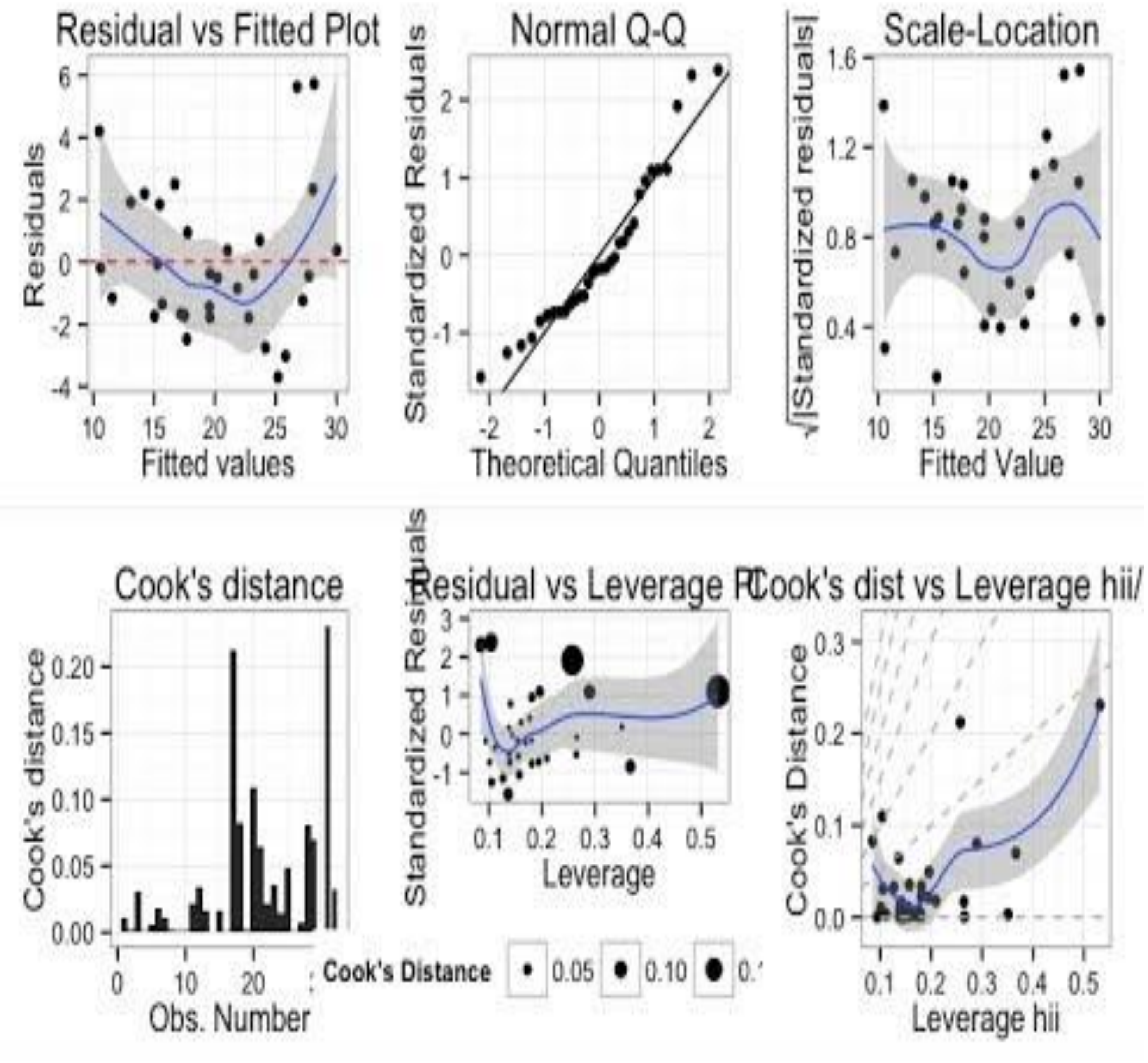


Condiciones de la RLM: los diagnósticos

Óscar Centeno Mora

Preámbulo

- Luego del proceso de seleccionar las variables o predictores que deben estar en la RLM, y habiendo estimado los coeficientes, es momento de ver si se cumplen o no las condiciones del modelo.

- ¿Qué entendemos por condiciones o supuestos del modelo?

- En el tema de la regresión lineal múltiple mencionamos los pasos a llevarse a cabo en la estimación o aplicación de una ecuación de regresión:
  1. Relación entre las variables
  2. Estimación de la recta de mejor ajuste (modelo de regresión)
  3. **Diagnóstico del modelo de regresión**
  4. Medidas remediales
  5. Estadísticas de bondad y ajuste
  6. Inferencia y prueba de hipótesis de los coeficientes del modelo

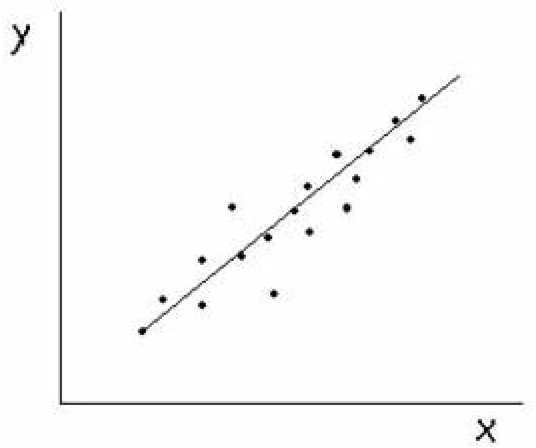
Preámbulo

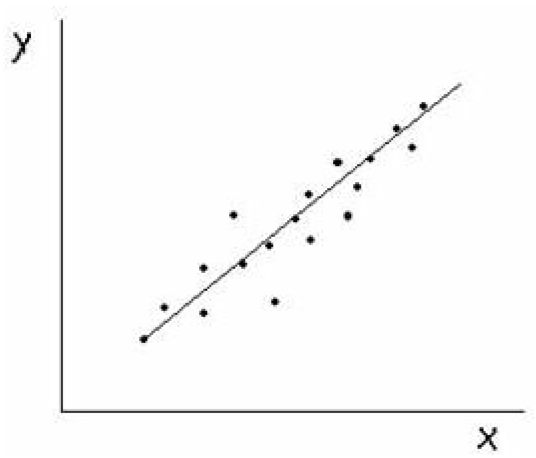
<u>Regresión bivariada</u>

Una variable dependiente (Y)
Una variable independiente (X)

$$Y \sim X$$

$$\begin{bmatrix} Y1 \\ Y2 \\ \cdots \\ Y_n \end{bmatrix} \sim \begin{bmatrix} X1 \\ X2 \\ \cdots \\ X_n \end{bmatrix}$$





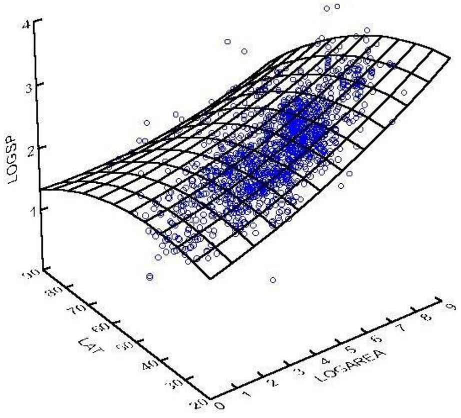
<u>Regresión multivariada</u>

Una variable dependiente
Dos o más variables independientes

$$Y \sim X$$

$$\begin{bmatrix} Y1 \\ Y2 \\ \cdots \\ Y_n \end{bmatrix} \sim \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}$$

¿se cumplen las condiciones ?



Preámbulo

Veremos como verificar las condiciones de la RLM:

$$\boxed{\varepsilon \sim N(0,\sigma^2 I)}$$

Índice

Parte I – SUPUESTOS DE LA RLM

Índice

Introducción

- Habíamos mencionado en el capítulo de la RLM, que la ecuación de regresión, aunque útil, no es del todo flexible, dado que supone unos condiciones que se deberían cumplir para su correcto uso:

  1. Relación lineal entre los predictores y la respuesta
  2. Variancia constante.
  3. Normalidad en sus residuos.
  4. No presencia de multicolinealidad.
  5. Homoscedasticidad o variancia constante de los errores.
  6. Independencia de errores o no autocorrelación de los errores (aplicable para datos temporales).

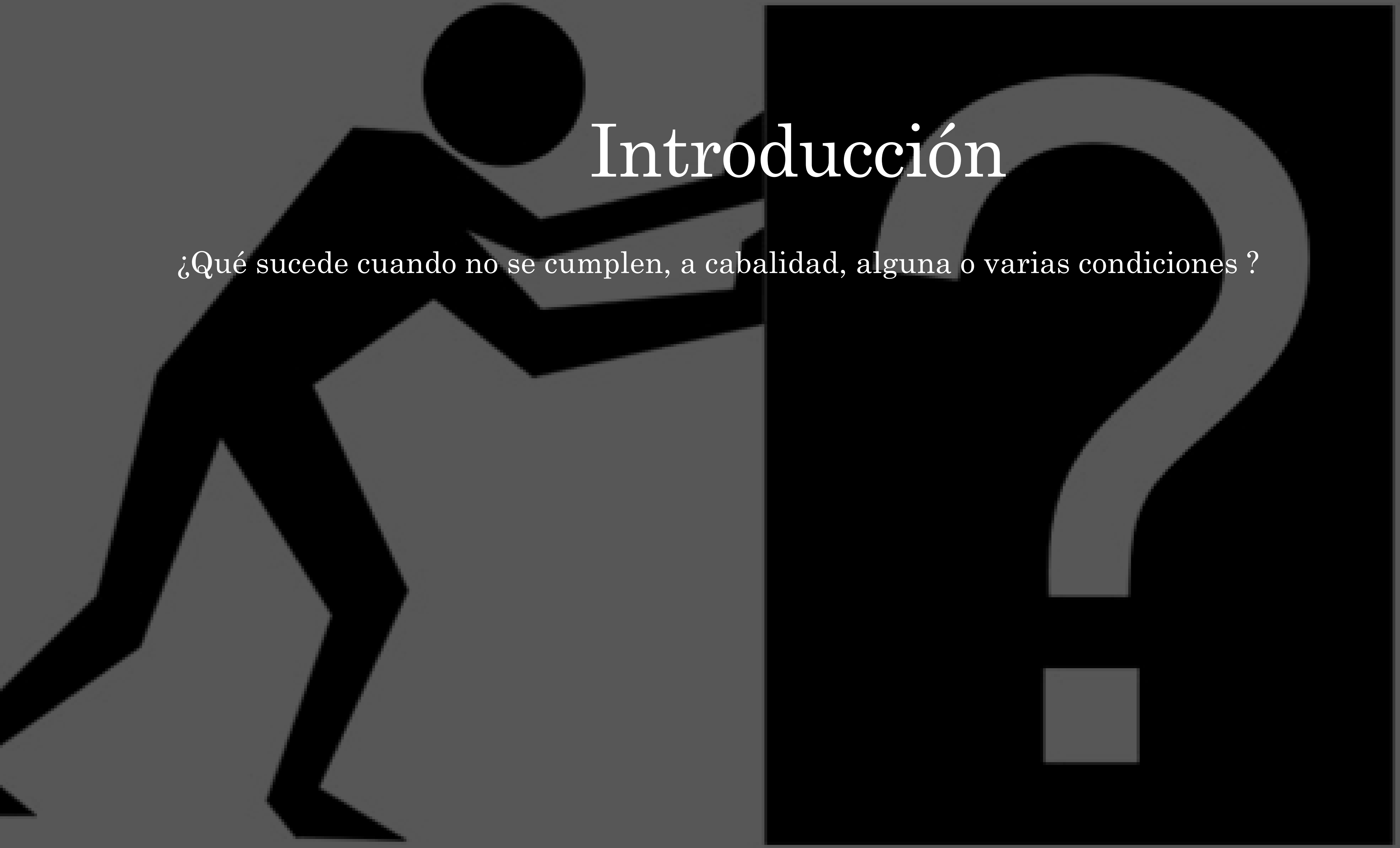
- Todas esas condiciones, se resumen en la siguiente ecuación de supuestos de la RLM:

$$\boxed{\varepsilon \sim N(0, \sigma^2 I)}$$

- Veremos como verificaremos las condiciones para la RLM.

Introducción

¿Qué sucede cuando no se cumplen, a cabalidad, alguna o varias condiciones ?



Índice

Prueba de normalidad de los residuos

- Al postular lo siguiente :

$$\boxed{\varepsilon \sim N(0,\sigma^2 I)}$$

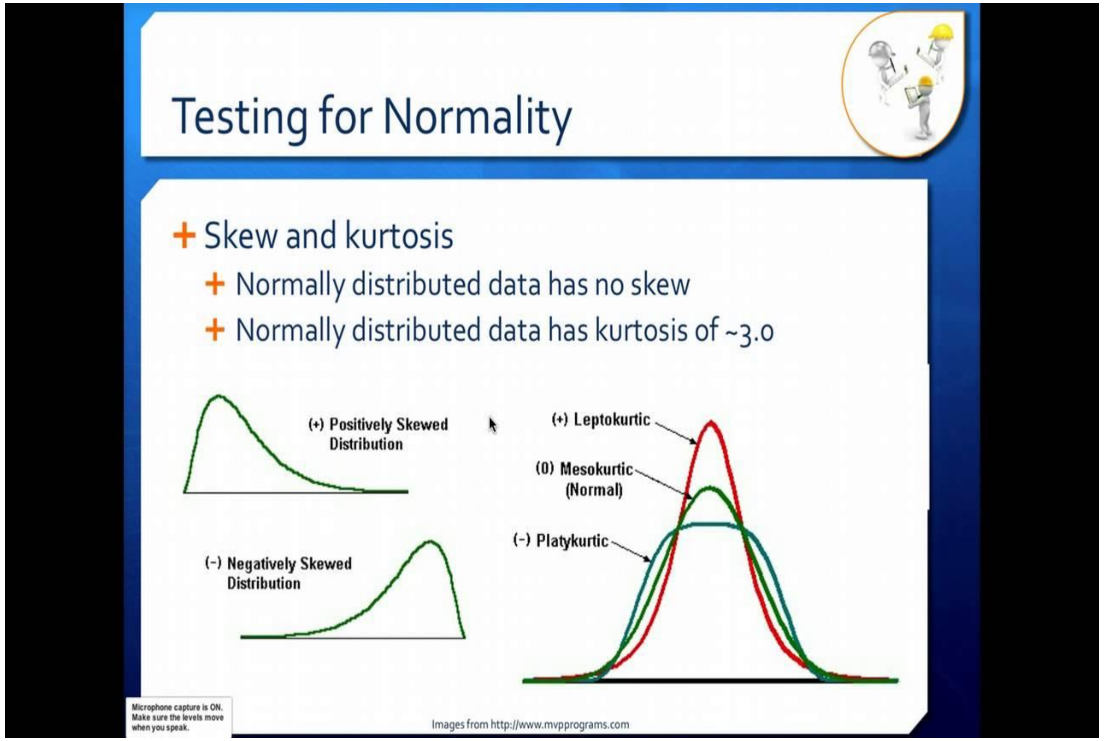
estamos indicando que los residuos (valores observados – valor ajustados), se distribuyen de como una nomal con media de cero, y con variancia constante

- Hay diversas formas de probar lo siguiente
  - ❑Pruebas de estadísticos como Kolgomorov – Smirnox, Shapiro Wiks, Shapiro Francia, etc.
  - ❑Método gráfico, como lo es el QQ Plot.

- Al realizar estas pruebas, es recommendable pasar tanto por pruebas de hipótesis, así como la aplicación del método visual de QQ Plot.

Prueba de normalidad de los residuos

- Aunque empezamos por explicar esta prueba, se considera que esta es la prueba menos flexible.
- Desviación o incumplimiento de los residuos normales puede ser causada por valores extremos, la multicolinealidad, la no linealidad entre la variable respuesta y los predictores, así como los otros supuestos.
- Aunque, a título personal soy de la línea de realizar todas las condiciones e inspeccionar cada uno de los casos, hay otros autores que dicen analizar esta prueba hasta el final … aunque es lo primero que suelen explicar…



Prueba de normalidad de los residuos

- Las pruebas de hipótesis sobre la normalidad se presentan aproximadamente de la siguiente forma:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Frisbee Throwing Distance (Metres)	.091	50	.200*	.963	50	.118

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

- En R:

```
> shapiro.test(normal)

        Shapiro-Wilk normality test

data:  normal
W = 0.97471, p-value = 0.5847
```

- ¿Cómo interpretaríamos dicha prueba?

Prueba de normalidad de los residuos

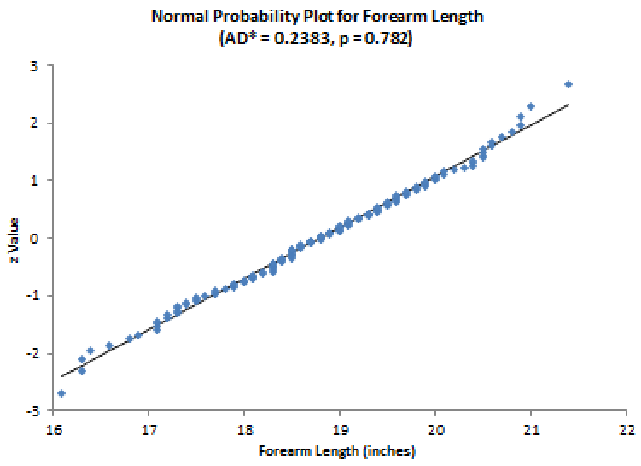
- Para las pruebas del QQ- Plot, se ordenan primeramente los residuos de menos a mayor:

$$\hat{\varepsilon}_{(1)} \leq \ldots \leq \hat{\varepsilon}_{(n)}$$

- Luego calculamos los cuantiles teóricos de una curva normal estándar:

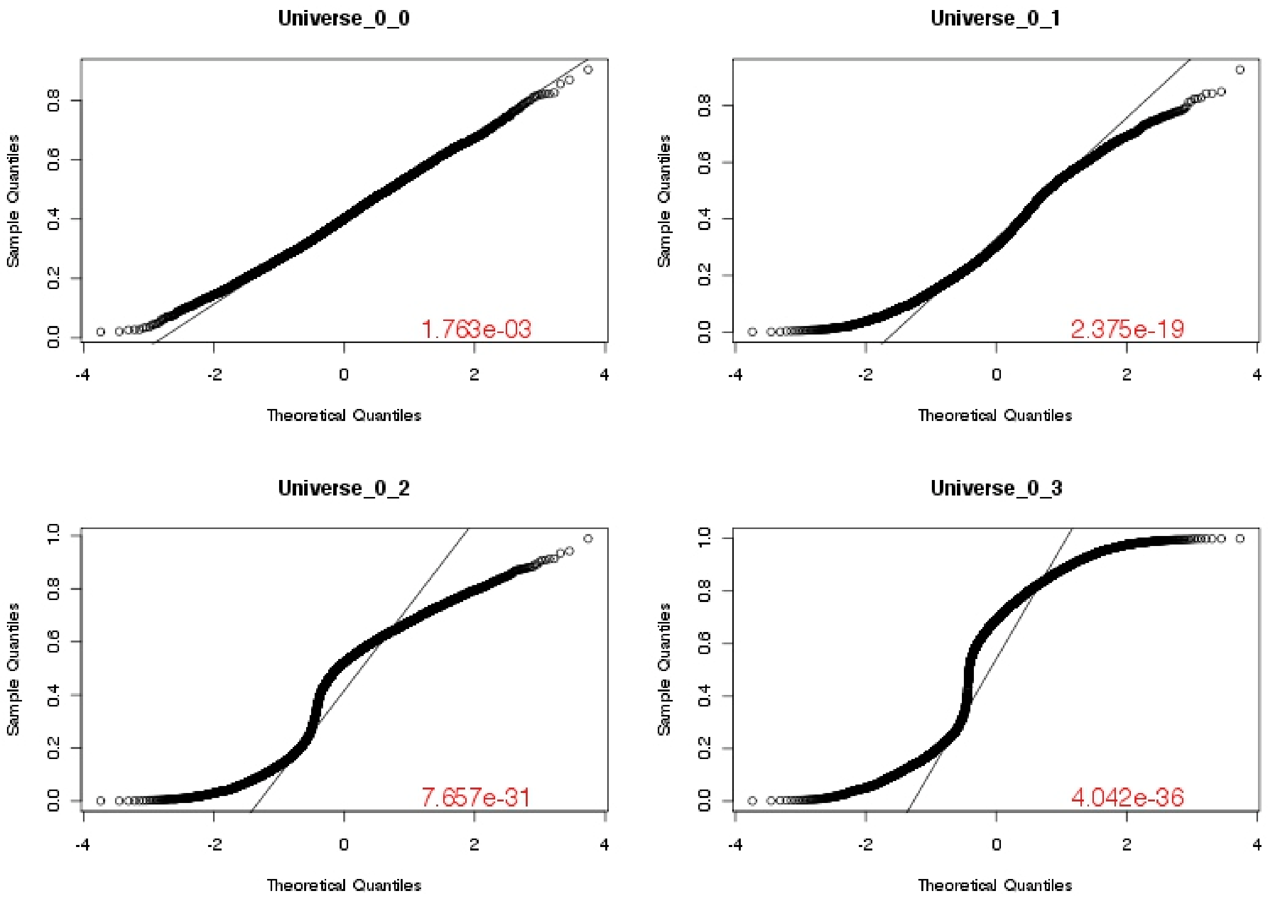
$$u_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$$

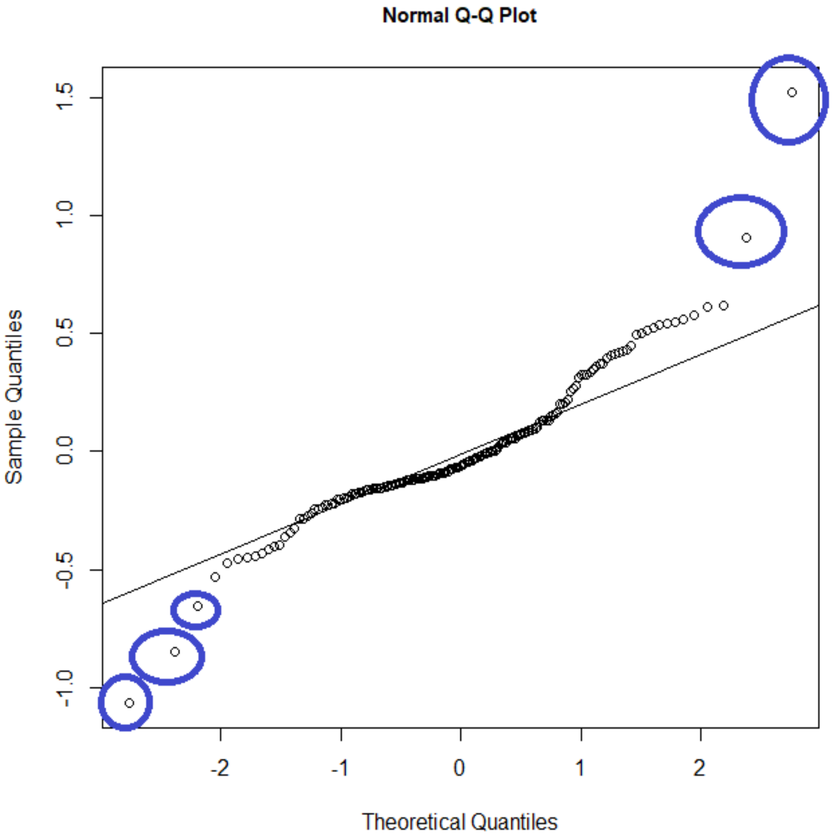
- Finalmente, se grafican los residuales ordenados contra los cuantiles. Si hay normalidad en los residuos, se esperaría una relación próxima entre estos dos.



Prueba de normalidad de los residuos

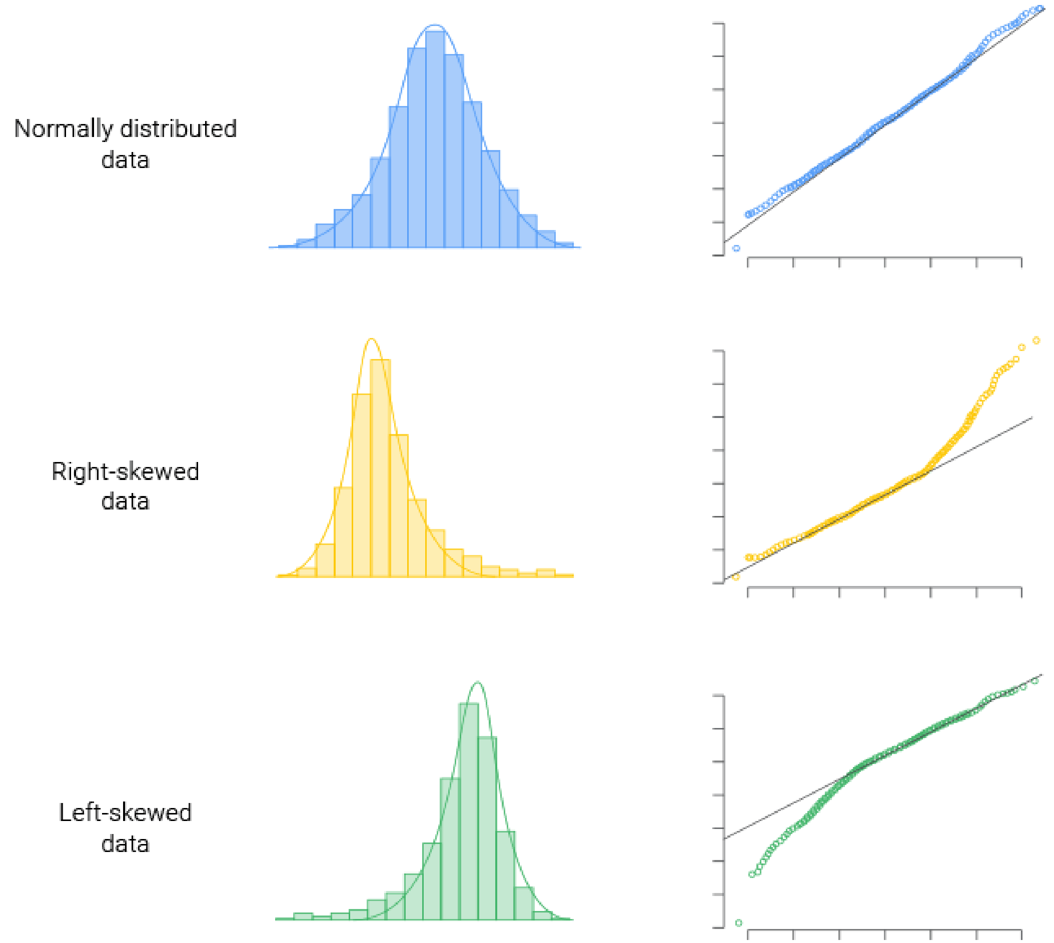
- Sin embargo, perfectamente podemos tener una clara desviación respeto a la normalidad…





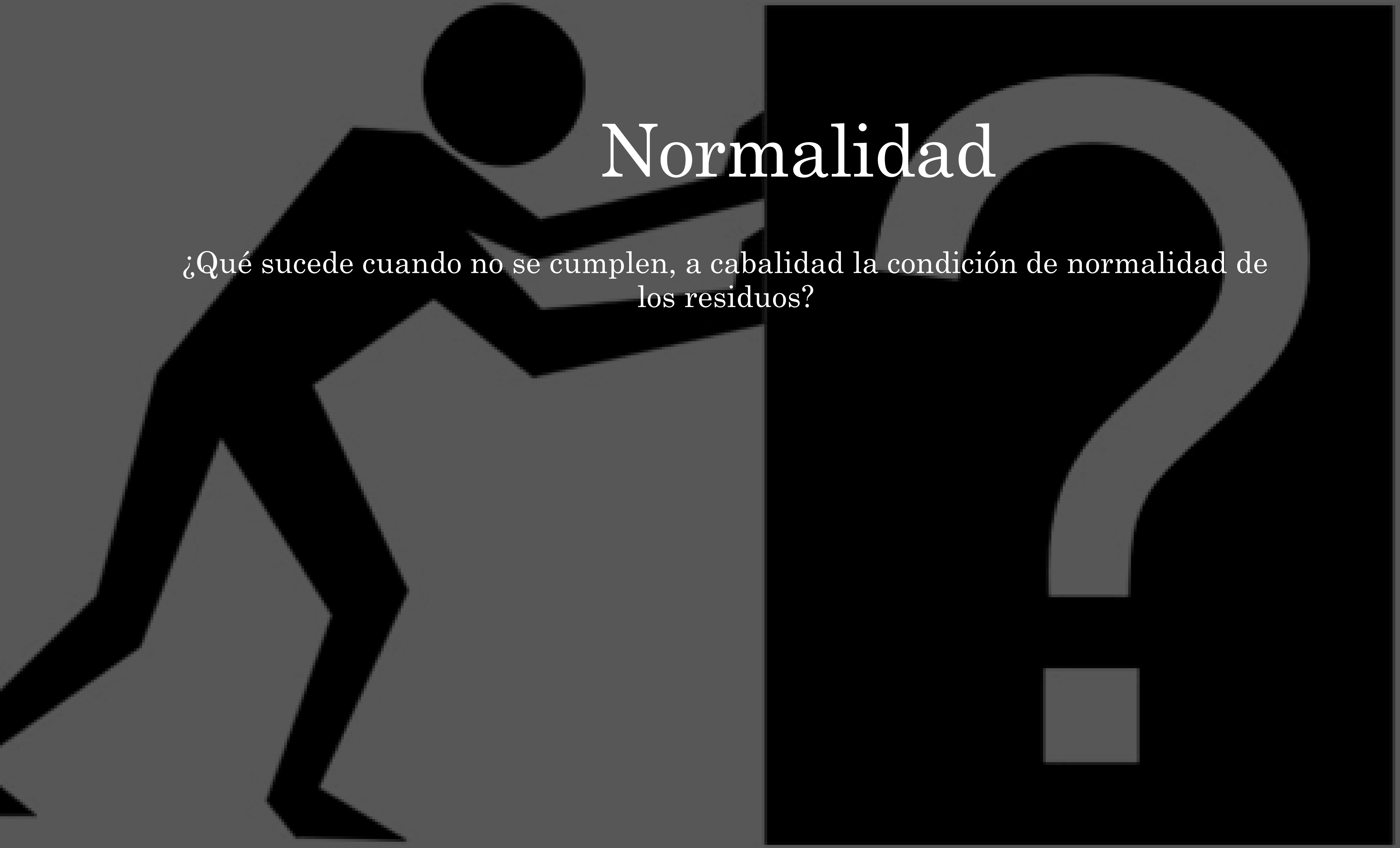
Prueba de normalidad de los residuos

- Sin embargo, perfectamente podemos tener una clara desviación respeto a la normalidad…



Normalidad

¿Qué sucede cuando no se cumplen, a cabalidad la condición de normalidad de los residuos?



Índice

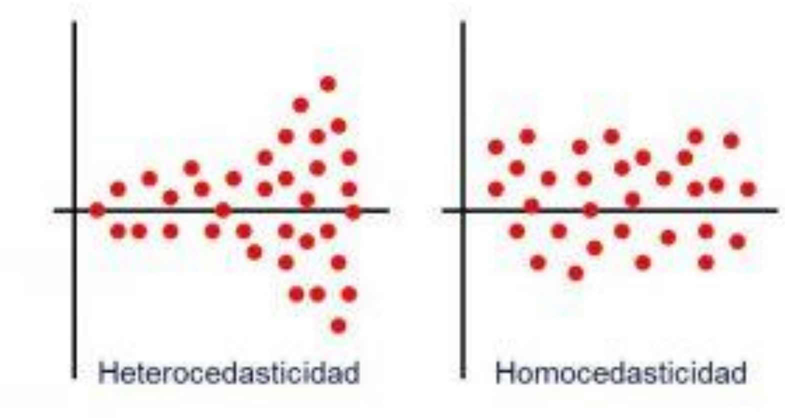
Variancia constante en los residuos: la homocedasticidad

- Al postular lo siguiente :

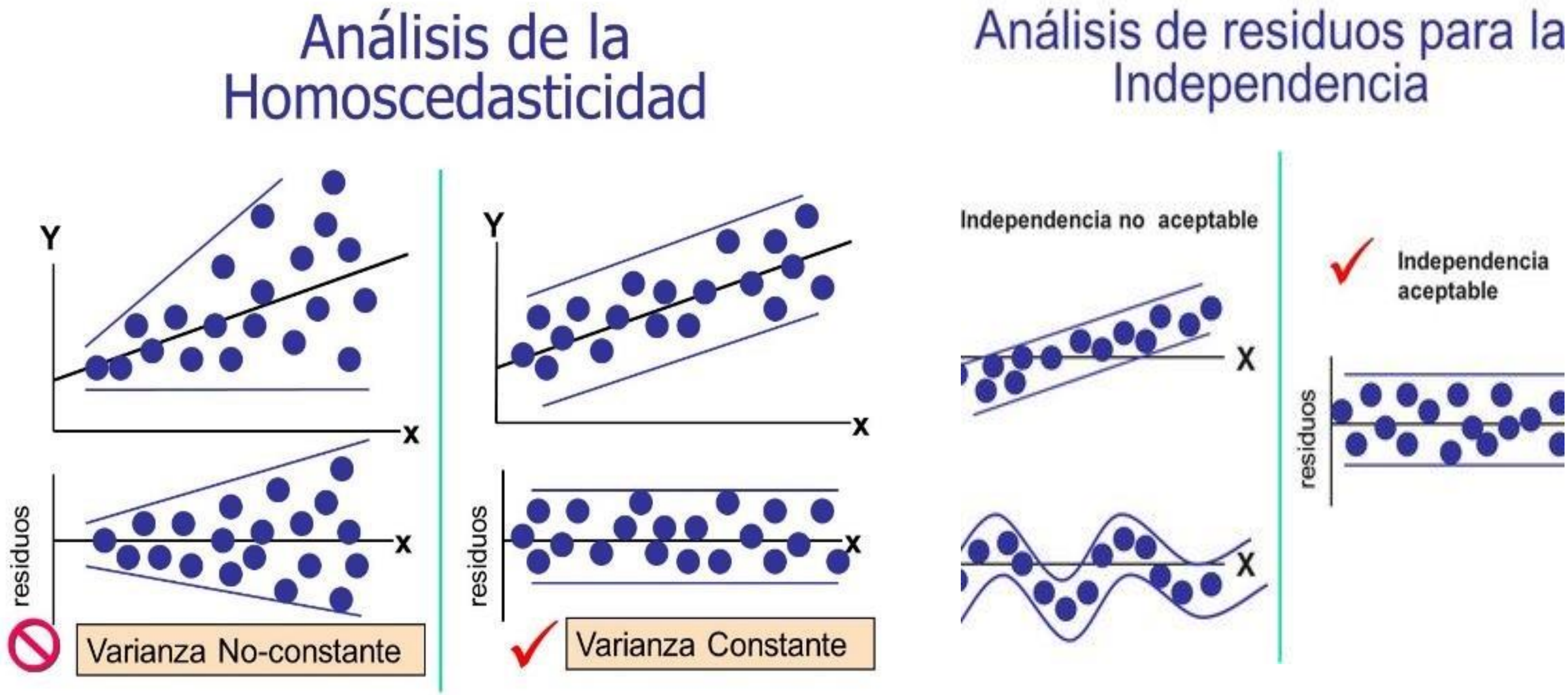
$$\boxed{\varepsilon \sim N(0, \sigma^2 I)}$$

nuestros residuos mantienen una estrucura constante… y si no, es un modelo heterocedastico o con heterocedasticidad.

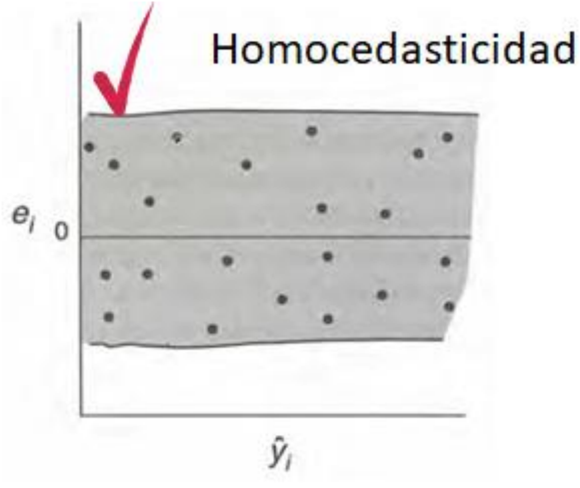
- La **homocedasticidad** es una característica de un modelo de regresión lineal implica que la varianza de los errores es constante a lo largo de su evaluación residual.
- En la **heterocedasticidad** la varianza de los errores de las variables explicativas <u>**no**</u> es constante a lo largo de todas las observaciones.

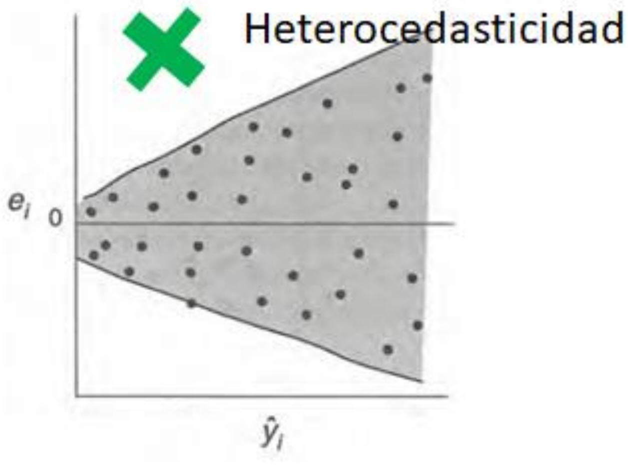


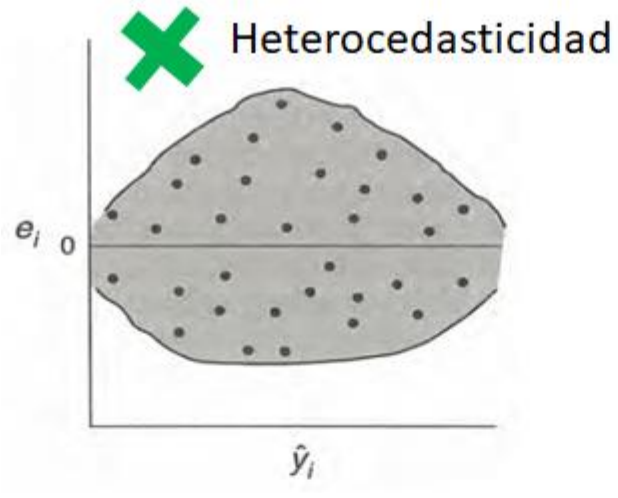
Variancia constante en los residuos: la homocedasticidad

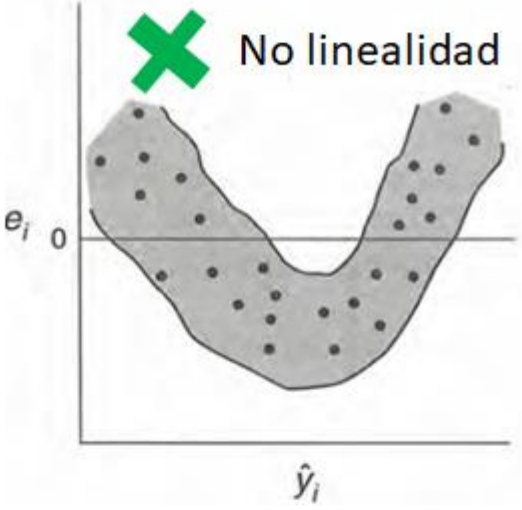


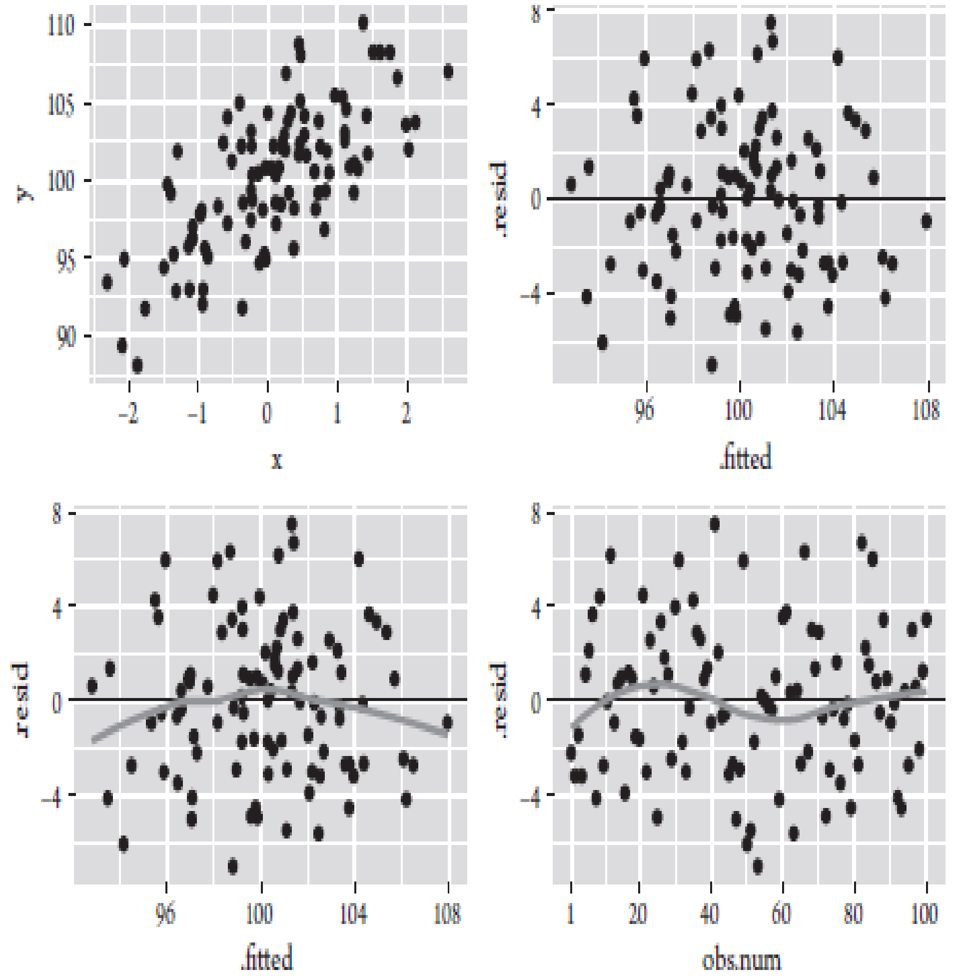
Variancia constante en los residuos: la homocedasticidad











Variancia constante en los residuos: la homocedasticidad

- En un modelo homocedástico, que es lo que se espera, los errores tiene una variancia constante para cada valor de X_i. Y si no es así, tendremos residuos no constantes, y pues, no es lo que se desea para la RLM (modelo heterosedástico).

- Hay diversas formas de verificar lo anterior: pruebas estadísticas y mediante el análisis de los residuos (si la variancia es constante, el gráfico debe aparecer como una secuencia aleatoria de puntos).

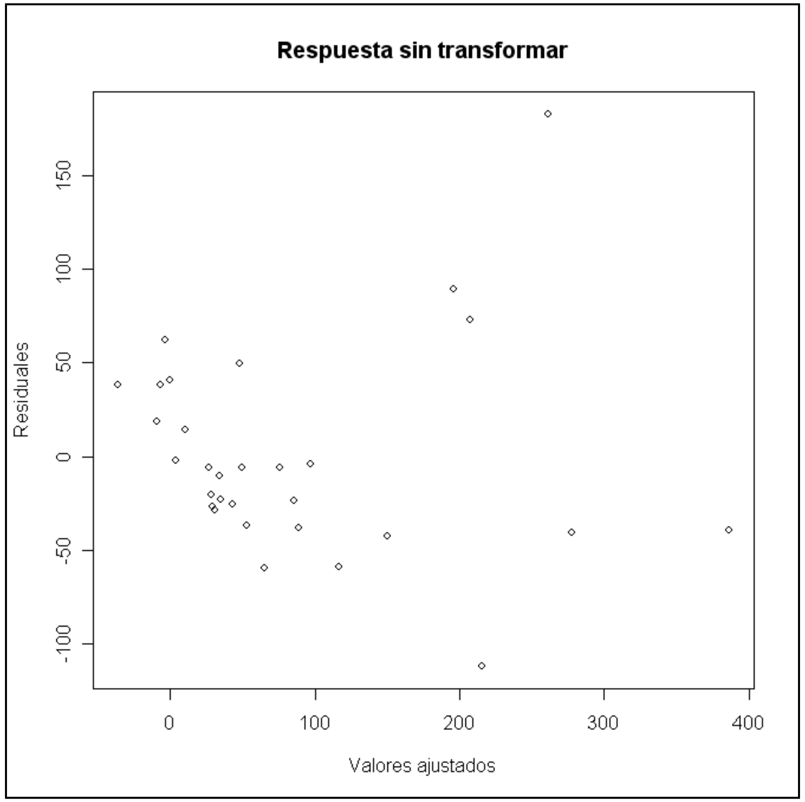
- Al igual que la vez pasada, se recomienda llevar a cabo ambas prueba.

- Se pueden usar los valores absolutos o raíz cuadrada de los residuales ya que el signo de los residuales no es tan importante. Estos valores se usan especialmente cuando no hay muchas observaciones.

Variancia constante en los residuos: la homocedasticidad

- Una prueba formal para determinar si existe evidencia de heteroscedasticidad es el test de Breush-Pagan. Asume que los errores son independientes y normales:

$$\chi^2 = \frac{SC\,\mathrm{Re}\,g}{2} \div \left(\frac{SCE}{n}\right)^2$$

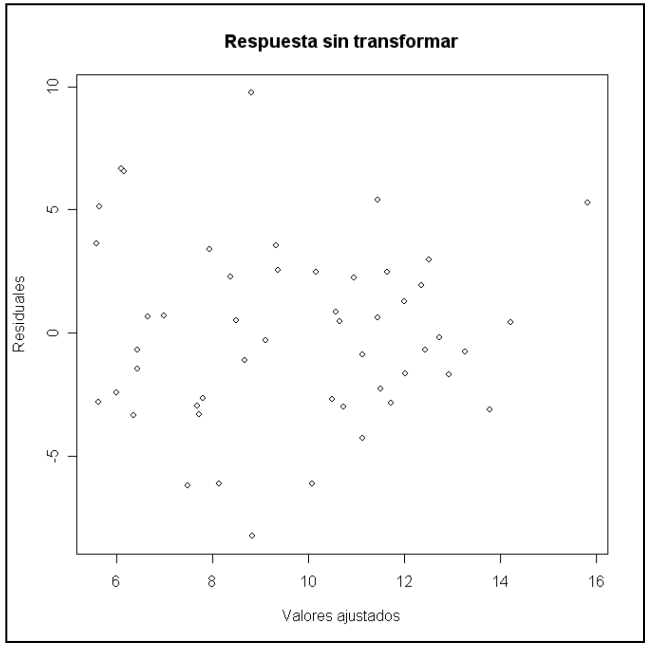


Non-constant Variance Score Test

Variance formula: ~ fitted.values

Chisquare = 15.07   Df = 1   p = 0.0001

Evidencia de variancia no constante



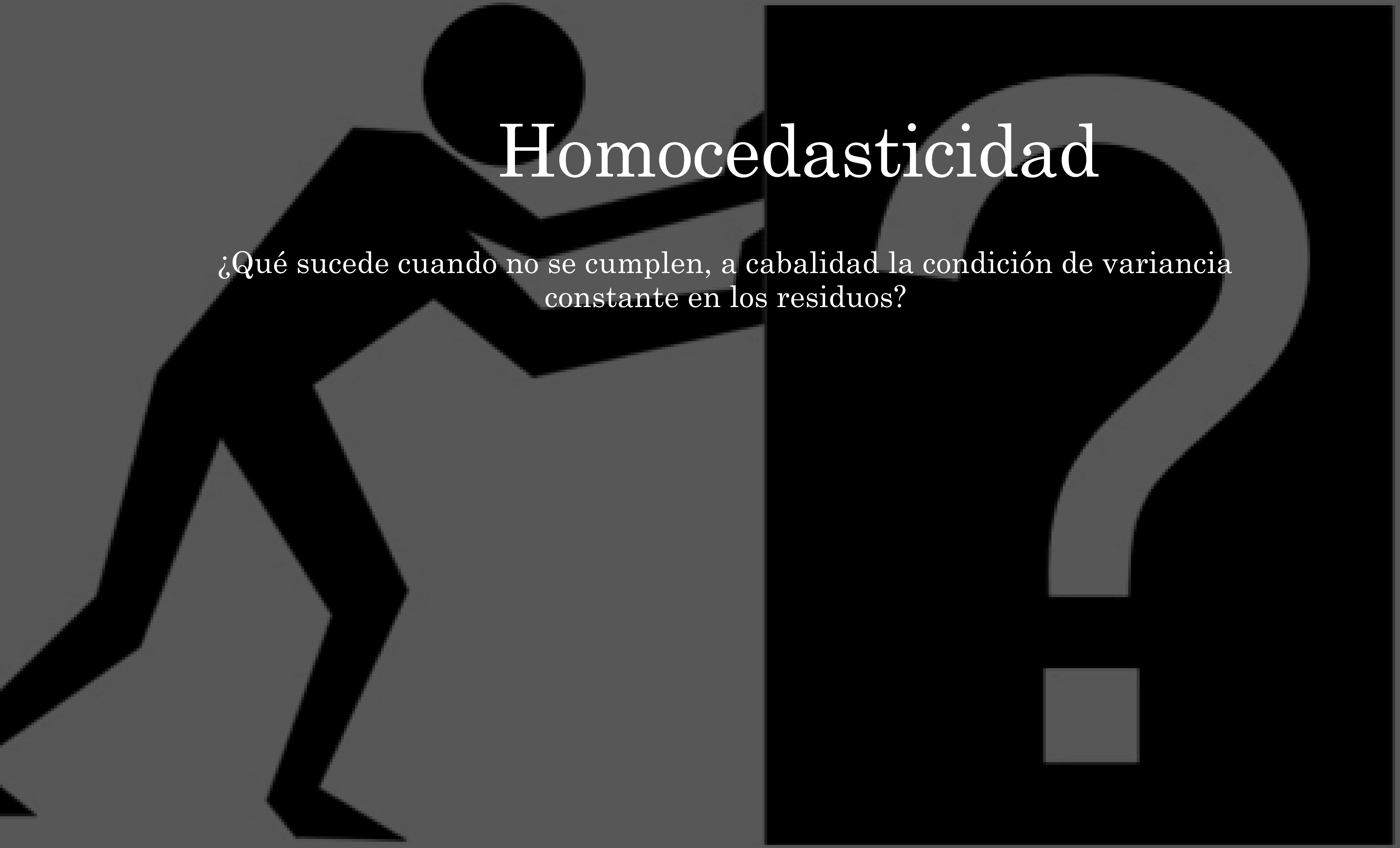
Non-constant Variance Score Test

Variance formula: ~ fitted.values

Chisquare = 2.27   Df = 1   p = 0.13

Homocedasticidad

¿Qué sucede cuando no se cumplen, a cabalidad la condición de variancia constante en los residuos?



Índice

La multicolinealidad

- Al realizar una RLM, nos gustaría, dentro de un mundo ideal, que nuestras variables independientes tengan independencia u ortogonalidad entre ellas, esto quiere decir, que la correlación sea lo más pequeña o nula entre pares.

- ¿Algún ejemplo para comprender esto?

- Al querer verificar si hay presencia de multicolinealidad en la RLM, debemos verificar si esta está o no presente, dado que podría sesgar fuertemente los resultados de los coeficientes (valor estimado, su error estándar, etc.).

- Definimos la multicolinealidad como : "*La **multicolinealidad** es la relación de dependencia lineal fuerte entre más de dos variables explicativas en una regresión múltiple que incumple el supuesto de Gauss-Markov cuando es exacta. En otras palabras, la **multicolinealidad** es la correlación alta entre más de dos variables explicativas*."

La multicolinealidad

- La multicolinealidad es una situación en la que se presenta una fuerte correlación entre los predictores del modelo:

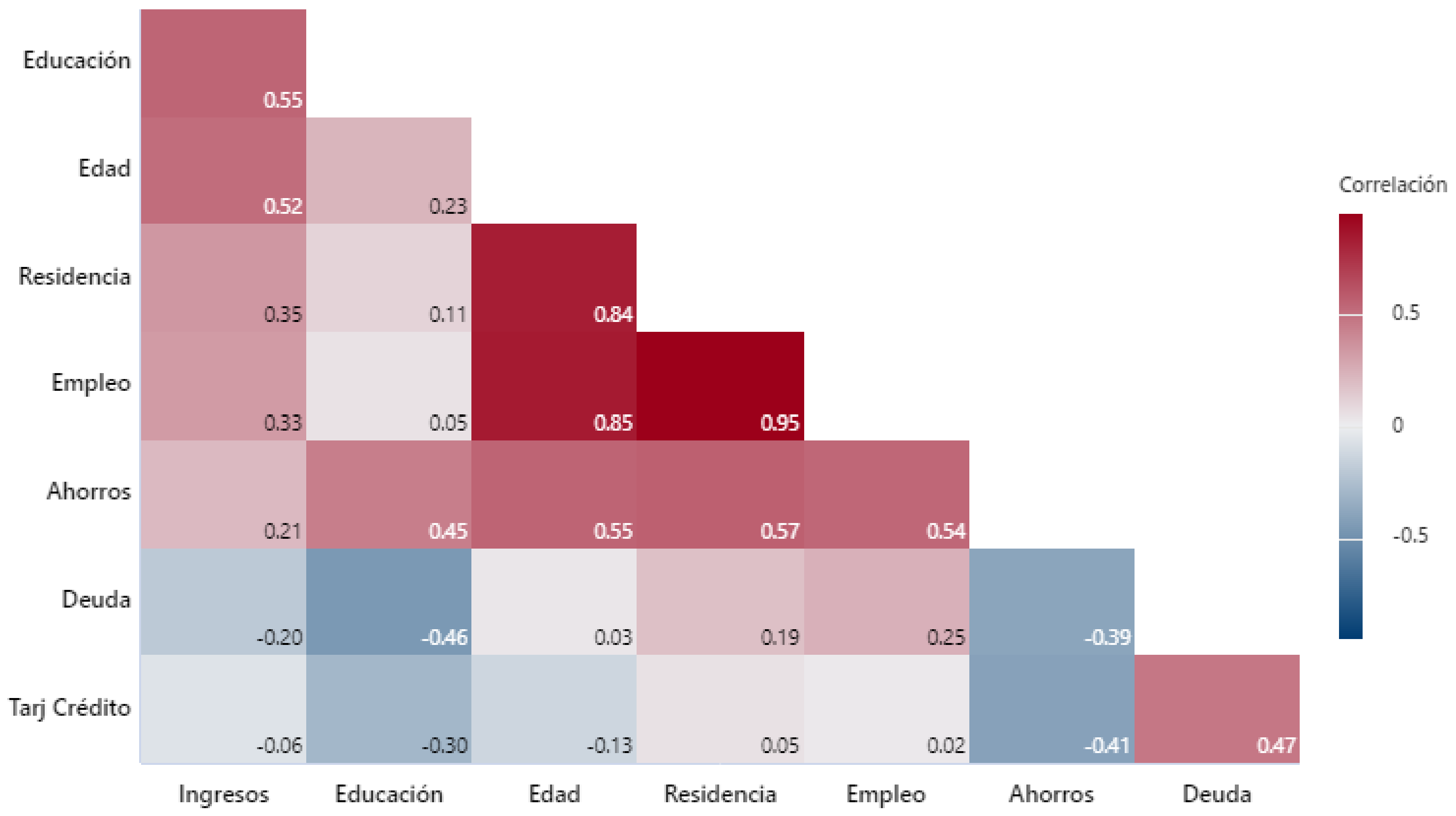
$$Y \sim X$$

$$\begin{bmatrix} Y1 \\ Y2 \\ \cdots \\ Y_n \end{bmatrix} \sim \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

- La matriz de las X, al posser relación entre sus columnas, produce que en el proceso de invisibilidad sea un contraste que produce a posteriori el sesgamiento en la RLM.
- Con alta multicolinealidad, pero no perfecta, los coeficientes poseen grandes errores estándar, lo que hace que los coeficientes no puedan ser estimados con gran precisión.
- Otro efecto, es que se pueden tener estimaciones sensibles a la muestra, por lo que pequeños cambios en los valores de Y o de X pueden dar lugar a cambios importantes en las estimaciones.
- Ahora, la pregunta sería: ¿cómo podemos verificar la multicolinealidad?

La multicolinealidad

- Considero que la primera idea de si nos toparemos o no con la multicolinealidad, es mediante el uso del correlograma, y verificar la posibilidad en encontrarnos con altos niveles de correlación.



La multicolinealidad

- Lo siguiente es ir, introduciendo, un coeficiente, uno a uno, y ver si hay o no cambio en los valores de los Betas, y errores estándares asociados bastante altos.

Variables incluidas	β_1	β_2
Triceps	0,86	---
Thigh	---	0,86
Triceps y Thigh	0,22	0,66
Triceps, Thigh y Midarm	4,33	-2,86

MUY CAMBIANTES

NO SIGNIFICATIVOS

EFECTO OPUESTO AL ESPERADO

Correlación entre Triceps y Thigh = $\boxed{0.92}$

CORRELACION MUY ALTA

$SCRes(X_1) = 352.3$

$SCRes(X_1/X_2) = 3.47$

$SCRes(X_2) = 381.97$

$SCRes(X_2/X_1)) = 33.17$

REDUCCION MUY PEQUEÑA

La multicolinealidad

- La última forma es mediante el uso del indicador VIF o el factor de inflación de la variancia. Este mide la cantidad en la que la variancia de un coeficiente se ve inflada en relación a una situación donde los predictores no estén correlacionados.

- Su fórmula es la siguiente:

$$\hat{V}(\hat{\beta}_i) = \frac{\sigma^2}{(n-1)s_i^2} \times \boxed{\frac{1}{1-R_j^2}}$$

Correlación múltiple de la regresión de X_j con los otros predictores

FACTOR

- Según "teoría", se dice que valores de VIF mayor a 10 se considera como un indicador de que hay multicolinealidad que está influenciando los estimadores de mínimos cuadrados. La raíz cuadrada del factor indica cuánto se expande el intervalo de confianza para un determinado coeficiente relativo a uno similar con datos no correlacionados.

Predictor	Coefficient	*t*-value	*p*-value	VIF
Age	0.079	14.324	<0.001**	2.122
REL	0.508	9.841	<0.001**	2.107
NFM	−0.021	−1.451	0.147	1.137
WI	0.695	24.110	<0.001**	1.454
TEBC	−0.159	−6.111	<0.001**	2.232
AAFB	0.158	6.287	<0.001**	2.384
AAFM	0.059	3.161	0.002**	2.380
HEL	0.071	7.906	<0.001**	1.862

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.155	0.132	1.18	0.243	
%Fat	0.00557	0.00409	1.36	0.176	14.93
Weight kg	0.01447	0.00285	5.07	0.000	33.95
Activity	0.000022	0.000007	3.08	0.003	1.05
%Fat*Weight kg	-0.000214	0.000074	-2.90	0.005	75.06

Multicolinealidad

¿Qué sucede cuando no se cumplen, a cabalidad la condición de independencia entre los predictores o variables independiente?

Índice

Linealidad entre la Y y la(s) X_i

- Antes, recordemos que a la RLM, la llamamos ***Regresión Lineal Múltiple***, y tenemos el término de "***Lineal***", lo cual nos indica que debería haber una linealidad entre la variable dependiente, Y, y los predictores, X.

- Y deberíamos hacernos la gran pregunta: ¿son todas las relaciones asimétricas lineales ? Y por supuesto, la respuesta es un rotundo **NO**.

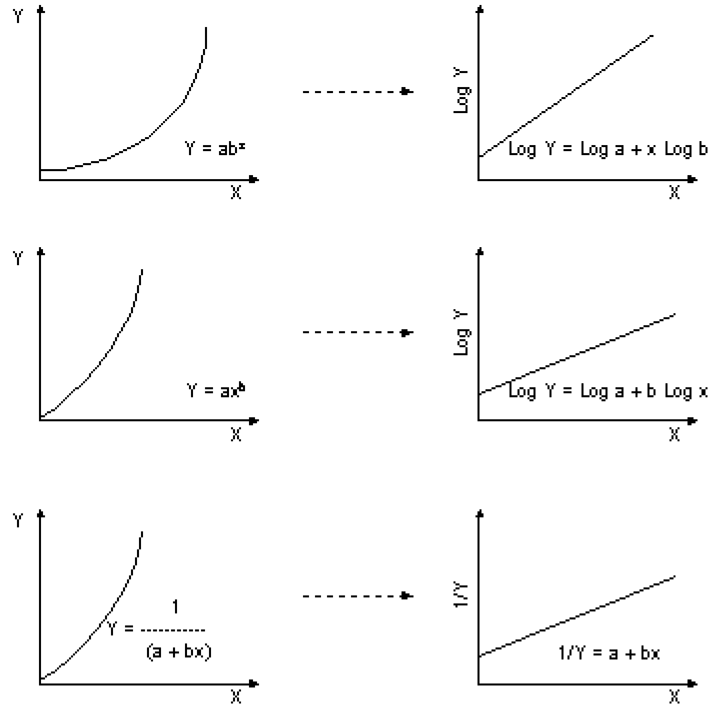
- ¿ Cómo podríamos corroborar entonces la linealidad en una RLM?

- Para determinar si la función de regresión lineal es apropiada se usan los gráficos de los residuales contra los predictores, o equivalentemente, contra los valores ajustados.

- A mí me gusta graficar la variable Y, y los predictores, X, también… pero lo anterior nos ayuda a determinar el tipo de transformación a utilizar.

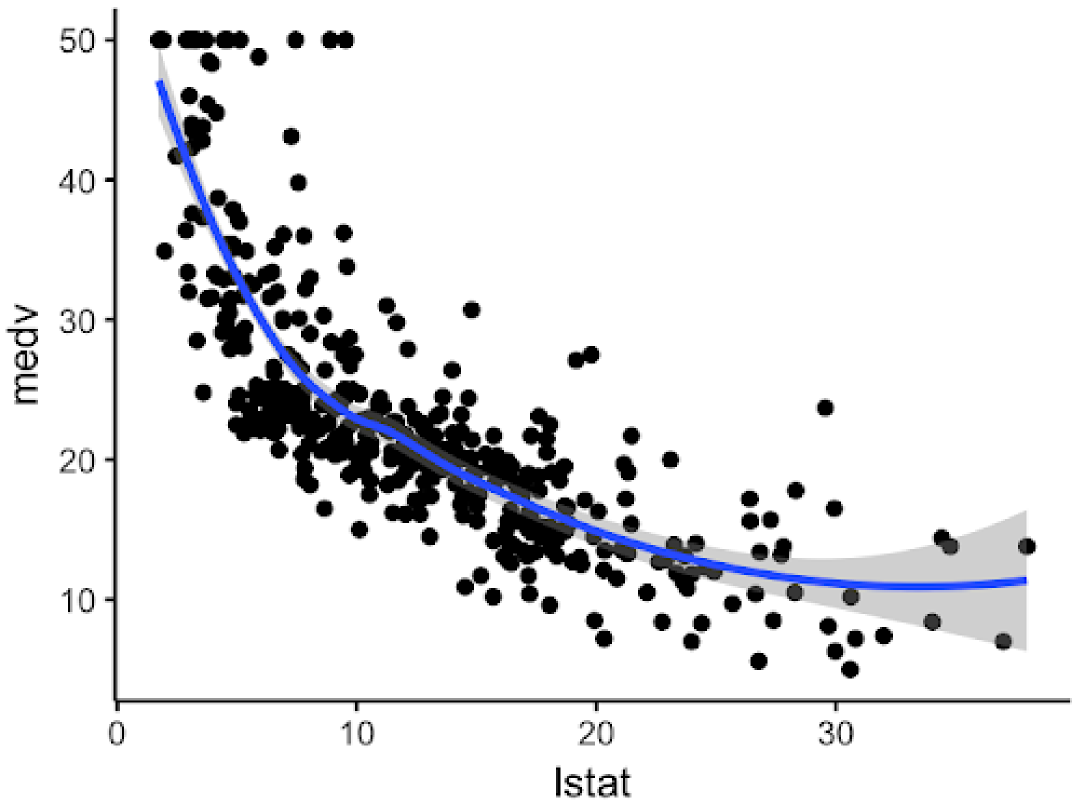
Linealidad entre la Y y la(s) X_i

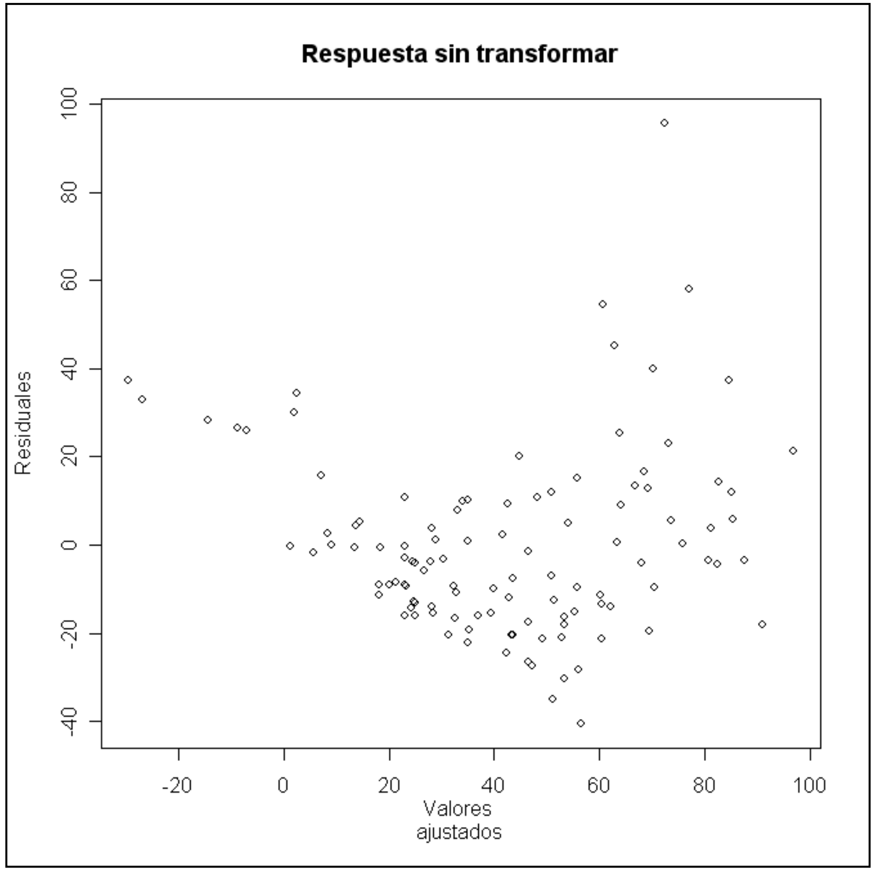
- ¿Es el fin del mundo si no poseemos la linealidad ? Pues no…
- Realmente este es el caso menos problemático de todo, dado que podemos optar por transformar las variables, o podemos optar por otro tipo de método de estimación.
- Lo que queremos, en una aplicación de RLM, es linealizar la respuesta (eso lo veremos en el siguiente capítulo):



Linealidad entre la Y y la(s) X_i

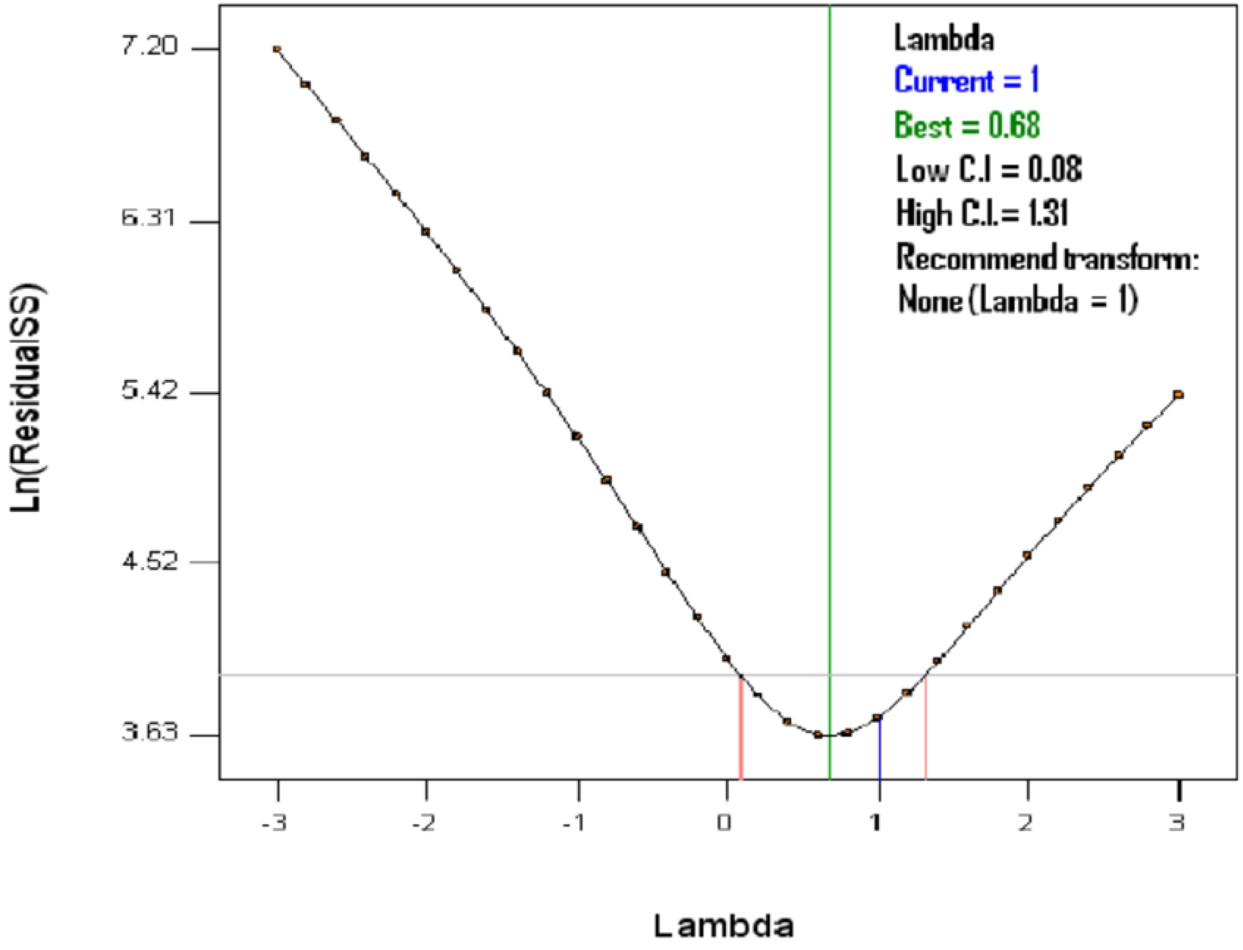
- Podemos verificar si existe o no linealidad prácticamente a de la visualización de los gráficos:





Linealidad entre la Y y la(s) X_i

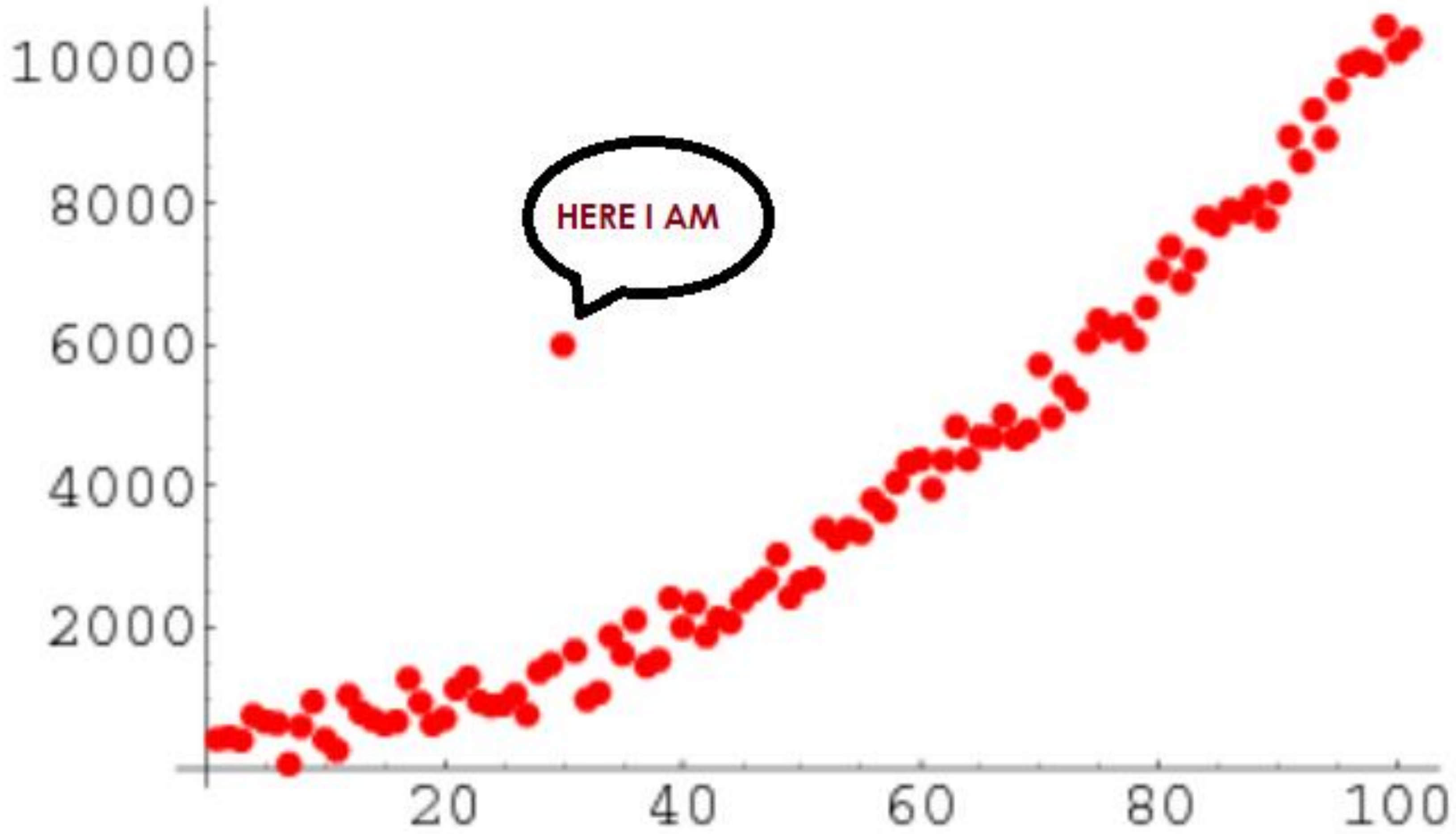
- También, podemos ver si se sugiere un tipo de transformación mediante el valor lambda de Cox, aunque analizaron esto en mayor detalle para el siguiente capítulo.



λ	Transformed Data
-2	y^{-2}
-1	y^{-1}
-0.5	$1/\sqrt{y}$
0	ln(y)
0.5	$\sqrt{y}$
1	y
2	y^2

Ausencia de linealidad entre Y - X

¿Qué sucede cuando no se cumplen, a cabalidad la condición de linealidad entre la variable dependiente y una o varias variables independientes?



Parte II – Valores extremos o de influencia

Índice

Valores extremos y de influencia

- ¿Por qué, en cursos de estadística básica, habíamos comentado la importancia de cuál medida de posición seleccionar, y el caso de las distribuciones simétricas y con asimetría ?

Valores extremos y de influencia

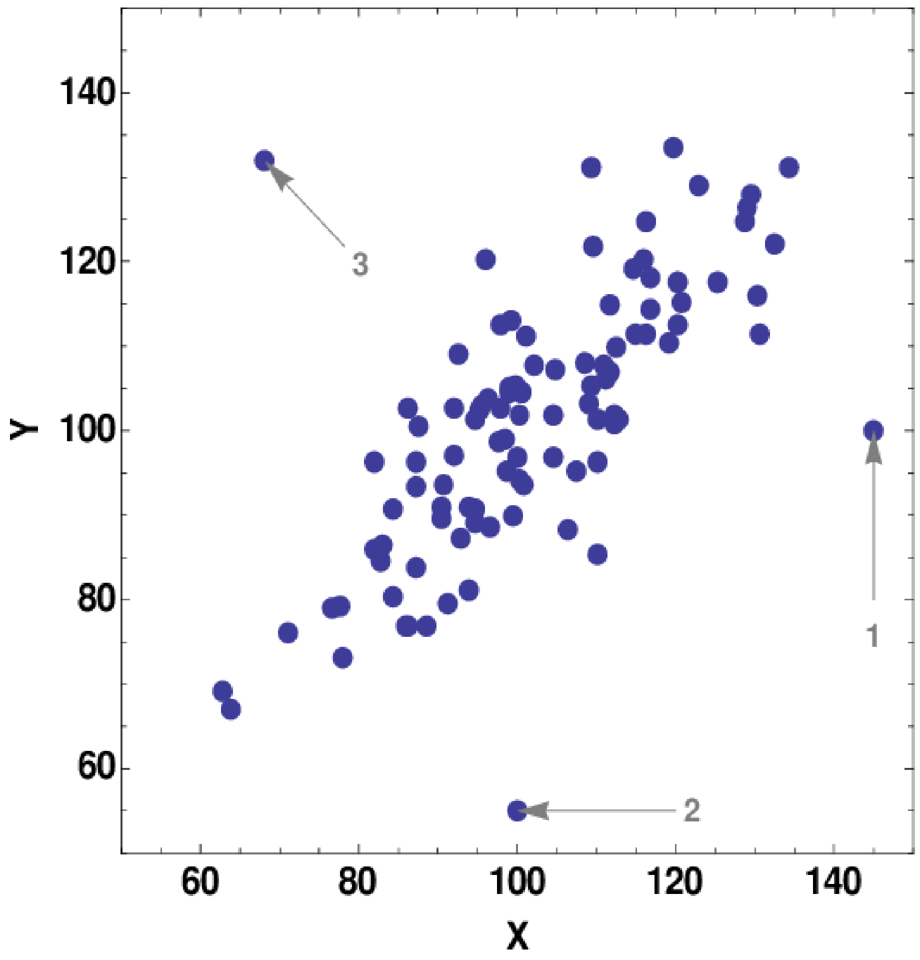
- La utiliza la RLM, realmente hacemos proyecciones o estimaciones a la media, o el promedio, lo cuál el análisis si hay o no presencia de valores extremos, es de suma importancia.
- Muchos autores de libros recomiendan comenzar por el análisis de los valores extremos, o los outliers. Y sin embargo, empiezan por por las propiedades o diagnóstico del modelo…
- Creo que todo se debe inspeccionar a la vez…
- Antes de proseguir, cómo definimos una valores extremo en el contexto de una regresión lineal múltiple:

“En el análisis de regresión, un valor atípico es una observación para la cual el residuo es de gran magnitud en comparación con otras observaciones en el conjunto de datos. La detección de valores atípicos y puntos influyentes es un paso importante del análisis de regresión.”.

“In regression analysis, an outlier is an observation for which the residual is large in magnitude compared to other observations in the data set. The detection of outliers and influential points is an important step of the regression analysis.”

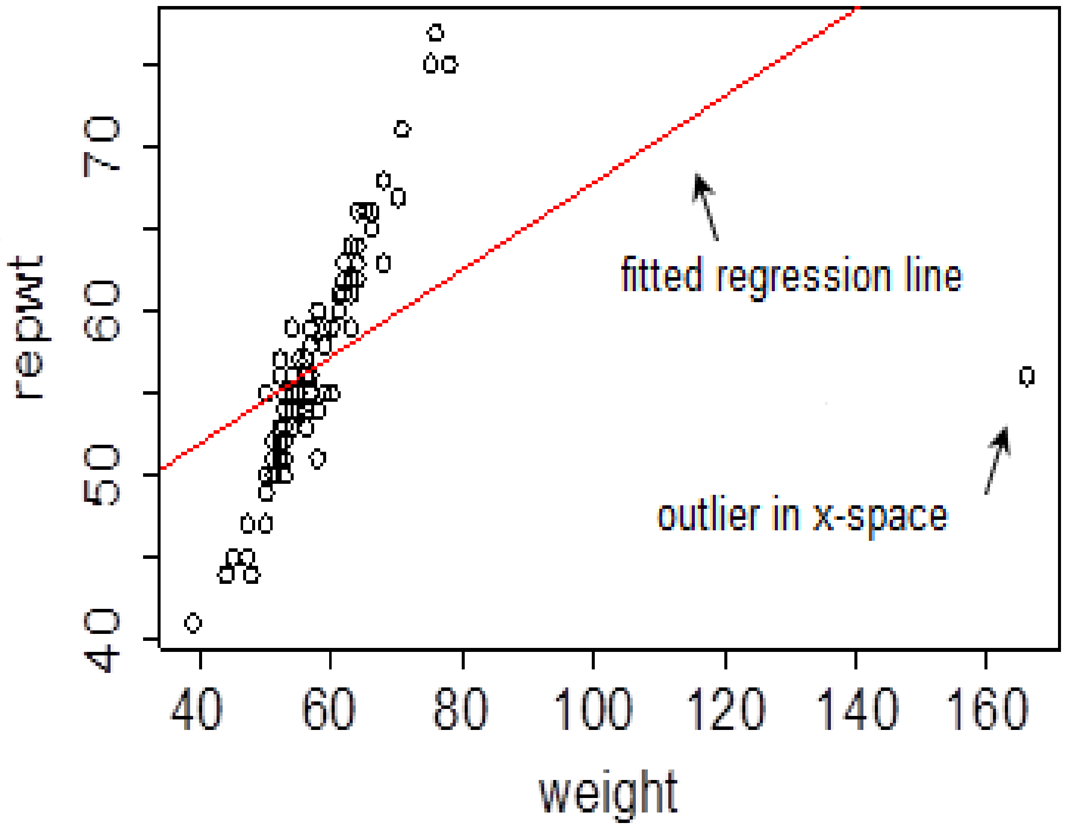
Valores extremos y de influencia

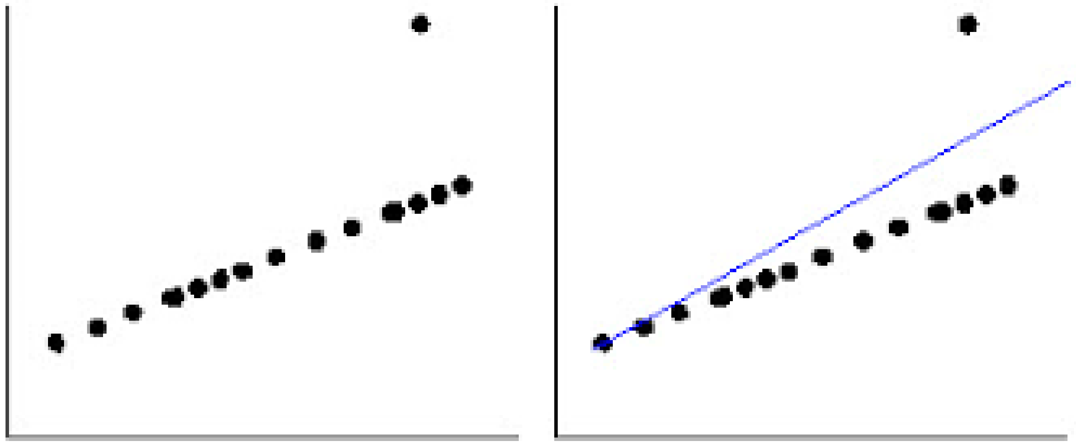
- Veamos casos de la presencia de valores extremos en una RLM:



Valores extremos y de influencia

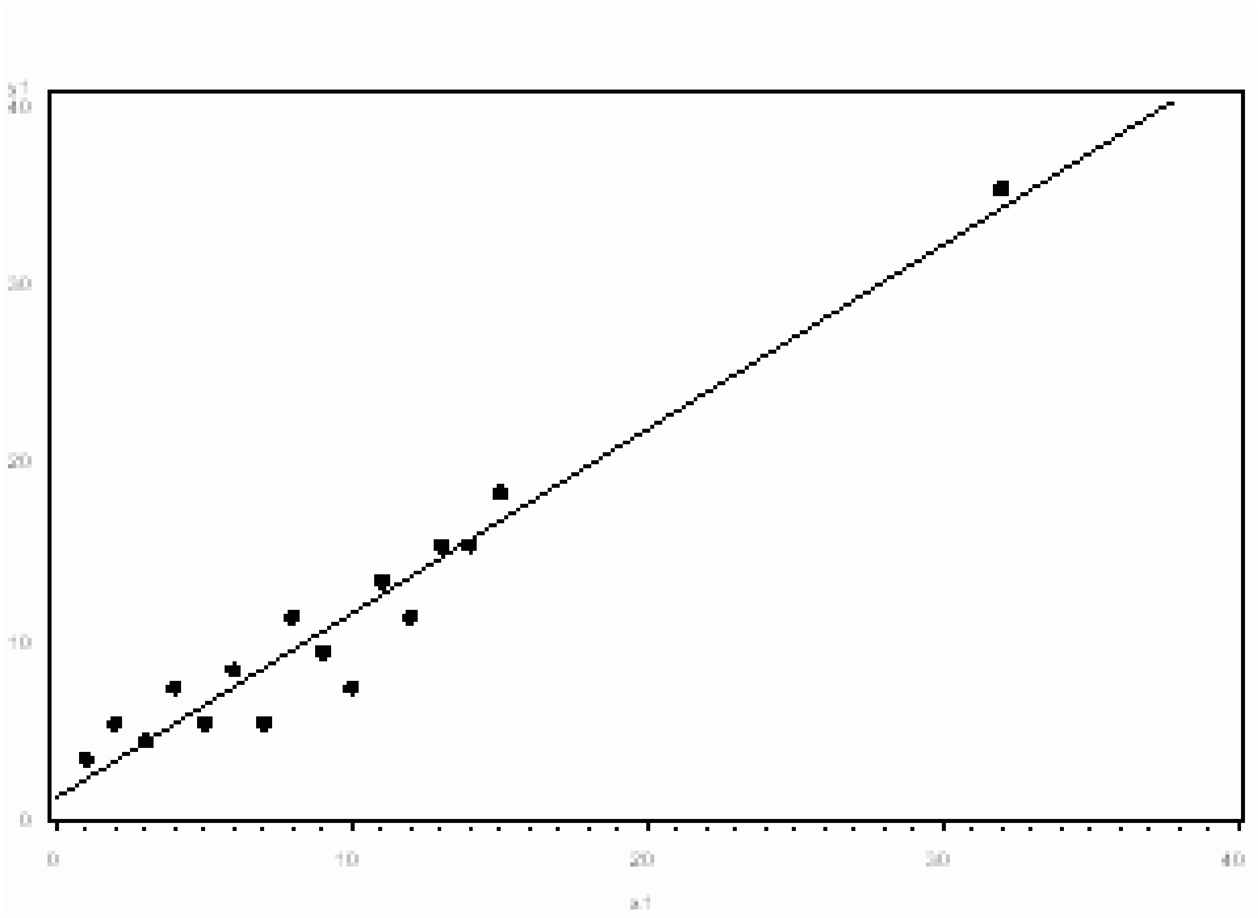
- ¿ Es relevante el efecto de un valor extremo en una RLM? El siguiente gráfico nos muestra que si…





Valores extremos y de influencia

- También puede darse el caso de que haya un valor extremo, pero este no sea un problema real para la estimación de la RLM:



Valores extremos y de influencia

- Para analizar la presencia de valores extremos o valores atípicos, se pueden analizar mediante gráficas los residuales ε_i, y así ver sus valores y / o magnitudes, o se pueden analizar también los residuales estudidentizados r_i:

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

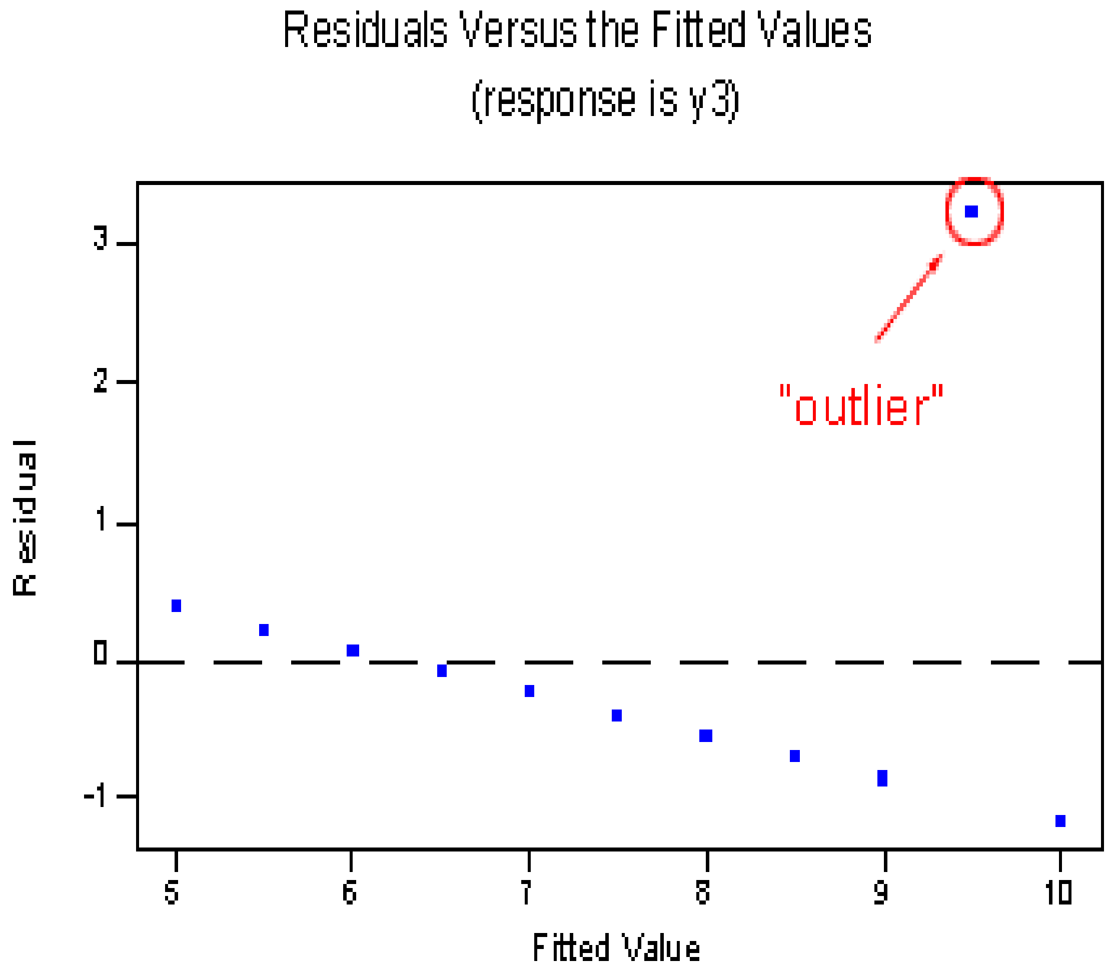
- Estos, al igual que los residuos ε_i, se pueden analizar en un gráfico de residuos.

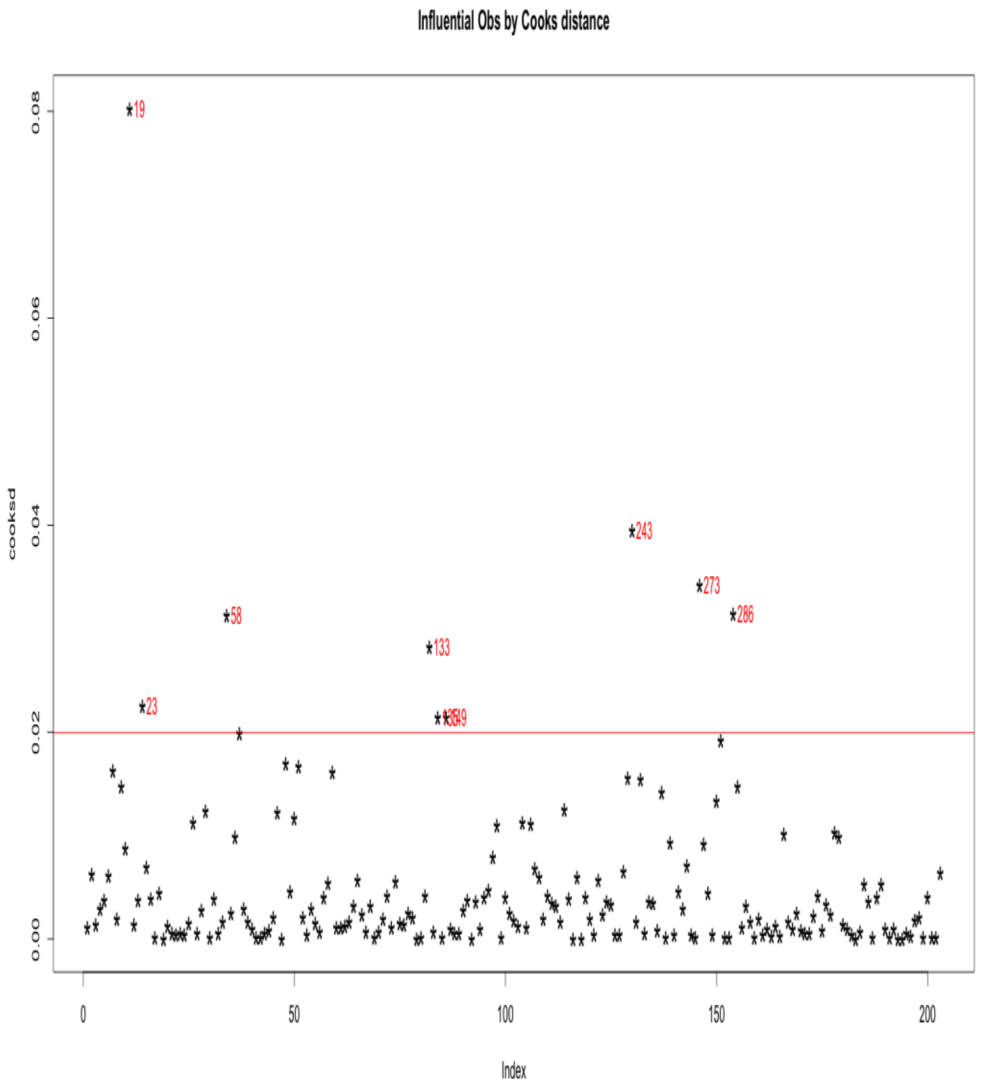
- Existen también los residuos estudidentizados externamente, los cuales se basan en los residuales obtenidos al ajustar la función de regresión eliminando el i-ésimo caso.

$$\left.\begin{aligned} t_i &= \frac{d_i}{E.E.(d_i)} \\ d_i &= y_i - \hat{y}_{i(i)} \end{aligned}\right\} \quad t_i = \frac{\varepsilon_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = \varepsilon_i \left[\frac{n-p-1}{SC\,\mathrm{Re}\,s(1-h_{ii}) - \varepsilon_i^2}\right]^{1/2}$$

- Los residuales estudentizados externamente se pueden utilizar para identificar valores de Y que se alejan del resto, y sobre todo ver el efecto en no tomar en cuenta cierto valores en el ajuste de la RLM. Se suele utilizar el criterio de Bonferroni de $t_{1-\alpha/2n,n-p-1}$

Valores extremos y de influencia





Valores extremos y de influencia

- Ahora, ¿qué es un caso de influencia?
- Según la siguiente definición:

"Un punto influyente es un valor extremo o atípico que afecta en gran medida la pendiente de la línea de regresión (el o los valores de los betas). Una forma de probar la influencia de un valor atípico es calcular la ecuación de regresión con y sin el valor atípico."

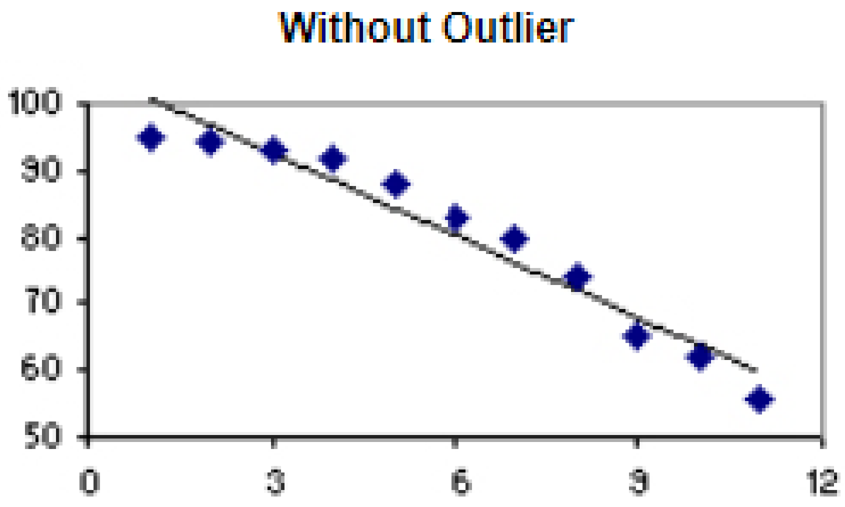
"An influential point is an outlier that greatly affects the slope of the regression line. One way to test the influence of an outlier is to compute the regression equation with and without the outlier."

- Entonces, "…diay Óscar, un punto de influencia es lo mismo que un punto de influencia ?..."

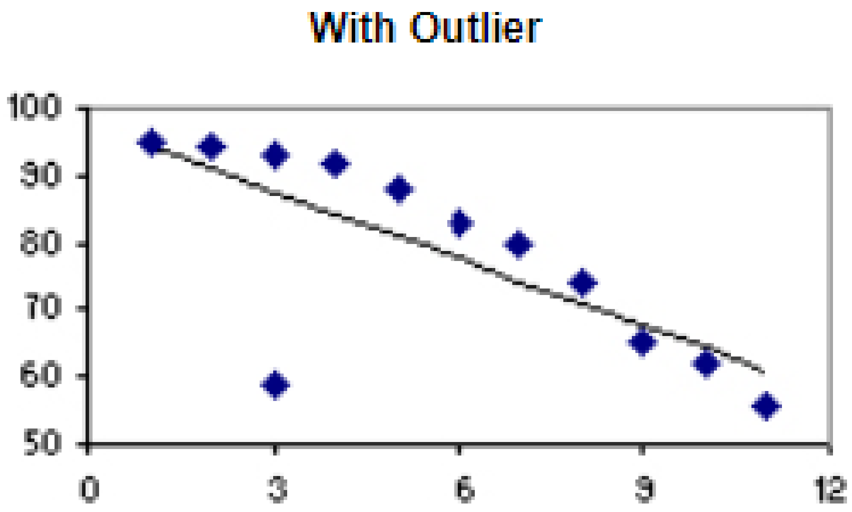
R/ "SI y NO: un punto de influencia, por definición, es un valor extremo, pero lo importante es el efecto que este puede llegar a tener en la estimación de la RLM".

Valores extremos y de influencia

- Las gráficas de dispersión son idénticas, excepto que una gráfica incluye un valor atípico. Cuando el valor atípico está presente, la pendiente es más plana (-4,10 frente a -3,32); por lo que este valor atípico se consideraría un punto influyente.



Regression equation: $\hat{y} = 104.78 - 4.10x$
Coefficient of determination: $R^2 = 0.94$

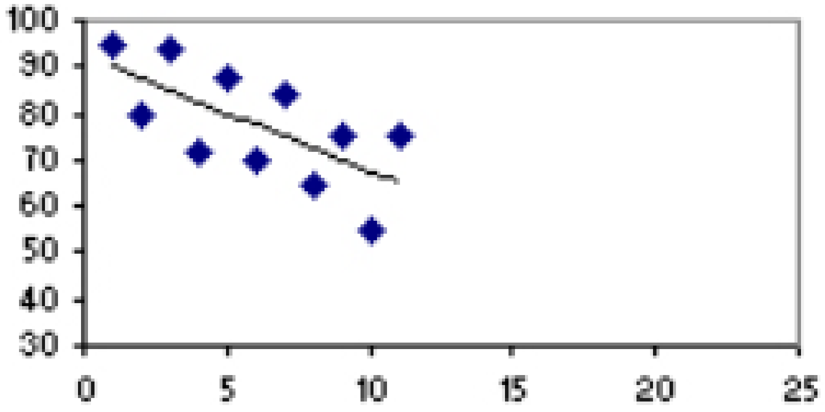


Regression equation: $\hat{y} = 97.51 - 3.32x$
Coefficient of determination: $R^2 = 0.55$

Valores extremos y de influencia

- Los gráficos a continuación comparan las estadísticas de regresión para otro conjunto de datos con y sin un valor atípico. Aquí, un gráfico tiene un único valor atípico, ubicado en el extremo superior del eje X (donde x = 24). Como resultado de ese único valor atípico, la pendiente de la línea de regresión cambia mucho, de -2,5 a -1,6; por lo que el valor atípico se consideraría un punto influyente.

Without Outlier

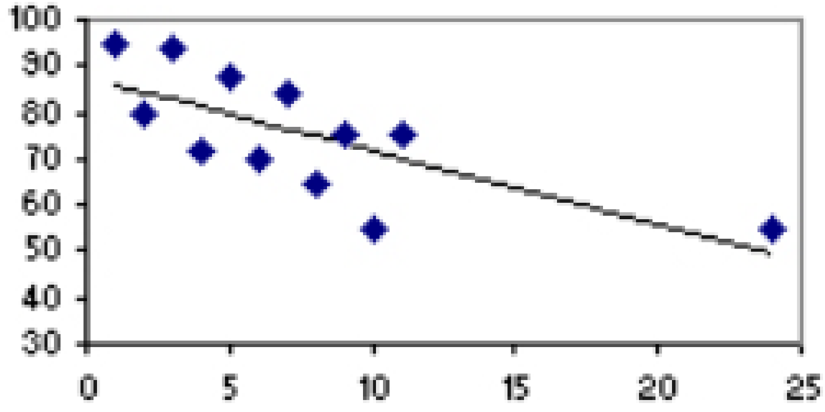


Regression equation: $\hat{y} = 92.54 - 2.5x$

Slope: $b_0 = -2.5$

Coefficient of determination: $R^2 = 0.46$

With Outlier



Regression equation: $\hat{y} = 87.59 - 1.6x$

Slope: $b_0 = -1.6$

Coefficient of determination: $R^2 = 0.52$

Valores extremos y de influencia

- A veces, un punto influyente hará que el coeficiente de determinación sea mayor; a veces, más pequeño. En el primer ejemplo anterior, el coeficiente de determinación es menor cuando el punto de influencia está presente (0,94 frente a 0,55). En el segundo ejemplo, es más grande (0,46 frente a 0,52).
- Si su conjunto de datos incluye un punto influyente, aquí hay algunas cosas a considerar.
  - Un punto influyente puede representar datos incorrectos, posiblemente el resultado de un error de medición. Si es posible, verifique la validez del punto de datos.
  - Compare las decisiones que se tomarían en función de las ecuaciones de regresión definidas con y sin el punto de influencia. Si las ecuaciones conducen a decisiones contrarias, tenga cuidado.
- Ahora la pregunta es, ¿cómo podemos detectar los puntos de influencia ?
- Cierto es que lo que vimos antes de valores extremos nos sirve, pero hay otros métodos…

Valores extremos y de influencia

- Las tres principales herramientas para detectar valores de influencia son:
  1. Sobre un solo valor ajustado (DFFITS).
  2. Sobre todos los valores ajustados (Distancia de Cook).
  3. Sobre los coeficientes (DFBETAS).

- Un caso se puede considerar de influencia si el valor absoluto de DFFITS excede 1 para conjuntos de datos pequeños a medianos, o excede $2\sqrt{p/n}$ para conjuntos de datos grandes:

$$(DFFITS)_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}} = t_i\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

- La distancia de Cook considera la influencia del i-ésimo caso sobre todos los n valores ajustados. Es una medida de influencia agregada

$$D_i = \frac{\varepsilon_i^2}{p\hat{\sigma}^2}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

Valores extremos y de influencia

- Se considera como medida de influencia la diferencia estandarizada entre el coeficiente de regresión b_k estimado basado en todos los n casos y el coeficiente obtenido omitiendo el i-ésimo caso ($\beta_{k(i)}$).

$$(DFBETAS)_{k(i)} = \frac{\beta_k - \beta_{k(i)}}{\hat{\sigma}_{(i)}\sqrt{c_{kk}}}$$

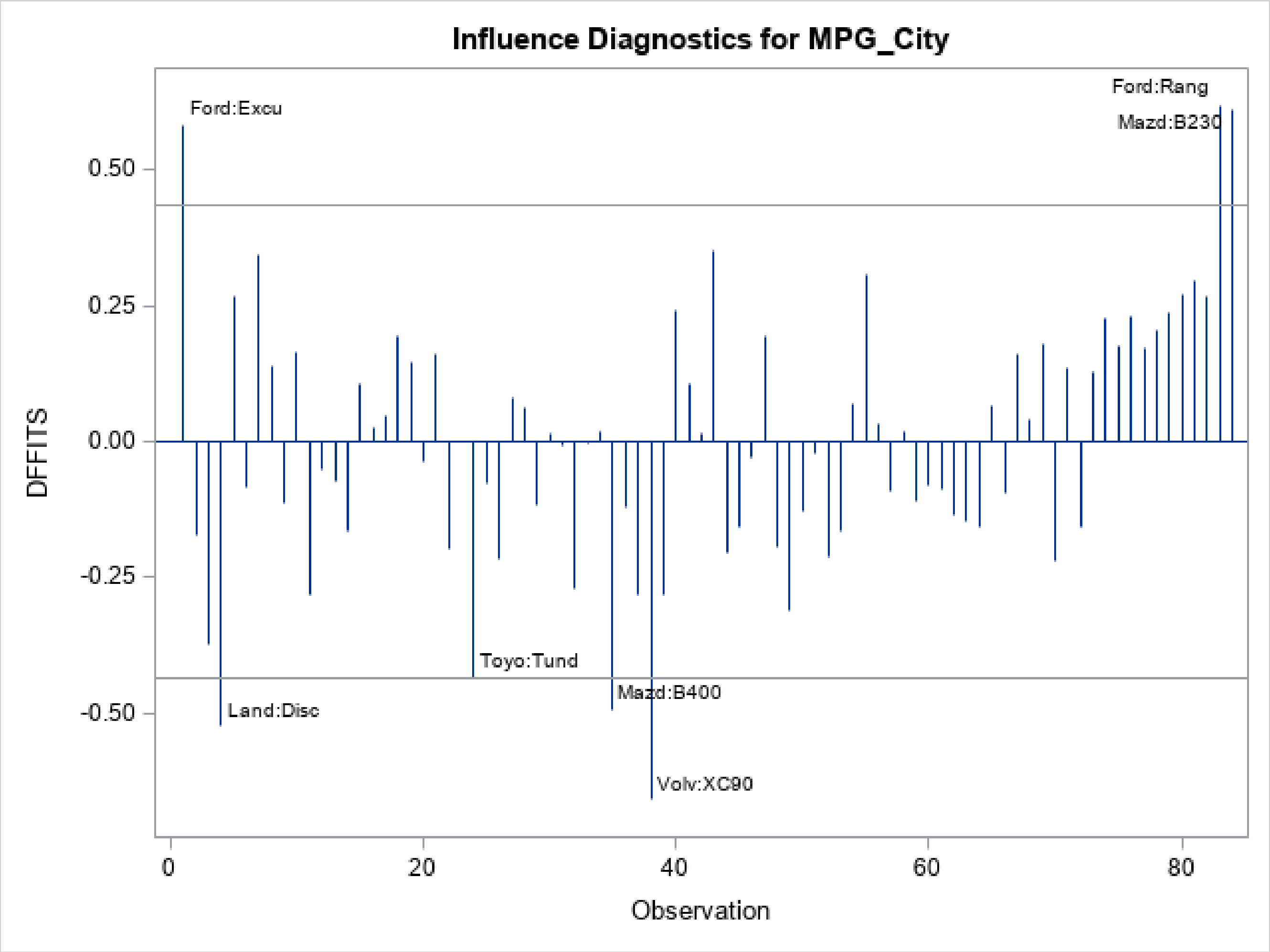
c_{kk} es el k-ésimo elemento de la diagonal de $(X^T X)^{-1}$

- Un valor absoluto de $(DFBETAS)_{k(i)}$ grande indica un gran impacto del caso i-ésimo en el k-ésimo coeficiente de regresión.

- Se considera grande si el valor absoluto es mayor a 1 en conjuntos de datos pequeños o medianos, o mayor a $2/\sqrt{n}$ para conjuntos de datos grandes.

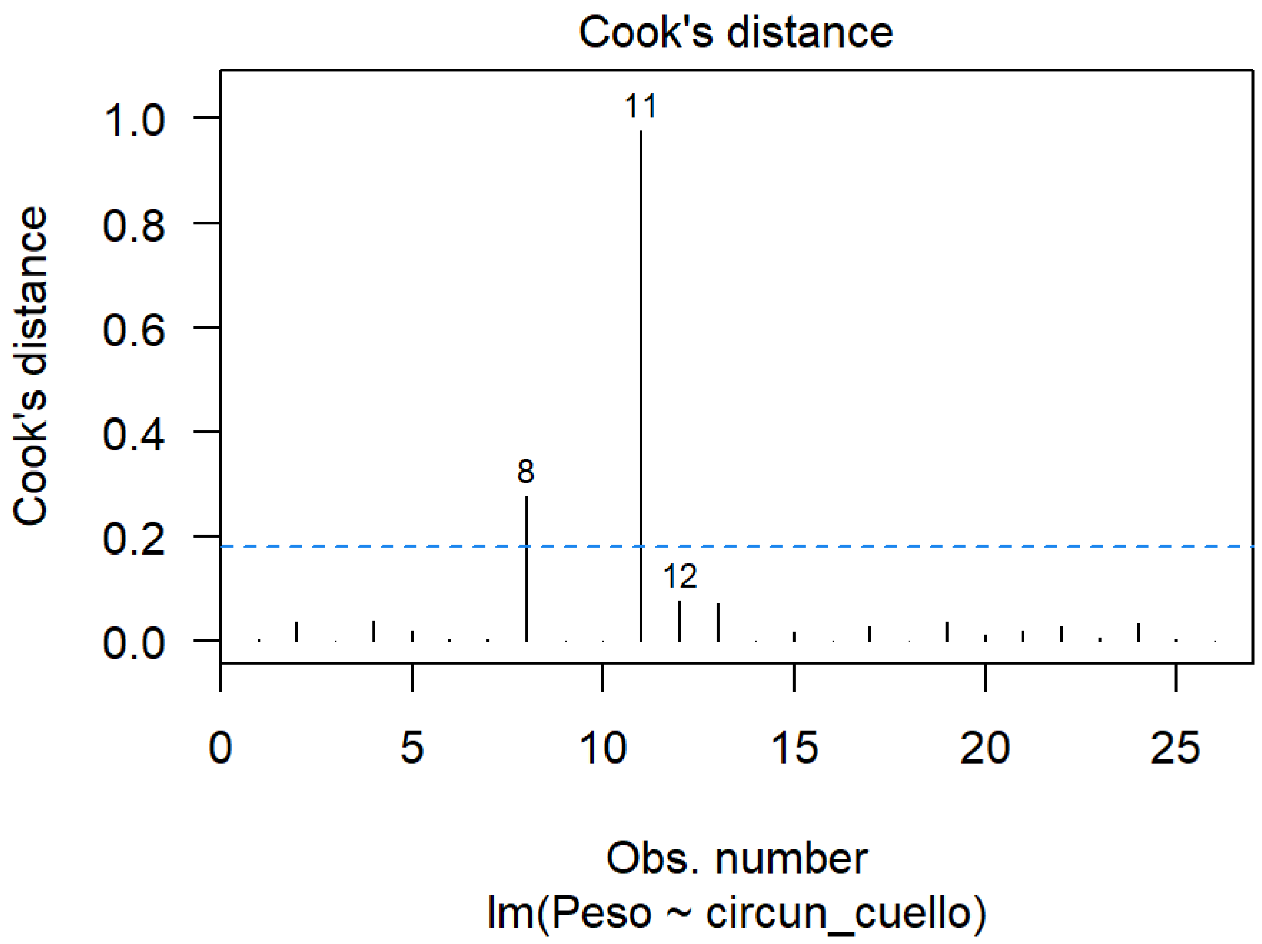
Valores extremos y de influencia

- Veamos algunos de los análisis:



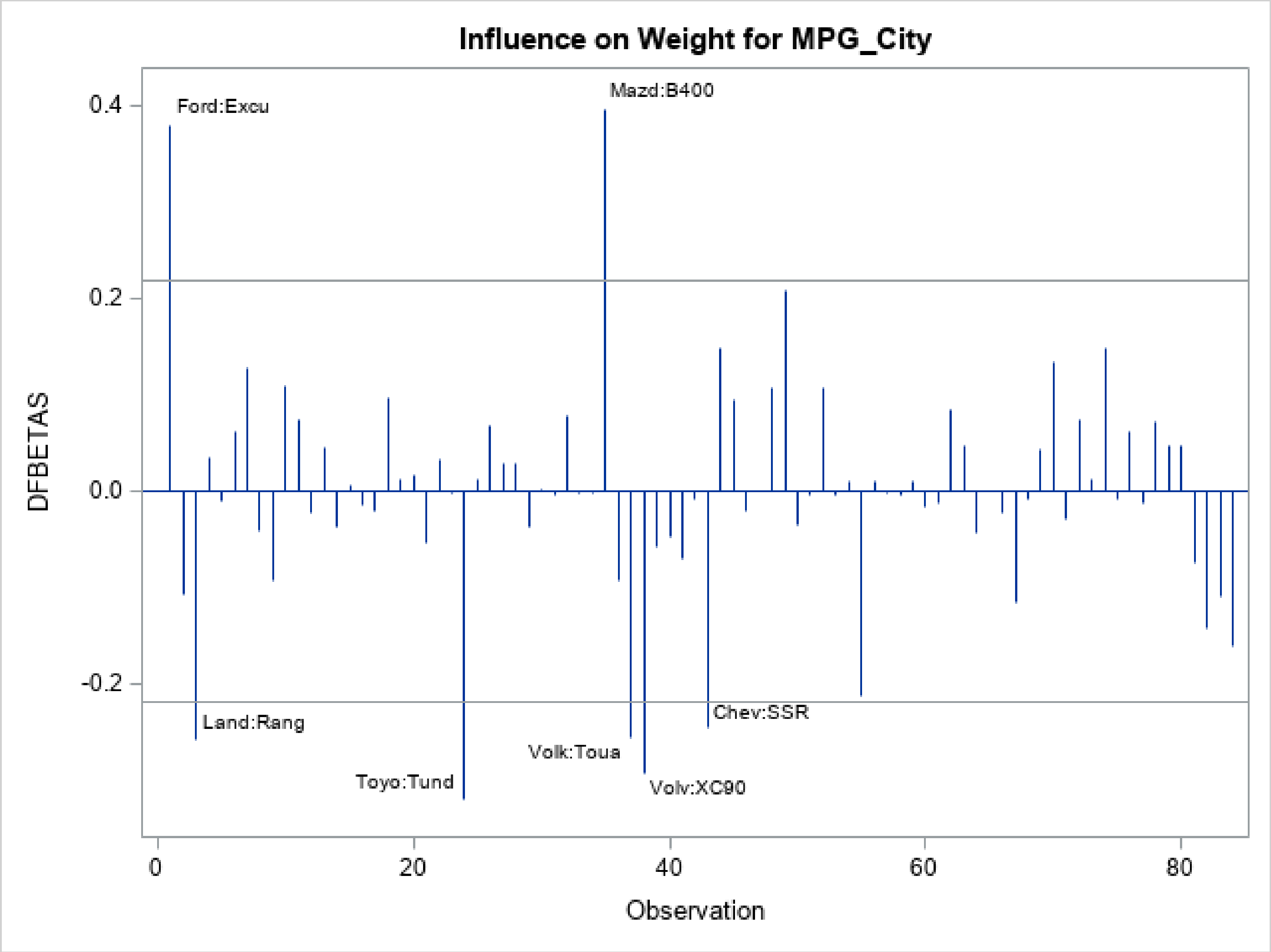
Valores extremos y de influencia

- Veamos algunos de los análisis:



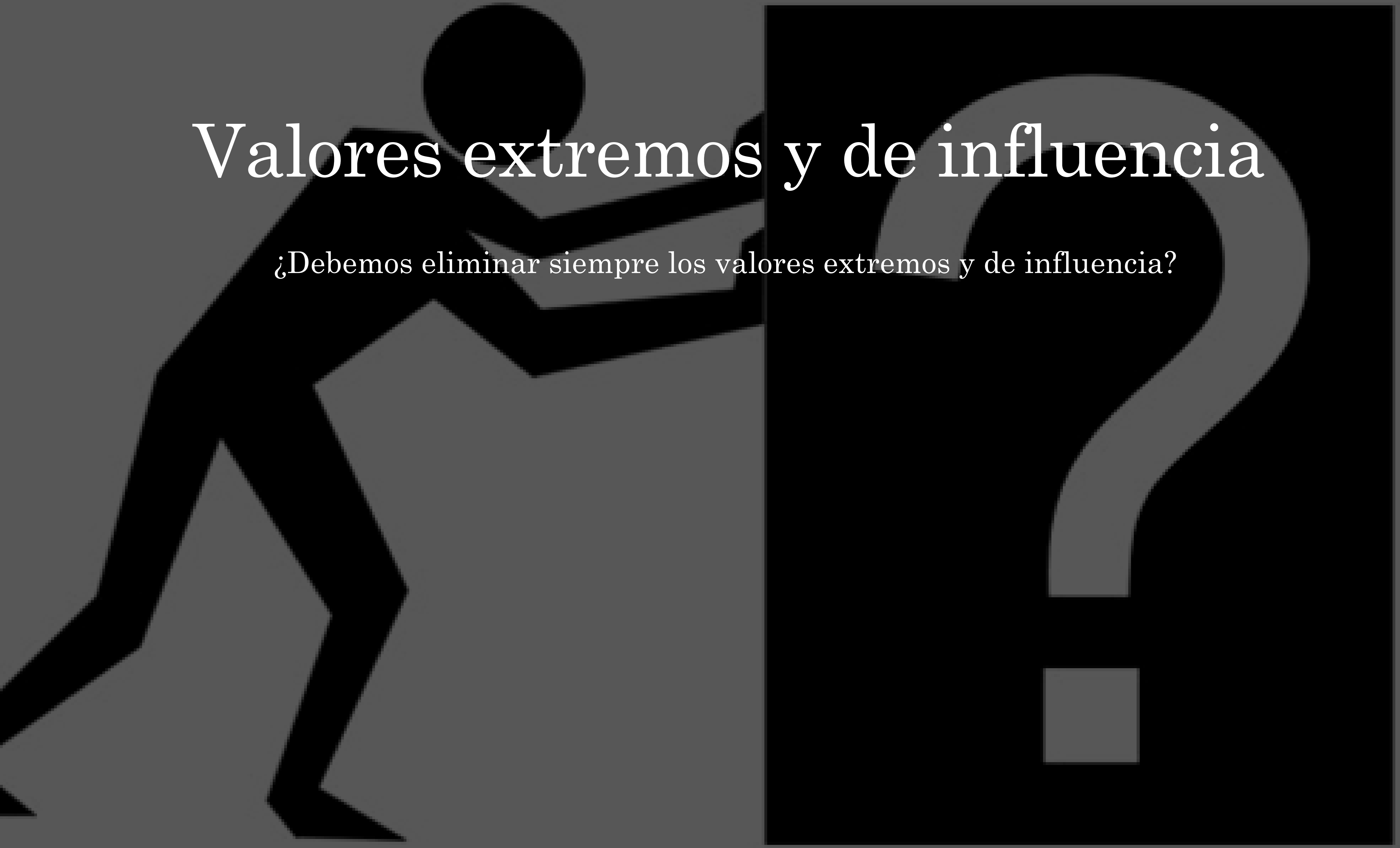
Valores extremos y de influencia

- Veamos algunos de los análisis:



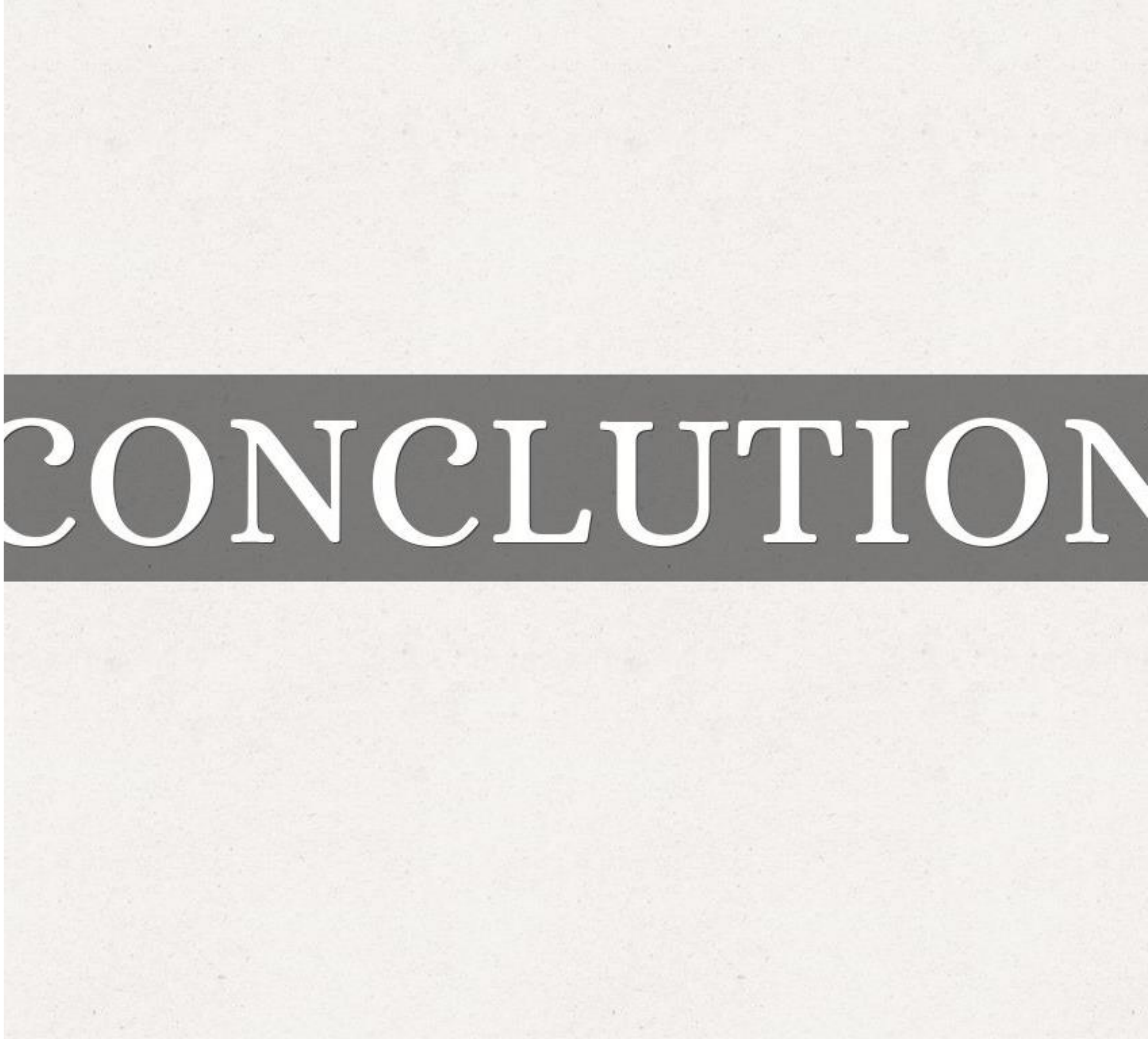
Valores extremos y de influencia

¿Debemos eliminar siempre los valores extremos y de influencia?



Conclusión

- El presente capítulo analizó las medias de diagnóstico o condiciones que debería poseer el análisis de la RLM.
- Se estudiaron los casos propios de los supuesto de la RLM, pero también el caso de poseer valores que distorcionan la adecuación de la RLM.
- Ahora la preguna… ¿y que debemos hacer?
- Veremos algunas medidas remediales que se podrían implementar en el siguiente capítulo.



Las medidas remediales serán nuestra solución…

Tal vez si… pero tal tez no…


The End