{0.05}{2n}$$

Observaciones

- Dos o mas observaciones atípicas pueden ocultarse entre ellas.
- Se deben buscar atípicos después de cada transformación aplicada.

Observaciones influyentes: definición

Un punto influyente es aquel que siendo removido del modelo causa un cambio importante en el ajuste del modelo.



Detección de observaciones atípicas

Consideraremos cuatro formas para detectar observaciones influyentes, los cuales se basan en los siguientes principios:

- La distancia de cada observación con respecto al conjunto de datos, ya que en general una observación que se encuentra lejos de la región de observación ocasiona que el modelo cambie de manera significativa.
- Cambios en las estimaciones $\hat{\boldsymbol{\beta}}$ cuando se elimina la i-ésima observación.
- Cambios en los valores ajustados $\hat{\mathbf{y}}$ cuando se elimina la i-ésima observación.
- Cambios en la estimación de la varianza de $\hat{\boldsymbol{\beta}}$.

Observaciones influyentes: *leverage* o apalancamiento

- Para detectar los puntos de muestreo que están alejados se utiliza la matriz sombrero, $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$.
- A los elementos de la diagonal de $\mathbf{H}$, h_{ii}, se les conoce como el *leverage* (apalancamiento) o la influencia de cada observación. De este modo un *leverage* grande nos habla de una observación alejada de la masa de los puntos de muestreo.
- Se puede mostrar que $\text{tr}(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = p$, en promedio los *leverages* toman valen p/n.
- Existe una regla muy utilizada que dice que aquellos h_{ii} que sean mayores a dos veces al promedio son puntos alejados y por tanto deben de ser analizados, esto es $h_{ii} > 2p/n$.

Observaciones influyentes: $DFBETAS_{i,k}$

- El foeficiente $DFBETAS_{i,k}$ permite medir cómo cambia la estimación de β_k cuando se elimina la i-ésima observacióndel conjunto de datos, $k = 1, \ldots, p$ e $i = 1, \ldots, n$.

$$DFBETAS_{i,k} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\hat{\sigma}_{(i)}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{kk}}}$$

  donde $\hat{\sigma}^2_{(i)}$ es la estimación de σ^2 por MCO sin considerar la i-ésima observación y $(\mathbf{X}'\mathbf{X})^{-1}_{kk}$ es el k-ésimo elemento de la diagonal de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$.

- Se sugiere que si $|\, DFBETAS_{i,k} \,| > 2/\sqrt{n}$, la i-ésima observación tiene un influencia sobre el coeficiente k.

Observaciones influyentes: $DFFITS_i$.

- El coeficiente $DFFITS_i$ permite medir la influencia de la i-ésima observación sobre su valor ajustado $\hat{y}_i$. El $DFFITS_i$ se calcula como:

$$DFFITT_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}$$

  donde $\hat{y}_{i(i)}$ es el valor ajustado para y_i, obtenido sin usar la i-ésima observación y h_{ii} es el i-ésimo elemento de la diagonal de la matriz $\mathbf{H}$.

- El denominador sirve para estandarizar, ya que se puede mostrar que $V(\hat{Y}_i) = \sigma^2 h_{ii}$.

- El $DFFITS_i$ se interpreta como el número de desviaciones estándar que cambia el valor ajustado $\hat{y}_i$ si se elimina la observación i. Se sugiere investigar toda observación tal que $|\, DFFITS_i \,| > 2\sqrt{p/n}$.

Observaciones influyentes: $COVRATIO_i$

- Se define la *varianza generalizada* de $\hat{\boldsymbol{\beta}}$ como el determinante de su matriz de covarianzas, es decir

$$VG(\hat{\boldsymbol{\beta}}) = \left| V(\hat{\boldsymbol{\beta}}) \right| = \left| \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \right|$$

- Para tener una medida de cuanta precisión se gana o se pierde por quitar una observación se define el coeficiente:

$$COVRATIO_i = \frac{\left| \sigma^2_{(i)}(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1} \right|}{\left| \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \right|}$$

- Si el tamaño de muestra es grande, Belsley, Kuh y Welsh (1980) sugieren considerar al punto i como influyente si

$$|\, COVRATIO_i - 1 \,| > 3p/n.$$