Unidad 5

Diagnosticos del modelo y medidas correctivas

Supuestos del modelo RLM

La teoría que hemos desarrollado en el curso se basa en los siguientes supuestos:

- Linealidad:
$$E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$
- Homocedasticidad:
$$V(Y|\mathbf{X}) = \sigma^2.$$
- Independencia de las observaciones:
$$Cov(Y_j, Y_k) = 0.$$
- Independencia lineal de las variables explicativas
$$r(X) = p + 1.$$
- Normalidad
$$\mathbf{Y}|\mathbf{X} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}).$$

Cuando no es factible suponer que nuestros datos satisfacen las cinco propiedades anteriores, los resultados obtenidos a través del curso **¡no son válidos!**.
El objetivo de esta sección es identificar desviaciones a los supuestos del modelo y cómo corregirlas cuando se presentan.

Los residuos del modelo

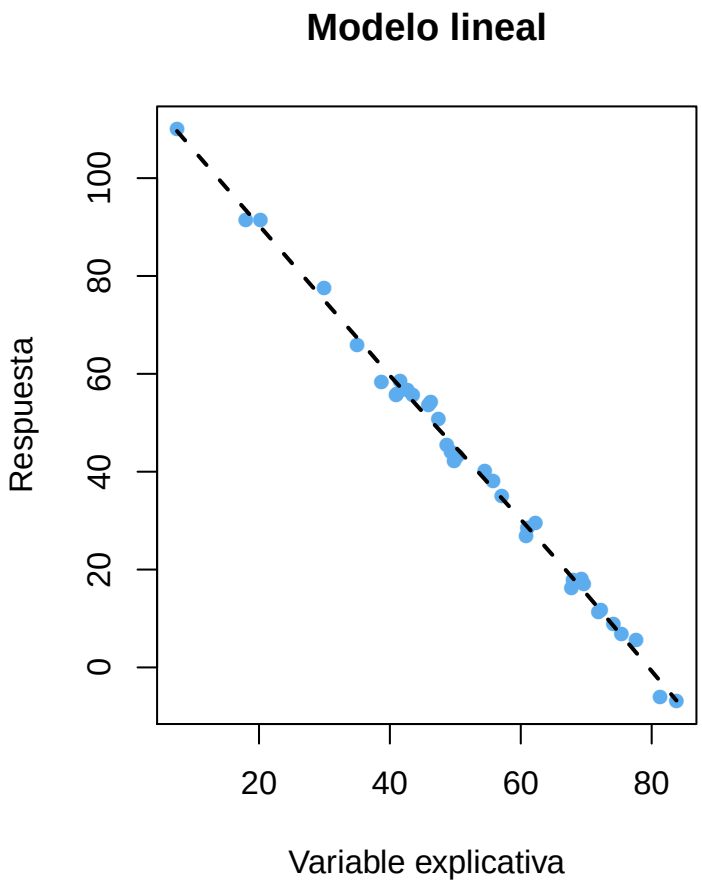
Algunas de las técnicas para detectar desviaciones a los supuestos del modelo se basan en los residuos. Antes de observar los datos

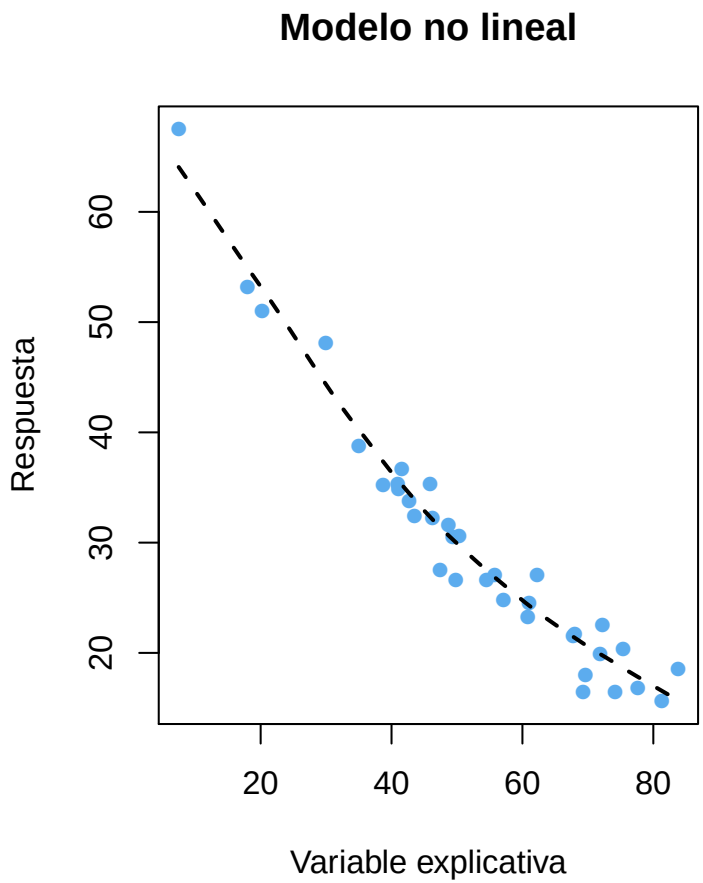
$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y},$$

por lo que $\mathbf{e} \sim N_n\left(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})\right)$. Como $\mathbf{I} - \mathbf{H}$, en general, no es diagonal, se sigue que los residuos del modelo no son independientes. Además la matriz $\mathbf{I} - \mathbf{H}$ no es de rango completo, por lo que la distribución de $\mathbf{e}$ es normal multivariada singular. Sin embargo, nos sirven para hacer muchos de los diagnósticos para detectar desviaciones a los supuestos del modelo.

Verificación del supuesto de linealidad

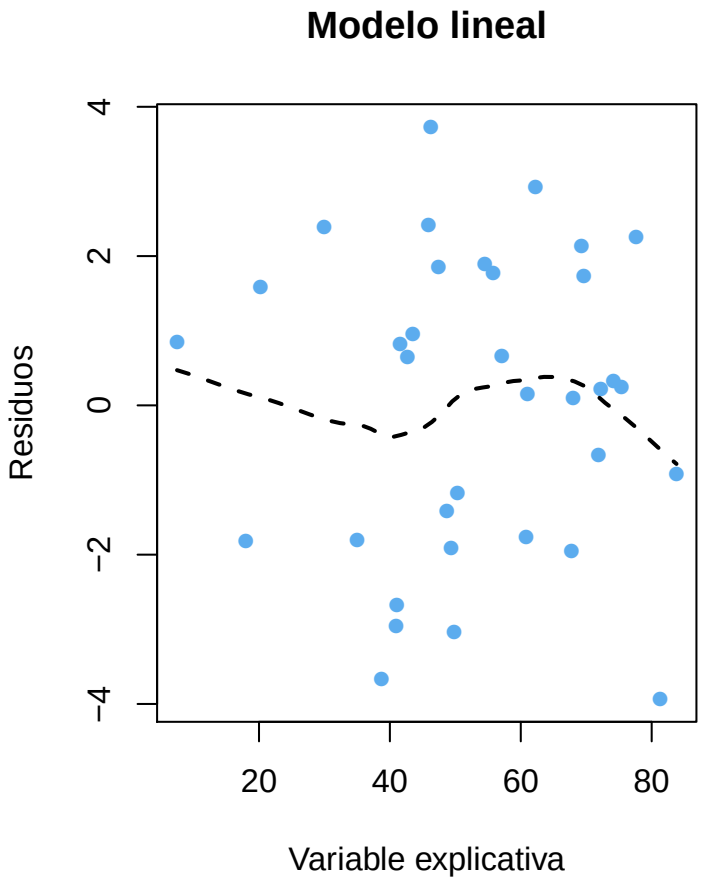
El primer supuesto a verificar es el de la linealidad del modelo. En el caso simple basta un gráfico de dispersión de y contra x.

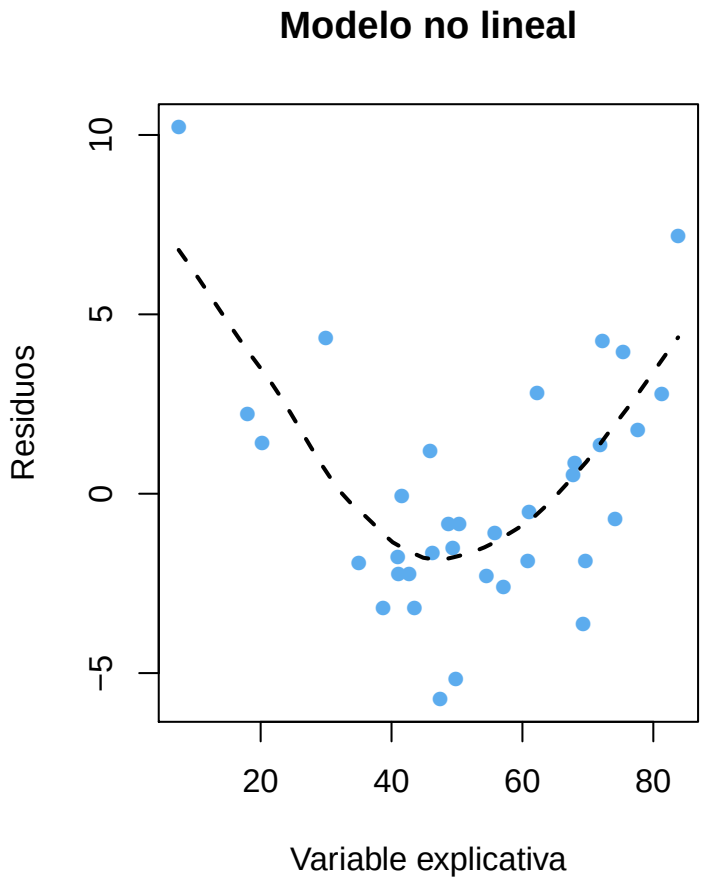




Verificación del supuesto de linealidad

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



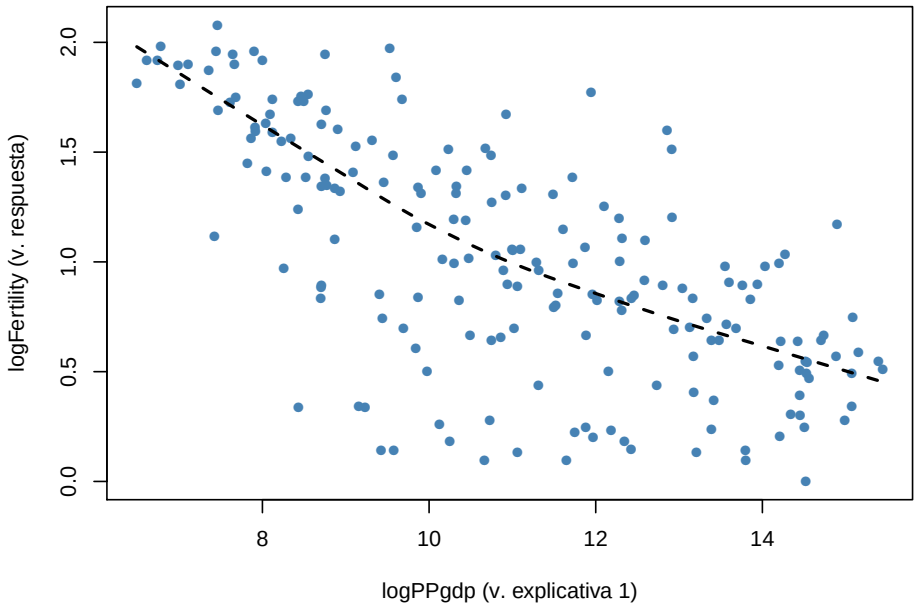


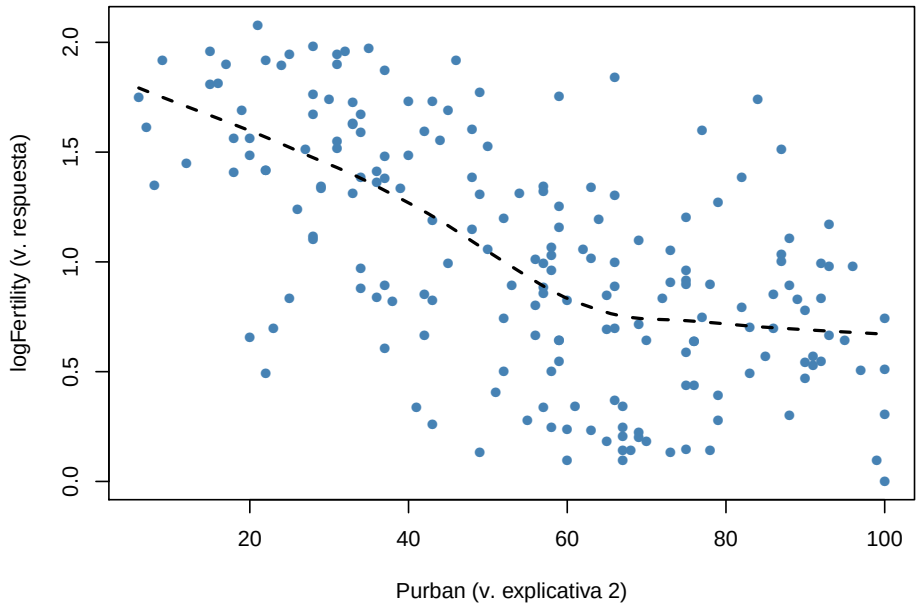
Verificación del supuesto de linealidad

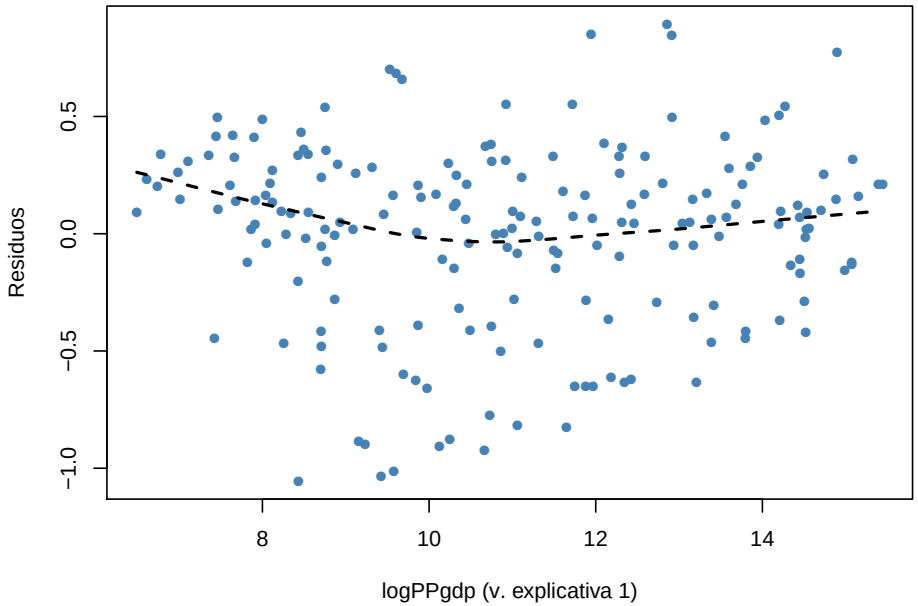
En el caso múltiple, se complica verificar la linealidad dado que se requiere graficar en dimensiones mayores y que puede haber asociaciones entre las variables explicativas.
Una solución simple es verificar las gráficas anteriores de respuesta y residuos contra cada X.
Las gráficas de respuesta vs. 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

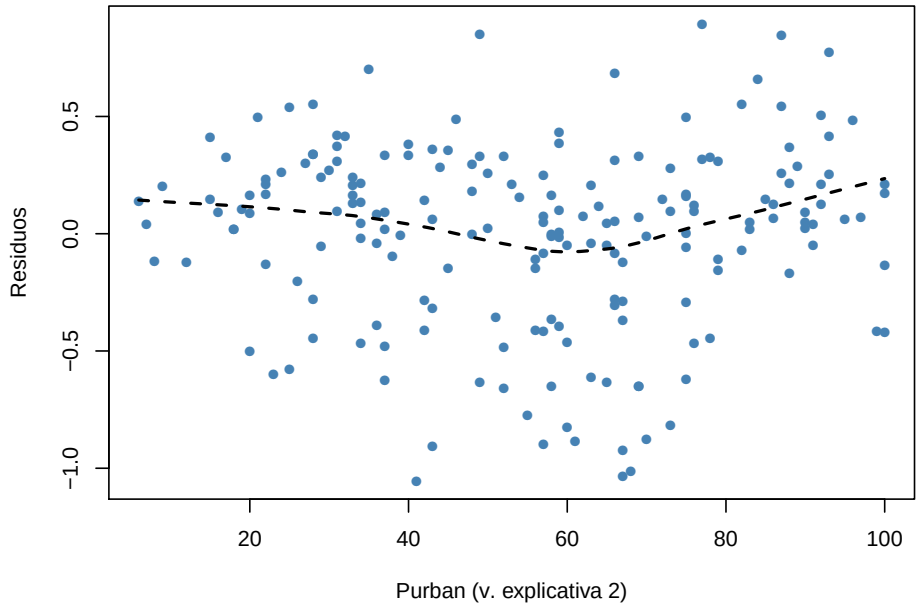
$$\texttt{logFertility} = \beta_0 + \beta_1 \texttt{PPgdp} + \beta_2 \texttt{Purban} + \epsilon.$$

Podemos *verificar* el supuesto de linealidad con las siguientes gráficas.


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
−1.0
−0.5
0.0
0.5
8
10
12
14
logPPgdp (v. explicativa 1)


Residuos
−1.0
−0.5
0.0
0.5
20
40
60
80
100
Purban (v. explicativa 2)

Verificación 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 consisderar únicamente X_k.

Ejemplo: falta de ajuste

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

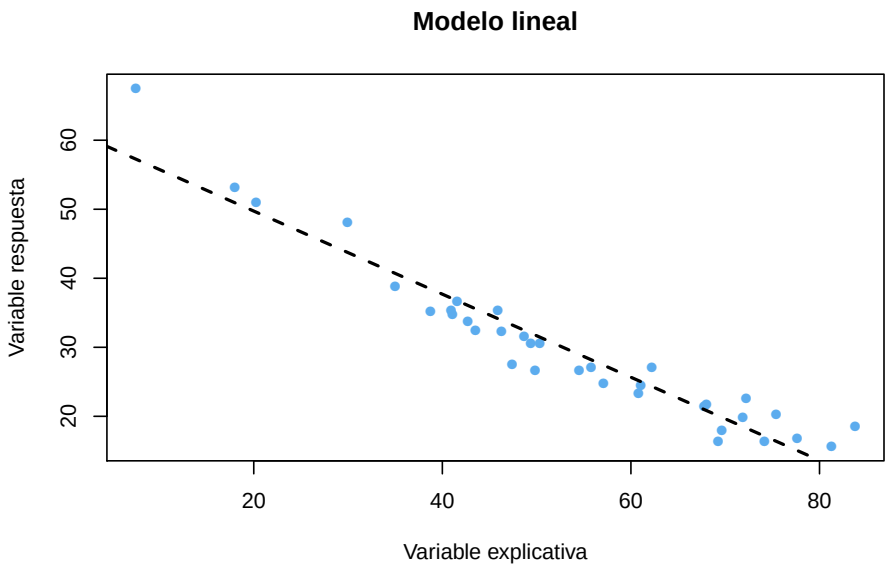
$$Y = 80 \exp(-0.02X) + \epsilon$$

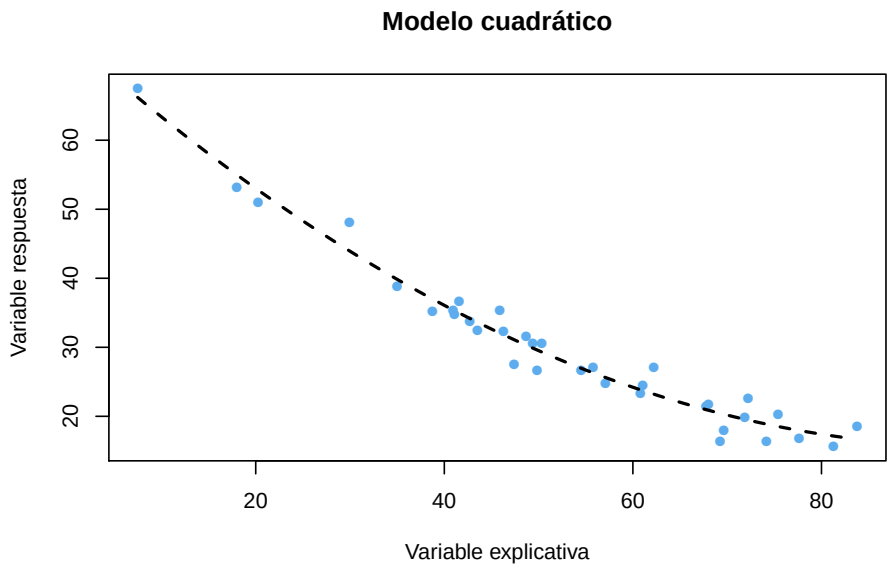
con $\epsilon \sim N(0, 4)$.
Aplicamos una prueba de falta de ajuste:

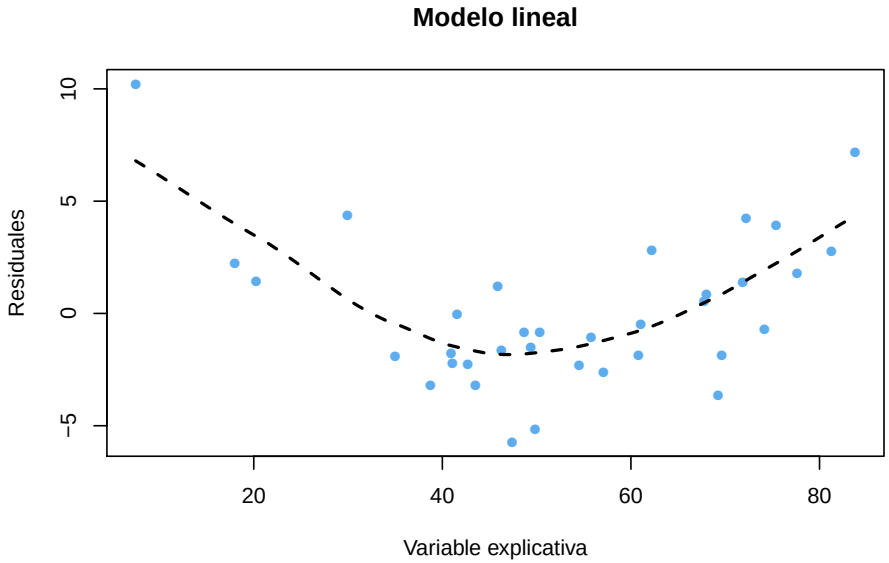
	Estimate	Std. Error	t value	Pr($>|t|$)
(Intercept)	75.0213	2.0754	36.15	0.0000
x	-1.2268	0.0857	-14.32	0.0000
x^2	0.0063	0.0008	7.48	0.0000

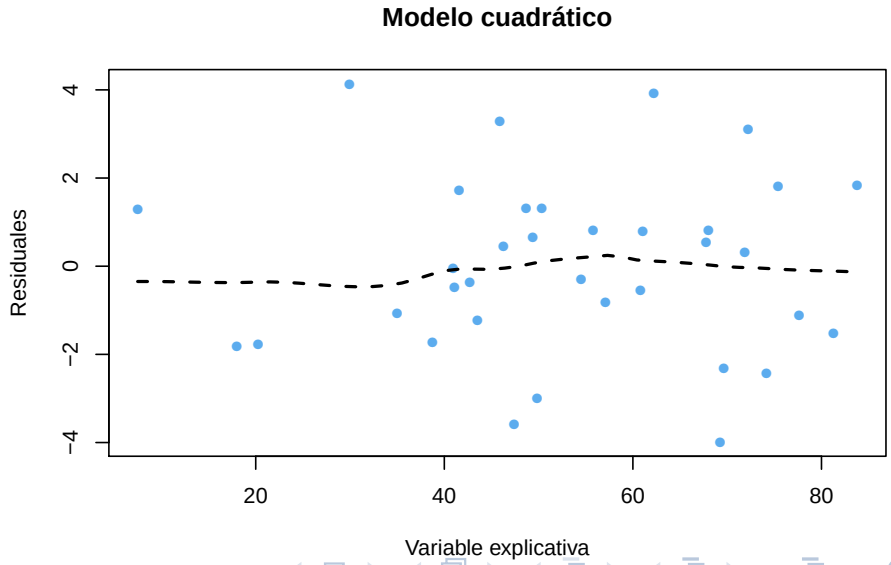
El resultado anterior indica que hay falta de linealidad en la v. explicativa x.

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).

Verificación del supuesto de homocedasticidad

Las principales desviaciones al supuesto de homocedasticidad se pueden resumir como

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

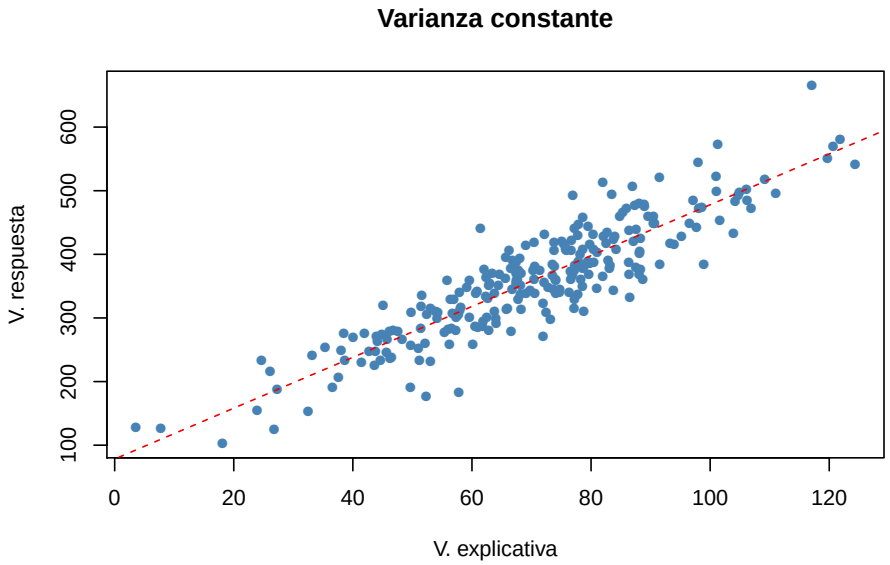
Es decir, la varianza de Y no es constante y (posiblemente) depende de $\mathbf{X}$.
¿En qué casos no se cumple el supuesto de varianza constante?

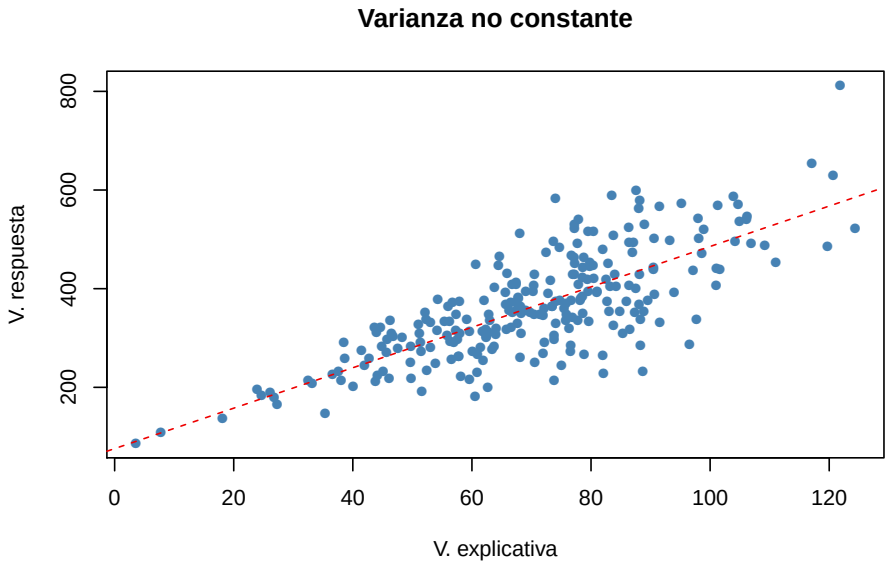
- Prácticamente en cualquier problema real.
- En la medición de magnitudes físicas, la precisión de las mediciones está relacionada con la magnitud que se desea medir.
- Cuando Y corresponde al total de m observaciones independientes con igual varianza σ^2, entonces $V(Y) = m\sigma^2$.
- Si Y es el promedio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

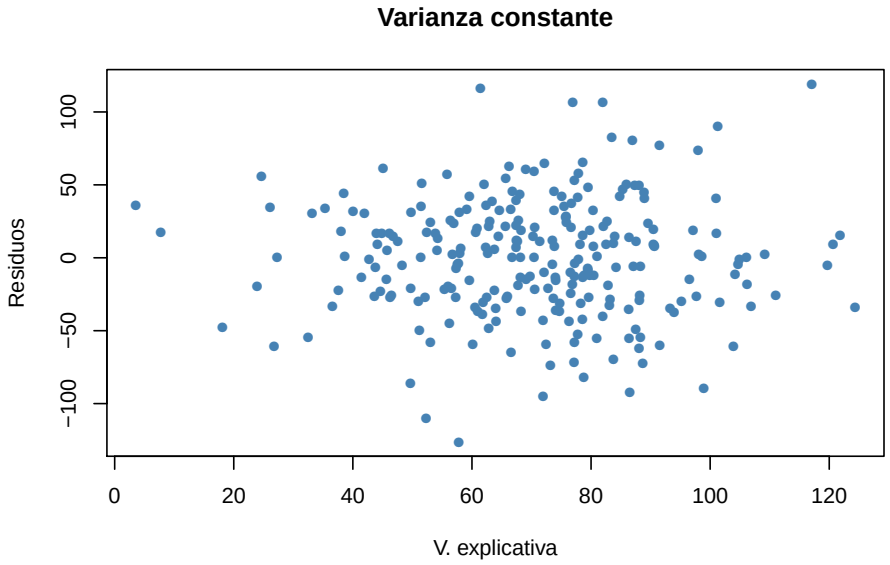
Verificación del supuesto de homocedasticidad

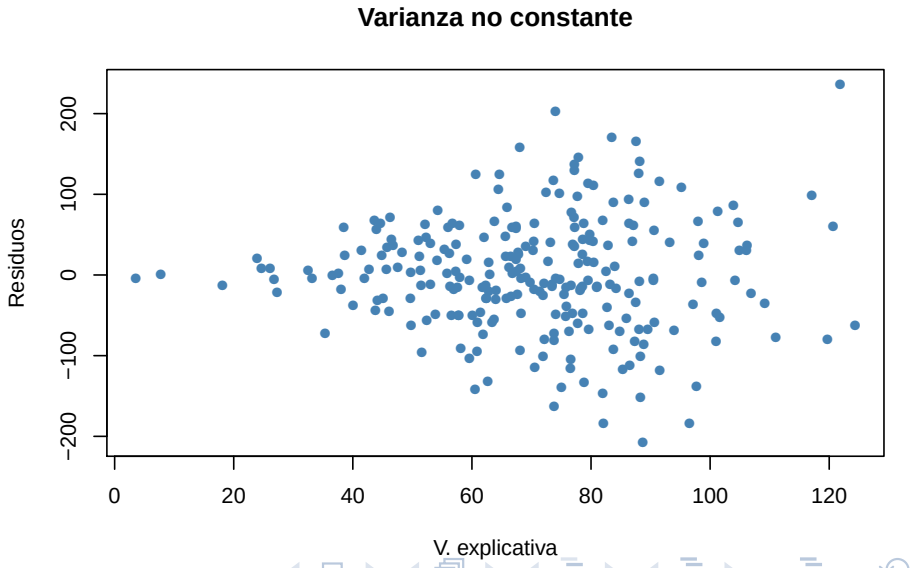
- Las desviaciones al supuesto de varianza constante se pueden detectar a partir de las gráficas de dispersión de la variable respuesta contra las explicativas o bien de los residuos contra las explicativas.
- Cuando no hay varianza constante, los residuos no se distribuyen uniformemente alrededor de 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 varianza estimada después de aplicar alguna corrección, pero menor que la varianza sin corregir.

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 $\hat{\sigma}^2$).
- Si el modelo anterior es significativo, entonces hay evidencia de que la varianza no es constante.
- La limitación de la prueba anterior es que sólo detecta cierto tipo de asociación entre la varianza y las variables explicativas, sin embargo ésta puede ser suficiente para detectar el problema y corregirlo.

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 re-emplazamos el supuesto de homocedasticidad por $V(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{W}$ donde mW 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

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

donde w_i es el i-ésimo elemento de la diagonal de $\mathbf{W}$. El estimador de β que resulta de minimizar $Q_{\mathbf{W}}$ se conoce como estimador de MCP. Si se asume normalidad conjunta en los errores, el EMV también coincide con el EMCP. En cualquier caso se puede mostrar que

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

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

Verificación del supuesto de homocedasticidad

Cuando se estima por MCP, se calculan los residuos ponderados

$$\hat{e}_i\mathbf{w} = \frac{1}{\sqrt{w_i}}\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}\right).$$

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.

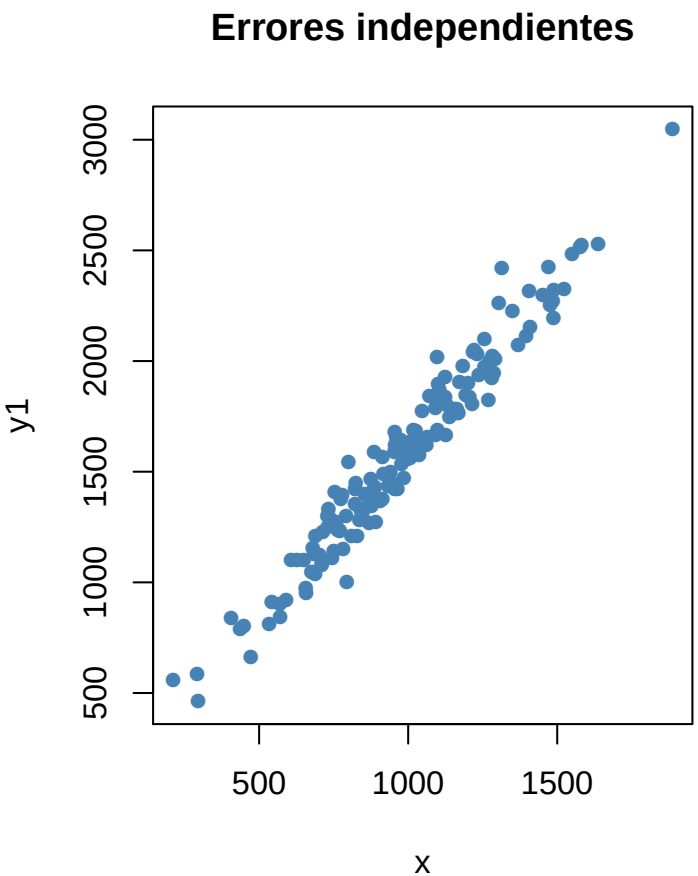
Verificación del supuesto de independencia

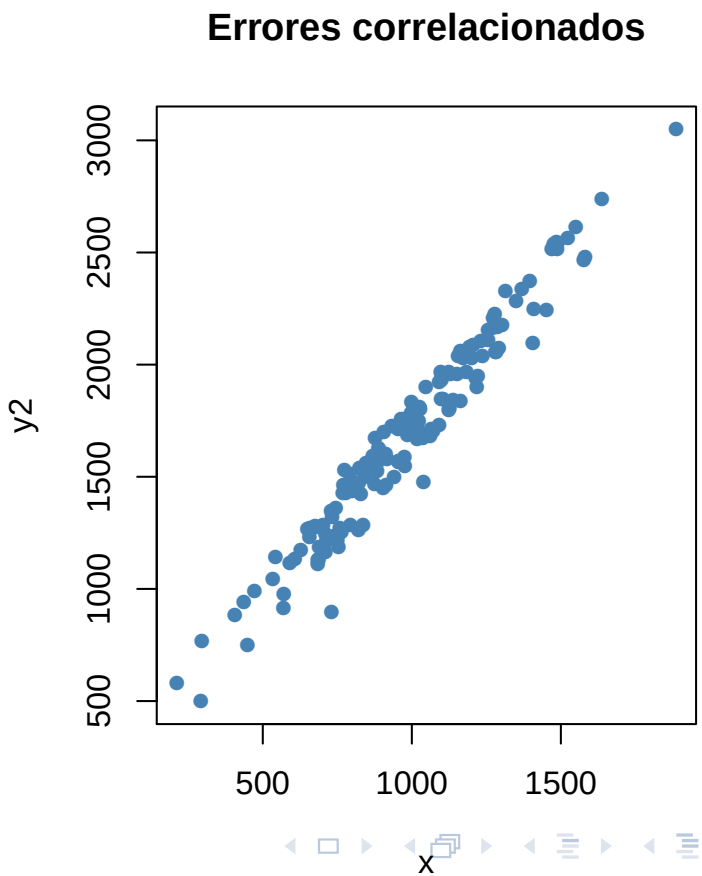
- Este supuesto es importante principalmente cuando los datos se obtuvieron con algún orden en el tiempo y las mediciones realizadas pueden estar influidas por mediciones previas.
- La presencia de errores correlacionados tiene efectos en la estimación de la varianza del modelo y de los estimadores. 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.
- En este caso nos concentraremos cómo detectar dependencia sobre el tiempo.

Verificación del supuesto de independencia

Consideremos dos conjuntos de datos con pares de observaciones (X, Y). El primer conjunto cumple con todos los supuestos del modelo RLS mientras que el segundo tiene errores correlacionados en el tiempo. En ambos casos el modelo es

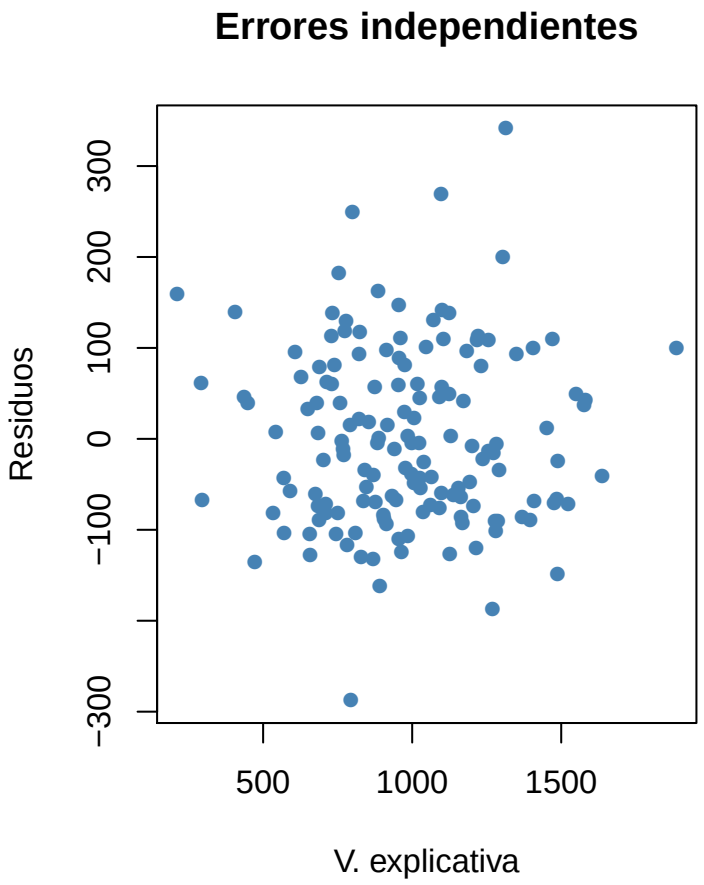
$$E(Y|X) = 100 + 1.5X$$

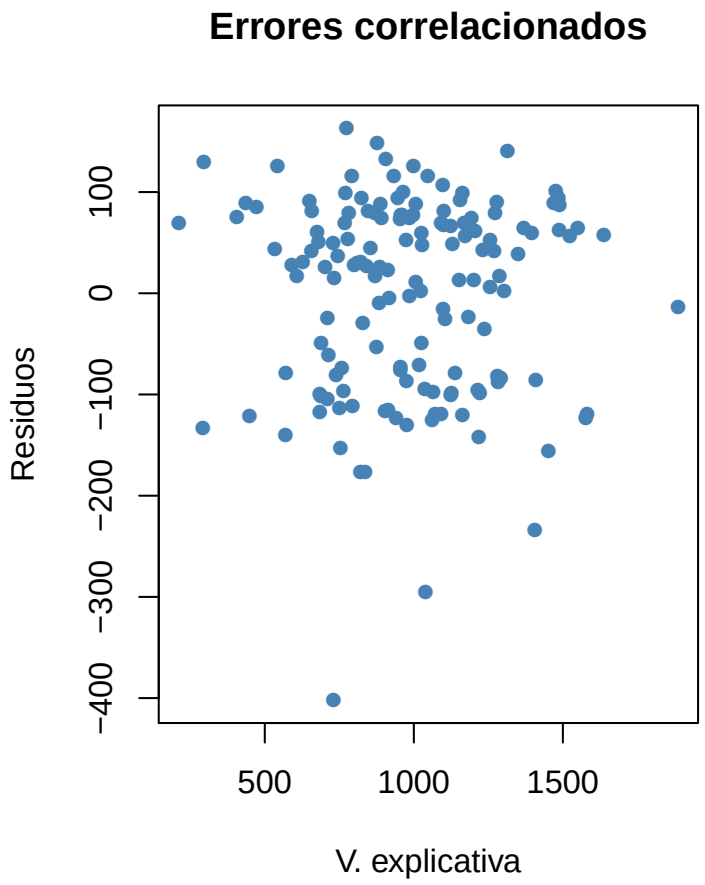




Verificación del supuesto de independencia

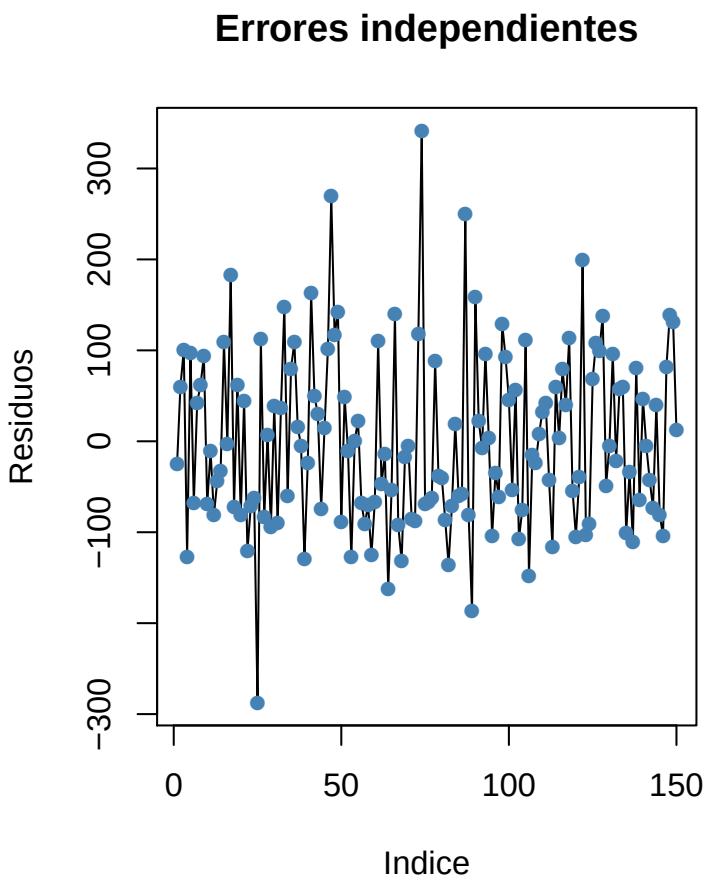
Cuando exploramos los residuos contra la variable explicativa, no parece haber desviaciones a los supuestos del modelo.

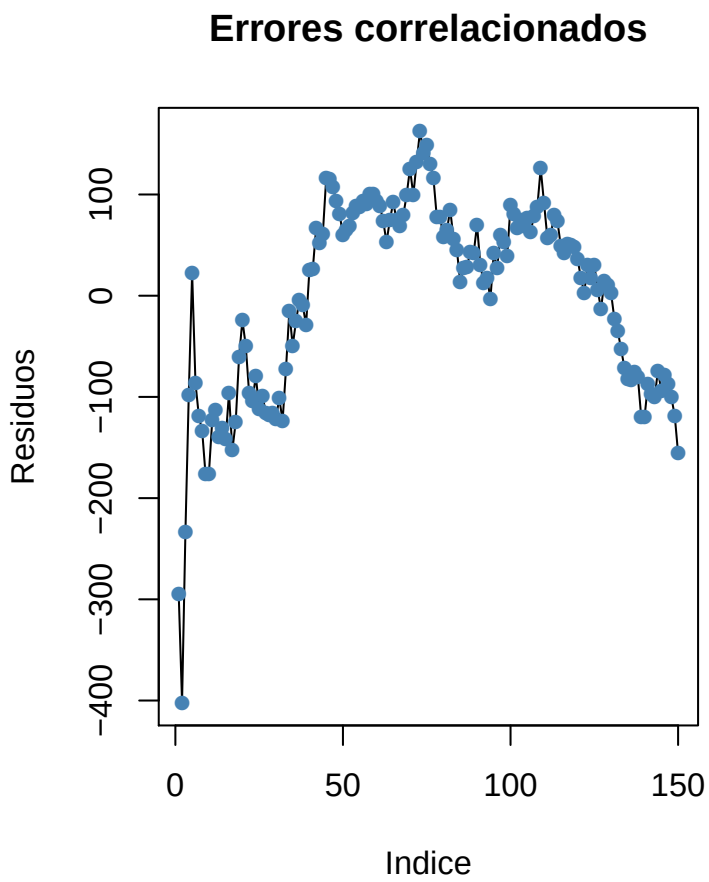




Verificación del supuesto de independencia

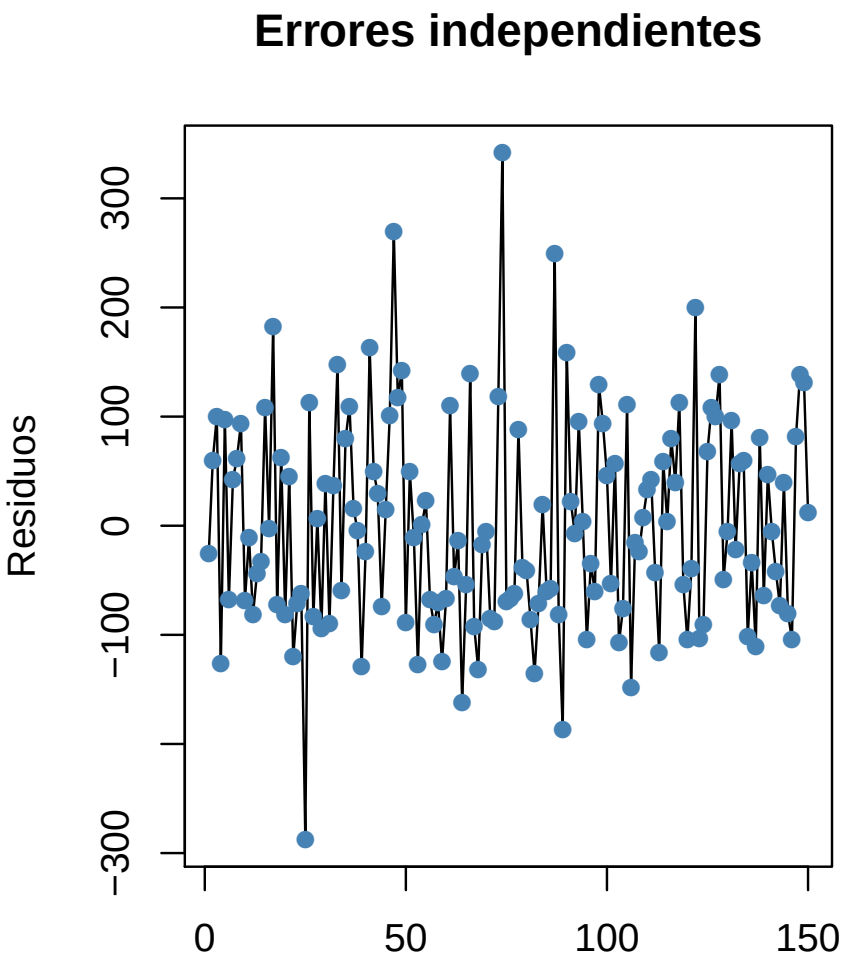
Los patrones aparecen cuando graficamos los residuos contra el tiempo.

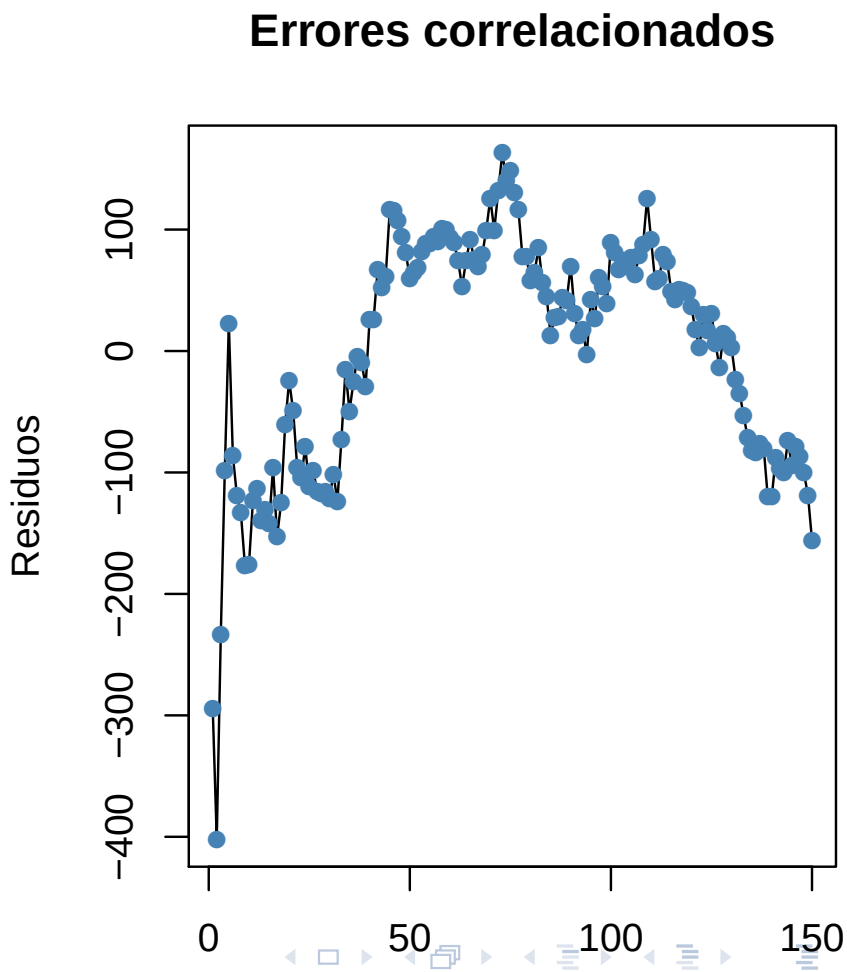




Verificación del supuesto de independencia

Además de graficar los residuales contra el tiempo, es posible graficar su función de autocorrelación, que mide la dependencia temporal de los errores a distintos rezagos.





Verificación del supuesto de independencia

Para corroborar las sospechas, podemos aplicar la prueba de rachas (para comprobar aleatoriedad) o la prueba de Durbin-Watson (para probar autocorrelación).

La prueba de rachas se encuentra en el paquete `randtests` y se ejecuta con la función `runs.test`.

```
> library(randtests)
> runs.test(residuos1)
Runs Test

data: residuos1
statistic = -0.3277, runs = 74, n1 = 75, n2 = 75, n = 150, p-value = 0.7431
alternative hypothesis: nonrandomness

> runs.test(residuos2)
Runs Test

data: residuos2
statistic = -10.323, runs = 13, n1 = 75, n2 = 75, n = 150, p-value < 2.2e-16
alternative hypothesis: nonrandomness
```

Verificación del supuesto de independencia

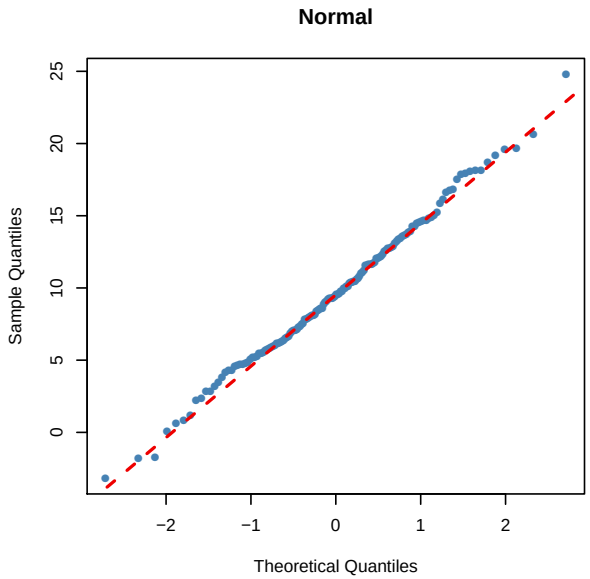
La prueba de Durbin-Watson se encuentra en el paquete `lmtest` y se ejecuta con la función `runs.test`.

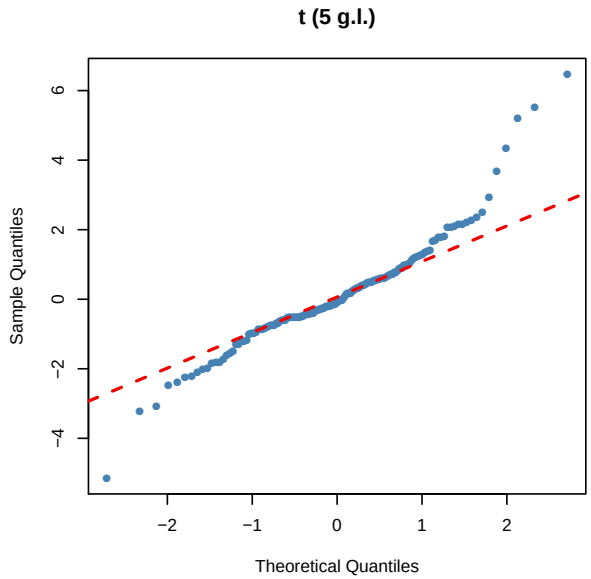
```
> library(lmtest)
> dwtest(modelo1)
Durbin-Watson test

data: modelo1
DW = 1.9784, p-value = 0.4474
alternative hypothesis: true autocorrelation is greater than 0

> dwtest(modelo2)
Durbin-Watson test

data: modelo2
DW = 0.12018, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

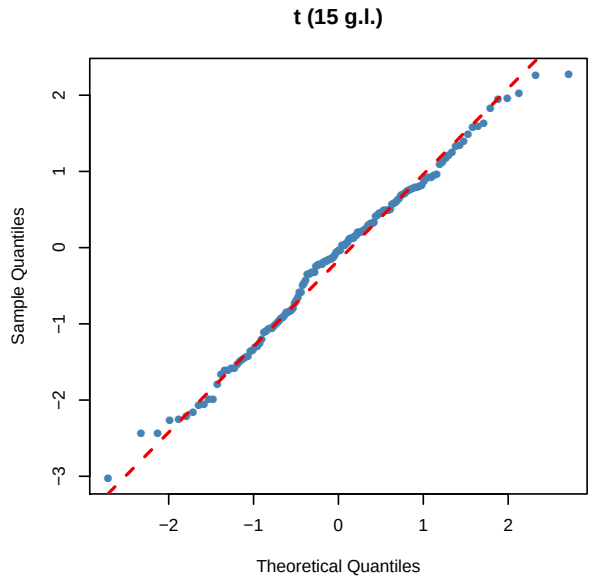
Verificación del supuesto de normalidad

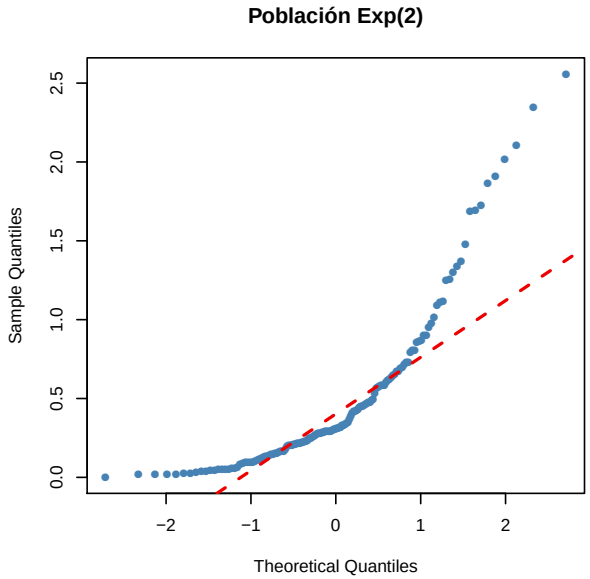
- Las desviaciones al supuesto de normalidad afectan la significancia de las pruebas y la confianza de los intervalos que hemos construido.
- Afortunadamente, cuando el tamaño de muestra es grande, los estimadores EMV se distribuyen aproximadamente normales. La *velocidad* de la convergencia es mayor 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 de normalidad.

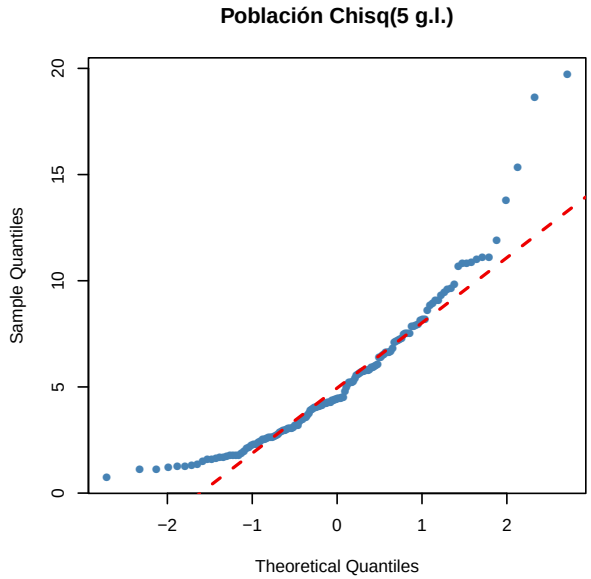
Verificación del supuesto de normalidad

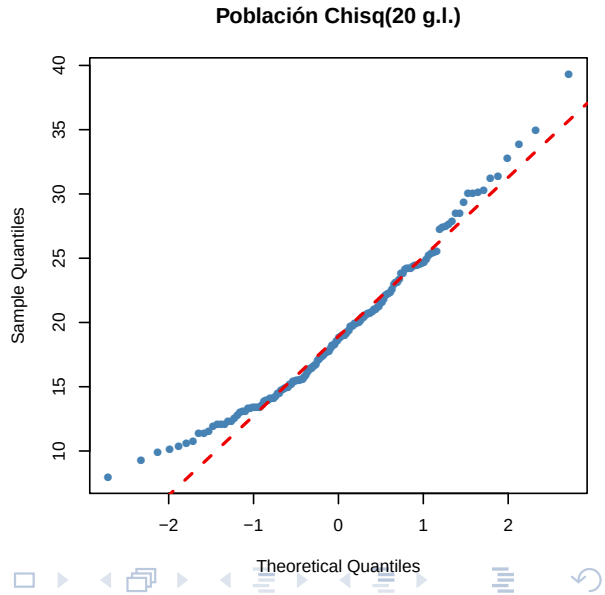












Verificación del supuesto de normalidad

- 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 β es sencillo, para la prueba ANOVA no lo es tanto.

Verificación del supuesto de independencia de las X

- Este supuesto se traduce en que la matriz de diseño $\mathbf{X}$ es de rango completo, ya que esto garantiza que $\mathbf{X}'\mathbf{X}$ es invertible y así obtener 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 invertibles, 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 formalmente no hay cómo decidir si hay o no problemas de multicolinealidad.

Verificación del supuesto de independencia de las X

- Otra forma de detectar la multicolinearidad es con el *índice de condición* de la matriz de diseño $\mathbf{X}'mX$, 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

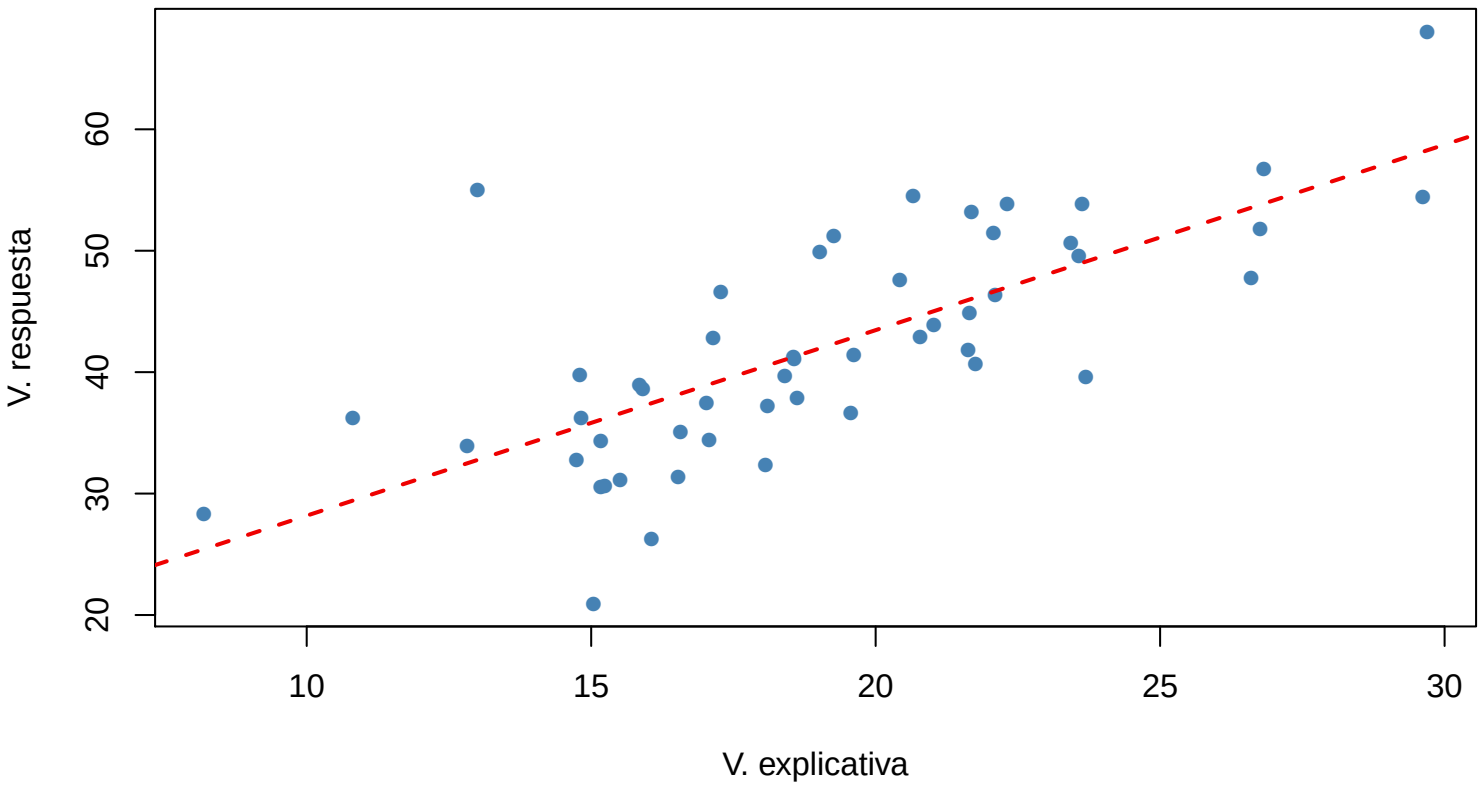
- También es posible hacer regresiones lineales de X_k contra el resto de las variables y obtener el coeficiente R_q^2 de dicha regresión. Un coeficiente R_q^2 cercano a 1, indica que hay un ajuste muy por lo que X_k es *casi* una combinación lineal del resto de las X.
- Existe una igualdad que permite entender la relación de la varianza del estimador de β_k y la dependencia linea de X_k.

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

- 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 $V(\hat{\beta}_k)$ que si X_k fuera ortogonal al resto de las X.
- En la práctica se suele tomar como punto de corte 5 o 10.
- Las posibles soluciones al problema de multicolinealidad son: eliminar las variables linealmente dependientes o aplicar componentes principales.

Observaciones atípicas

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

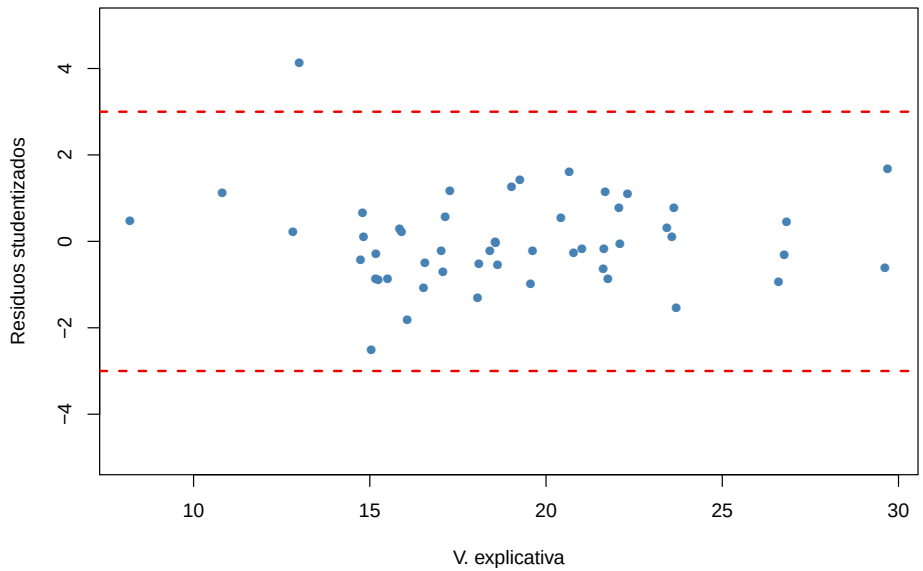


Observaciones atípicas

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

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

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

Para eliminar la posible influencia del dato atípico se utilizan los residuos de validación cruzada o *jackknife*, 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

Finalmente, se puede probar que bajo el supuesto de que normalidad en los errores $t_i \sim t_{n-p-1}$.
Luego, 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

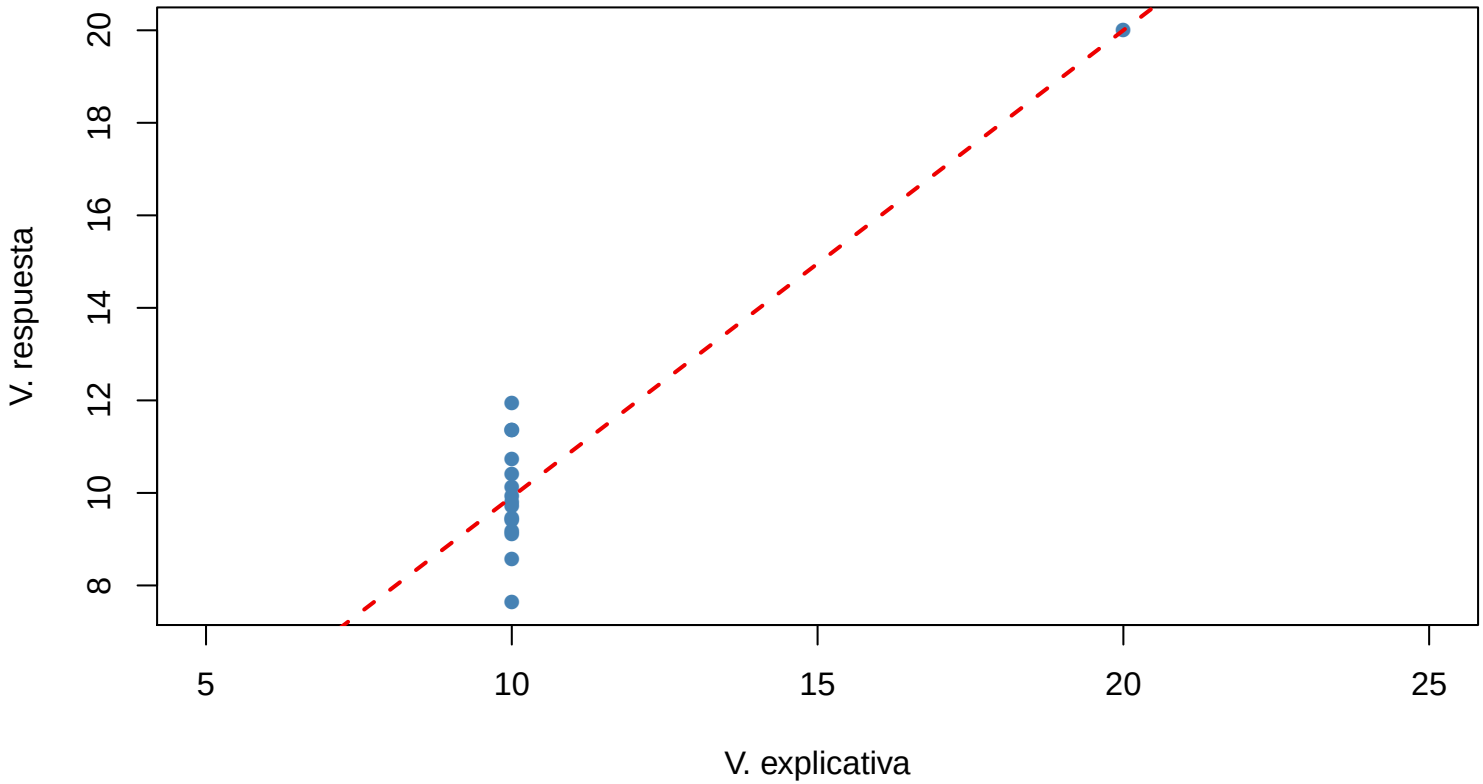
$$|t_i| > t_{n-p-1}^{(1-\alpha^*)} \qquad \text{con } \alpha^* = \frac{0.05}{2n}$$

Notas:

- Dos o mas *outliers* juntos pueden ocultarse entre ellos.
- Un *outlier* en un modelo puede dejar de serlo en otro cuando hubieron transformaciones, se recomienda hacer el análisis de *outliers* cada vez que se hace alguna transformación al modelo.

Observaciones influyentes

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



Observaciones influyentes

En general una observación que se encuentra lejos de la región de observación ocasiona que el modelo cambie de manera significativa (aunque no necesariamente). Para detectar los puntos de muestreo que están alejados se utiliza la matriz sombrero, $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t$. A los elementos de la diagonal de $\mathbf{H}$ (h_{ii}) se les conoce como el *leverage* 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. Ya que $tr(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = p$, en promedio los *leverages* toman el valor de $\frac{p}{n}$. 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. $\left(h_{ii} > 2\frac{p}{n}\right)$

Observaciones influyentes

Algunas medidas para detectar la influencia de una observación son:

- 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.

Belsey, Kuh y Welch introducen los coeficientes $DFBETAS_{i,k}$ y $DFFITS_i$. La primera es una estadística que indica cómo cambia la estimación de β_k cuando removemos la i-ésima observación, $k = 1, \ldots, p$ 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

El coeficiente $DFFITSi, j$ 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 usa 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 $\text{Var}\left(\hat{Y}_i\right) = \sigma^2 h_{ii}$.

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 obtervación tal que $|DFFITS_i| > 2\sqrt{p/n}$.

Observaciones influyentes

Otra forma de determinar la influencia de una observación es a través de la precisión $\hat{\beta}$. Se define la *varianza generalizada* de $\hat{\beta}$ como el determinante de su matriz de covarianzas, es decir

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

Finalmente para tener una medida de cuanta precisión se gana o se pierde por quitar una observación se definimos 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 que si

$$COVRATIO_i > 1 + 3p/n \quad \text{o} \quad COVRATIO_i < 1 - 3p/n$$

Entonces se debe considerar al punto i como influyente.

Unidad 6
Selección de modelos

Métodos de selección de variables

Cuando se tiene un número grande de variables potencialmente explicativas a una respuesta Y, se puede tener interés en *seleccionar* con algún procedimiento el *mejor* subconjunto de ellas, de manera que seamos capaces de ajustar el *mejor* modelo. El algoritmo general para llevar a cabo la selección del modelo se puede resumir en lo siguiente:

- Se emplea un criterio de selección de variables en particular
- El modelo resultante se revisa para verificar que las especificaciones funcionales sean correctas (Análisis de Residuales) y que no existan outliers ni observaciones de alta influencia.
- Si el modelo no pasa el punto anterior, se debe de repetir el proceso de selección de variables omitiendo el modelo resultante de las iteraciones anteriores.

Criterios para Evaluar Modelos de Regresión

A continuación se presentan criterios para evaluar y comprar modelos de regresión, debe de quedar claro que todos estos modelos deben de ser submodelos (incompletos) de un modelo general con todas las variables (saturado):

- R^2 ajustado
- Verosimilitud maximizada
- Criterio de información de Akaike (AIC):

$$AIC = 2k - 2\hat{\ell}$$

donde k es el número de parámetros en el modelo y $\hat{\ell}$ es el máximo del logaritmo de la verosimilitud.

- Criterio de información de Bayes (BIC):

$$BIC = k \log n - 2\hat{\ell}$$

donde k es el número de parámetros en el modelo, n el número de observaciones y $\hat{\ell}$ es el máximo del logaritmo de la verosimilitud.

Criterios para Evaluar Modelos de Regresión

Existen algoritmos que determinan el mejor modelo con base en comparar alguno de los criterios anteriores. Los métodos mas utilizados son:

- *Forward*: el modelo inicial no tiene variables auxiliares y en cada iteración se van agregando según mejore o no empeore el ajuste del modelo.
- *Backward*: el modelo inicial contiene todas las variables auxiliares y en cada iteración se van eliminando variables según mejore (o no empeore tanto) el ajuste del modelo.
- *Stepwise*: es una combinación de los algoritmos anteriores, la diferencia es que en cada caso puede eliminar o agregar variables al modelo. Este método tiene el problema de que en alguna iteración se llegue a un ciclo repetido sacando y metiendo la misma variable, en cuyo caso se decide terminar el algoritmo en ese momento.

En `R`, la función `step` se utiliza para selección de modelos. Esta función utiliza por defecto el AIC y se puede especificar si se quiere hacer *forward*, *backward* o *stepwise*.