



# Regresión Lineal

Máster en Data Science y Big Data en Finanzas (MDS\_F)  
Máster en Data Science y Big Data (MDS)

José Ramón Sánchez Leo

[rsanchez@afi.es](mailto:rsanchez@afi.es)

# Índice

1. Introducción
  - 1.1 Problemas de regresión
2. Regresión lineal simple
  - 2.1 Estimación de los coeficientes
  - 2.2 Precisión de las estimaciones de los coeficientes
  - 2.3 Precisión del modelo
3. Regresión lineal múltiple
  - 3.1 Estimación de los coeficientes
  - 3.2 Determinación de la bondad del ajuste

# Índice

- 4. Métodos de diagnóstico
  - 4.1 Hipótesis
- 5. Evaluación del modelo
- 6. Selección de características
- 7. Problemas de un modelo de regresión
  - 7.1 Uso de variables categóricas
  - 7.2 Relaciones no lineales
  - 7.3 Identificación de *outliers*
  - 7.4 Colinealidad

# 1. Introducción

## 1. Introducción

# Motivación

Queremos aconsejar a una compañía sobre cómo mejorar las ventas de un determinado producto (en miles de unidades).

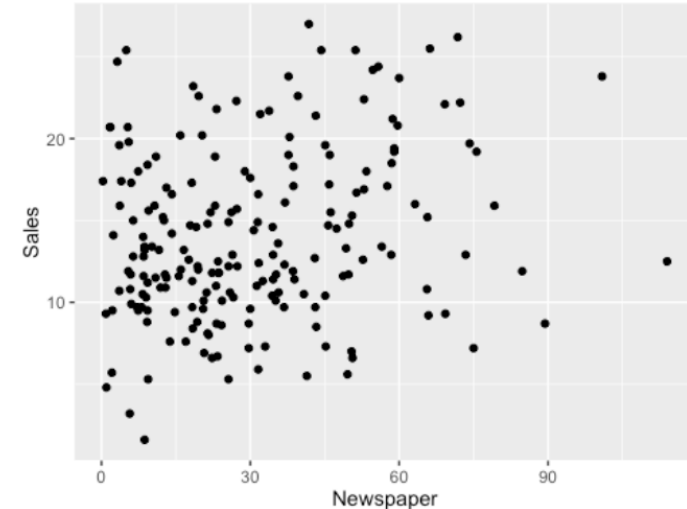
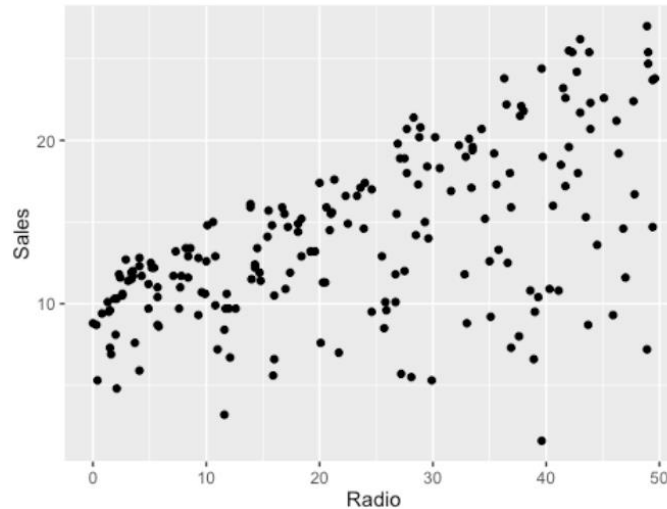
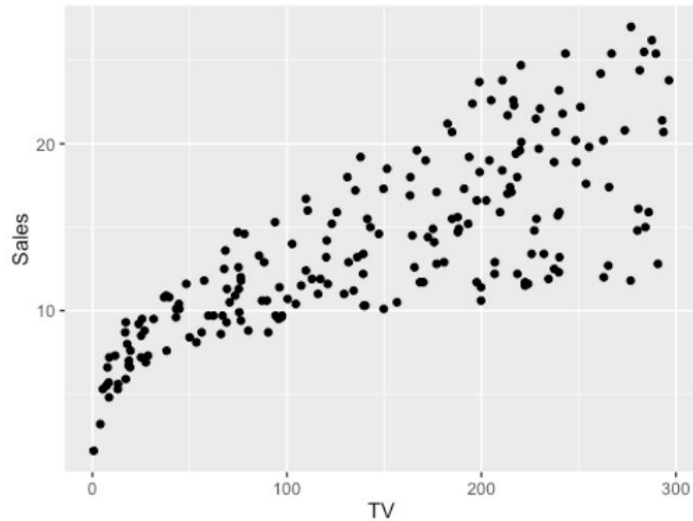
Para conseguirlo, nos proporcionan un set de datos que contiene las ventas del producto en 200 mercados diferentes, junto con el presupuesto de publicidad en televisión, radio y periódicos en cada uno de tales mercados (en millones de dólares).

## 1. Introducción

# Motivación

Queremos aconsejar a una compañía sobre cómo mejorar las ventas de un determinado producto (en miles de unidades).

Para conseguirlo, nos proporcionan un set de datos que contiene las ventas del producto en 200 mercados diferentes, junto con el presupuesto de publicidad en televisión, radio y periódicos en cada uno de tales mercados (en millones de dólares).



## 1. Introducción

# Motivación

Para responder al cliente, lo ideal sería encontrar una función  $f$  que cumpla lo siguiente:

$$Sales = f(TV, Radio, Newspaper)$$

y, si  $P$  es el presupuesto total del que dispone nuestro cliente, buscamos aproximar  $Sales$  en el conjunto:

$$\{TV, Radio, Newspaper \geq 0, TV + Radio + Newspaper \leq P\}$$

El problema es que encontrar una tal  $f$  no sencillo y en la mayoría de los casos es imposible porque:

- Los datos estarán afectados por **ruido**.
- Habrá más **variables relevantes** en el estudio.
- Tendremos una cantidad **insuficiente** de datos.

## 1. Introducción

# Motivación

En este escenario, a las variables de presupuesto se les llama **variables de entrada** (inputs, predictores, features, variables explicativas o variables independientes). Se denotará como  $X.$   $= x_1, x_2, \dots, x_n$

Por otro lado, a la variable *Sales* la notaremos como **variable de salida** (output, respuesta, variable objetivo, target o variable dependiente). A partir de ahora la notaremos como  $Y.$



## 1. Introducción

# Problemas de regresión

En el caso de regresión, tendremos una variable objetivo **cuantitativa** (generalmente numérica continua),  $Y$ , y contaremos con  $p$  predictores, que notaremos como  $X_1, \dots, X_p$ .

Nuestra hipótesis es que existe una relación entre  $Y$  y  $X_1, \dots, X_p$  de la forma:

$$Y = f(X_1, \dots, X_p) + \epsilon = f(X) + \epsilon$$

Donde  $f$  es una función desconocida y  $\epsilon$  es un término de error independiente de  $X$ , mientras que  $f$  representa la información que  $X$  contiene de  $Y$ .

## 1. Introducción

# Problemas de regresión

Un problema de regresión consiste en encontrar una función **hipótesis**  $h$  que aproxima  $f$  de la mejor forma posible con los datos con los que contamos.

Existen diversos motivos por los que uno querría estimar dicha hipótesis, que principalmente serán **realizar predicciones** o para **inferir relaciones**.

# Realizar predicciones

En ocasiones conoceremos un conjunto de entradas  $X$  pero no conoceremos la respuesta  $Y$ . Si hemos construido de forma razonable nuestra hipótesis  $h$ , tenemos que  $h(X)$  será un estimador insesgado de  $Y$ . Podremos por tanto predecir  $Y$  como:

$$\hat{Y} = h(X)$$

La bondad de nuestra estimación puede estimarse a partir de:

$$E \left[ (Y - \hat{Y})^2 \right] = E \left[ (f(X) + \epsilon - h(X))^2 \right] \approx E \left[ (f(X) - h(X))^2 \right] + V(\epsilon)$$

En la expresión anterior,  $E \left[ (f(X) - h(X))^2 \right]$  representa el **error reducible**.

Cuanto mejor sea nuestra hipótesis  $h$ , menor será el error reducible. (Por ello cuando construimos  $h$  tratamos de reducir el error reducible lo máximo posible.)

## 1. Introducción

# Entender relaciones

También podría interesarnos para **entender cómo variaría  $Y$  por los cambios de  $X$** .

Por ejemplo, podríamos querer responder preguntas del tipo:

- ¿Qué predictores tienen relación con la variable respuesta?
- ¿Qué tipo de relación hay entre la variable respuesta y cada uno de los predictores?
- ¿Se puede afirmar que el crecimiento de la variable respuesta con respecto a una feature es lineal? ¿Cuadrático? ¿Exponencial?

Para responder estas preguntas necesitamos conocer la forma de  $h$ .

## 1. Introducción

# Entender relaciones

Volviendo al ejemplo del *dataset* Advertising, para sugerir un plan de marketing a nuestro cliente, podríamos plantearnos responder a las siguientes preguntas:

- ¿Hay relación entre el presupuesto en publicidad y las ventas? Si es que la hay, ¿cómo de fuerte es? ¿Es lineal?
- ¿Qué canales de publicidad contribuyan a las ventas?
- ¿Con cuánta precisión podemos estimar el efecto de cada canal de publicidad en las ventas?
- ¿Con cuánta precisión podemos estimar ventas futuras en base al presupuesto en publicidad?
- ¿Existe interacción entre los distintos canales de publicidad?

El **modelo de regresión lineal** puede ayudarnos a responder a cada una de estas preguntas.



## 2. Regresión lineal simple

## 2. Regresión lineal simple

# Introducción a SLM

El modelo de **regresión lineal** es una de las formas más sencillas de estimar nuestra variable  $Y$  a partir de los predictores. Concretamente, en el modelo de regresión lineal simple contaremos con un único predictor  $X = (X_1)$ , por lo que asumimos que existe una relación como la siguiente:

$$Y \simeq \beta_0 + \beta_1 X_1$$

A partir de los datos disponibles estimaremos los **coeficientes**  $\hat{\beta}_0, \hat{\beta}_1$ , obteniendo como función  $h$ :

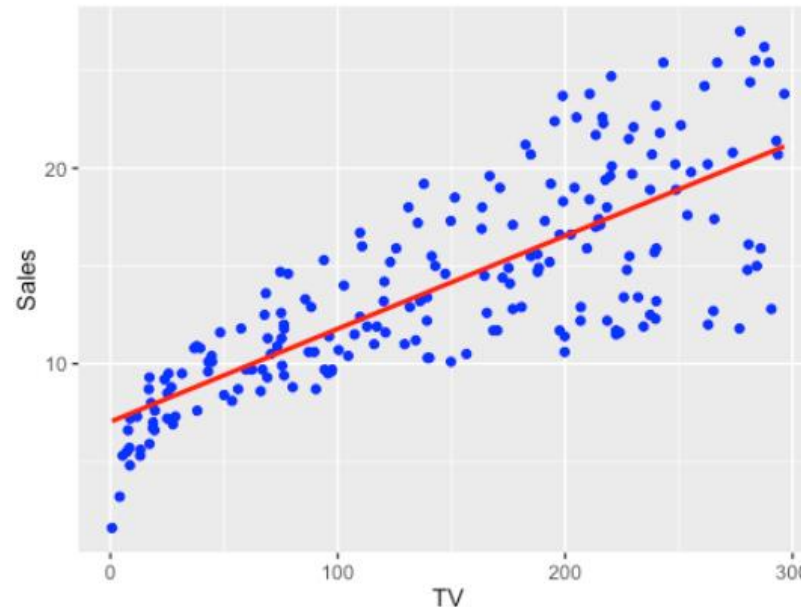
$$Y \simeq \hat{Y} = h(X) = \beta_0 + \beta_1 X_1$$

## 2. Regresión lineal simple

# Introducción a SLM

Supongamos que tenemos una serie de  $n$  puntos  $(x_1, y_1), \dots, (x_n, y_n)$ , nuestras observaciones. Determinar los coeficientes  $\hat{\beta}_0, \hat{\beta}_1$  es buscar una recta que pase muy cerca de todos nuestros puntos:

$$y_i \simeq \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad \forall i \in \{1, 2, \dots, n\}$$





## 2. Regresión lineal simple

# Estimación de coeficientes

Para calcular la recta de regresión usaremos el **método de mínimos cuadrados**.

Para ello, definimos el **residuo** o error en un punto como:

$$e_i = y_i - (\beta_0 + \beta_1 x_i)$$

Tendremos entonces que el error cuadrático medio es:

$$MSE = \frac{1}{n} \sum_{i=1}^n e_i^2$$

Para minimizar este error, se obtiene que los coeficientes deben ser:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

(La demostración se deja como ejercicio. Existe amplia literatura que la recoge).

## 2. Regresión lineal simple

# Precisión del modelo

Para medir la precisión de los coeficientes  $\hat{\beta}_0, \hat{\beta}_1$  debemos asumir la existencia de una relación lineal entre  $X$  e  $Y$  verificando:

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Nos referiremos a la expresión anterior como **recta de regresión poblacional**. Se puede demostrar además que los errores estándar asociados a los coeficientes estimados son:

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right], \quad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

En estos casos, se estima el **error residual estándar** como:

$$\sigma \simeq RSE = \sqrt{\frac{1}{n-2} \sum_{i=1}^n e_i^2}$$

## 2. Regresión lineal simple

# Precisión del modelo

Las expresiones anteriores permiten calcular **intervalos de confianza** y realizar contrastes de hipótesis relativos a los coeficientes del modelo, ya que:

$$\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim t_{n-2}, \quad j = 0, 1$$

Esto nos permite **contrastar si la respuesta está significativamente influida por alguna variable**, mediante el contraste con hipótesis nula:  **$H_0: \beta_j = 0$** .

## 2. Regresión lineal simple

# Precisión del modelo

La medida más utilizada para la bondad es la **raíz cuadrada del error cuadrático medio**:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n e_i^2}$$

Otra alternativa muy utilizada es el estadístico  $R^2$  que se define como:

$$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

que representa la **proporción de variabilidad de la variable respuesta que queda explicada por el modelo**.

## 2. Regresión lineal simple

# Precisión del modelo

La idea es que si la regresión tiene un ajuste suficientemente bueno, será debido a que la variable  $X$  explica buena parte de la variación que  $Y$  experimenta a lo largo de la muestra.

En este caso los residuos serán generalmente pequeños, la variación explicada en  $Y$  será un porcentaje elevado de su variación muestra total, y el coeficiente de determinación será próximo a 1.

## 2. Regresión lineal simple

# Lab + Ejercicio

- **Lab:** (*Lab1.Rmd*) utilizamos el dataset Advertising para construir un modelo de regresión lineal simple sobre la variable Sales utilizando la variable TV.
- **Ejercicio:** (*Ejercicio1.Rmd*) utiliza el dataset Auto, que tienes disponible en el fichero *auto.csv* o utilizando la librería ISLR. Resuelve las siguientes cuestiones:
  - Ajusta un modelo lineal de *mpg* frente a *horsepower* y comenta los resultados.
  - ¿Existe relación entre el predictor y la respuesta?
  - ¿Cómo de fuerte es esa relación?
    - ☐ ¿La relación entre el predictor y la respuesta es positiva o negativa?
    - ☐ ¿Cuál es el mpg predicho para un *horsepower* de 98? Da un intervalo de confianza del 99% para ese valor.
  - Representa gráficamente la respuesta y el predictor, así como el modelo que has ajustado.

(Envía tu resolución a [rsanchez@afi.es](mailto:rsanchez@afi.es) antes de que lo resolvamos en clase.)



# 3. Regresión lineal múltiple

### 3. Regresión Lineal Múltiple

## Estimación de los coeficientes

En el modelo anterior tan solo utilizábamos una variable predictora. Sin embargo, en general contaremos con más variables. Para explotar esta información, vamos a generalizar el modelo de regresión lineal simple.

Se supone que existe una relación lineal como sigue:

$$Y \approx \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

Si tenemos  $n$  observaciones  $(x_{i1}, \dots, x_{ip}; y_i)$ , definimos el **residuo** de la  $i$ -ésima observación como  $e_i = y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})$ .

Por lo que buscamos los coeficientes  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$  que minimicen el error cuadrático medio:

$$MSE = \frac{1}{n} \sum_{i=1}^n e_i^2$$



### 3. Regresión Lineal Múltiple

## Expresión matricial

Se puede expresar de forma matricial como:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}, y = \begin{pmatrix} 1 \\ y_1 \\ \vdots \\ y_n \end{pmatrix},$$

El objetivo es resolver:

$$\arg \min_{\hat{\beta}} \|y - X\hat{\beta}\|^2$$

cuya solución es:

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

### 3. Regresión Lineal Múltiple

## Bondad del ajuste

Al igual que en SLM, se puede medir la bondad del ajuste a partir del  $RMSE$  o  $R^2$ . Mejor que este último, se puede utilizar el estadístico  **$R^2$  ajustado**:

$$R_a^2 = R^2 - (1 - R^2) \frac{p}{n - p - 1}$$

Donde se penaliza el uso de muchos predictores.

( $n$ : número de observaciones;  
 $p$ : número de predictores)

### 3. Regresión Lineal Múltiple

## Lab

- **Lab:** (*Lab2.Rmd*) con el dataset Advertising ajustamos un modelo de regresión múltiple sobre la variable Sales utilizando todas las variables.



## 4. Métodos de diagnóstico

## 4. Métodos de diagnóstico

# Hipótesis

Los modelos de regresión lineal asumen las siguientes suposiciones:

- Relación lineal entre la variable objetivo y las variables predictoras.
- Los errores tienen media cero.
- Los errores tienen varianza constante (homocedasticidad).
- Los errores son independientes.

Concretamente, los residuos deben ser una variable aleatoria  $\epsilon \sim N(0, \sigma^2)$

**Remember:** En un contraste de hipótesis con un test asociado tomaremos la decisión de rechazar o no según el p-valor del estadístico del test:

$$\text{p-valor} < \alpha \Rightarrow \text{Rechazo } H_0$$

$$\text{p-valor} \geq \alpha \Rightarrow \text{No rechazo } H_0$$

( $\alpha$  nivel de significación)

# Diagnóstico de un modelo de Regresión Lineal

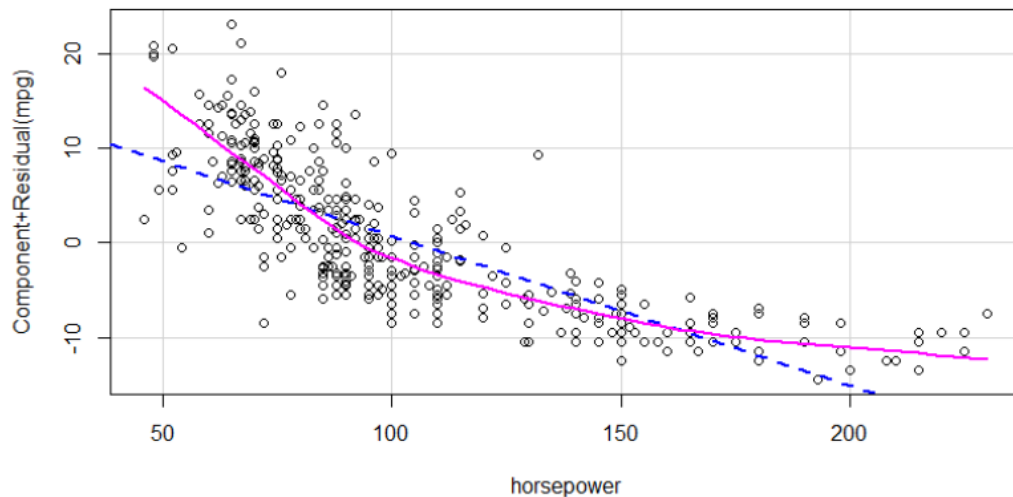
- **Diagnóstico cualitativo:** gráficas de residuos frente a variables, predictores frente a objetivo, ...
- **Diagnóstico cuantitativo:** uso de test de hipótesis o estimadores.

## 4. Métodos de diagnóstico

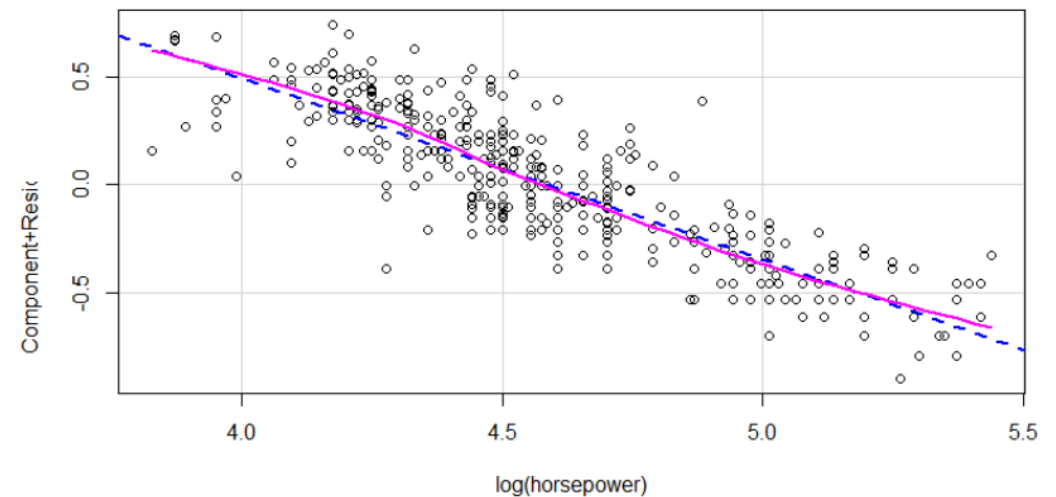
# Relación Lineal

**crPlots** (Función de la librería car)

```
lm_fit_mpg <- lm(mpg ~ horsepower , data = Auto)  
crPlots(lm_fit_mpg)
```



```
lm_fit_log_mpg<- lm(log_mpg ~ log(horsepower) , data = Auto)  
crPlots(lm_fit_log_mpg)
```



## 4. Métodos de diagnóstico

# Relación Lineal

Para determinar si realmente hay una relación lineal entre nuestra variable respuesta y los predictores se realiza el test de hipótesis con hipótesis nula:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

Un estadístico relevante en este caso será:

$$F = \frac{n - p - 1}{p} \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \hat{y})^2}$$

que, bajo la hipótesis nula, sigue una distribución **F de Snedecor** de parámetros  $(p, n - p - 1)$



## 4. Métodos de diagnóstico

# Relación Lineal

Se puede comprobar la hipótesis anterior mediante el **test de Rainbow**, cuya hipótesis nula es la siguiente:

$$H_0: F \sim F_{(p, n-p-1)}, \quad H_1: \text{existe linealidad}$$

```
> raintest(model)

Rainbow test

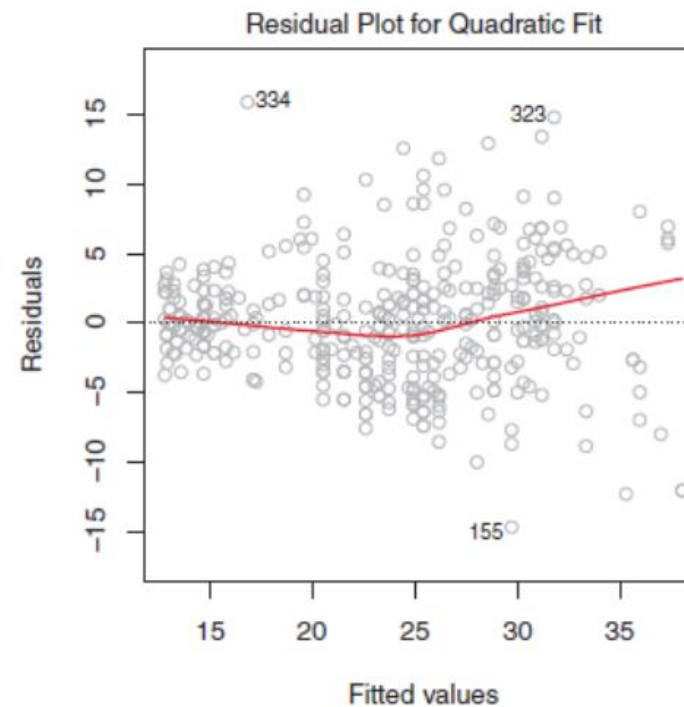
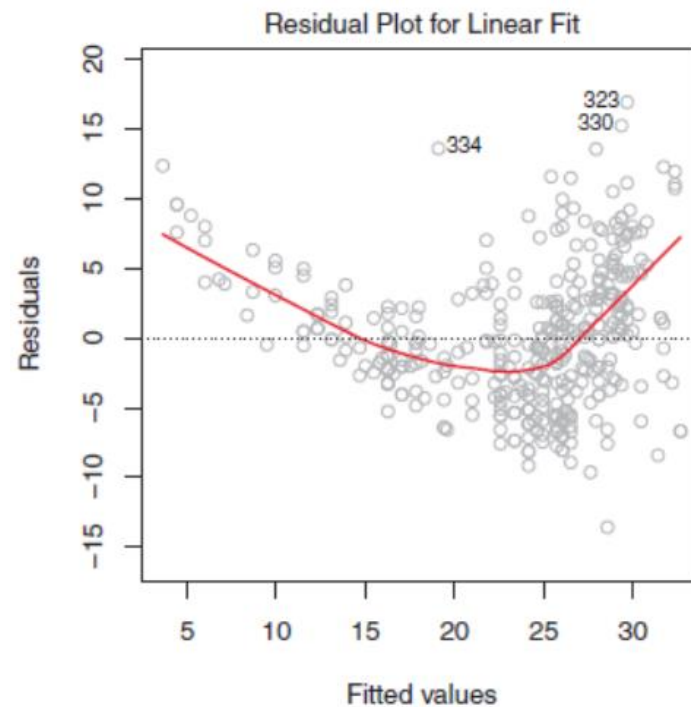
data: model
Rain = 2.9632, df1 = 196, df2 = 194, p-value = 7.152e-14
```

En este caso, se obtiene un p-valor muy pequeño, por lo que rechazamos la hipótesis nula.

(Para que se cumplan la hipótesis de linealidad de un modelo de regresión lineal debemos obtener un p-valor < 0.05, así rechazamos la hipótesis nula)

## 4. Métodos de diagnóstico

# Residuos: media nula



En la primera vemos que los errores están lejos de la media nula.....

## 4. Métodos de diagnóstico

# Residuos: media nula

Test de hipótesis:  $H_0: \mu = 0$ ,  $H_1: \mu \neq 0$

$$\frac{\bar{X} - \mu}{S\sqrt{n}} \simeq t_{n-1}$$

```
> t.test(X, mu=0)
```

One Sample t-test

```
data: X
t = 5.4554, df = 250, p-value = 1.174e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.01597462 0.03402538
sample estimates:
mean of x
 0.025
```

```
t.test(mpg_fit_who$residuals)
```

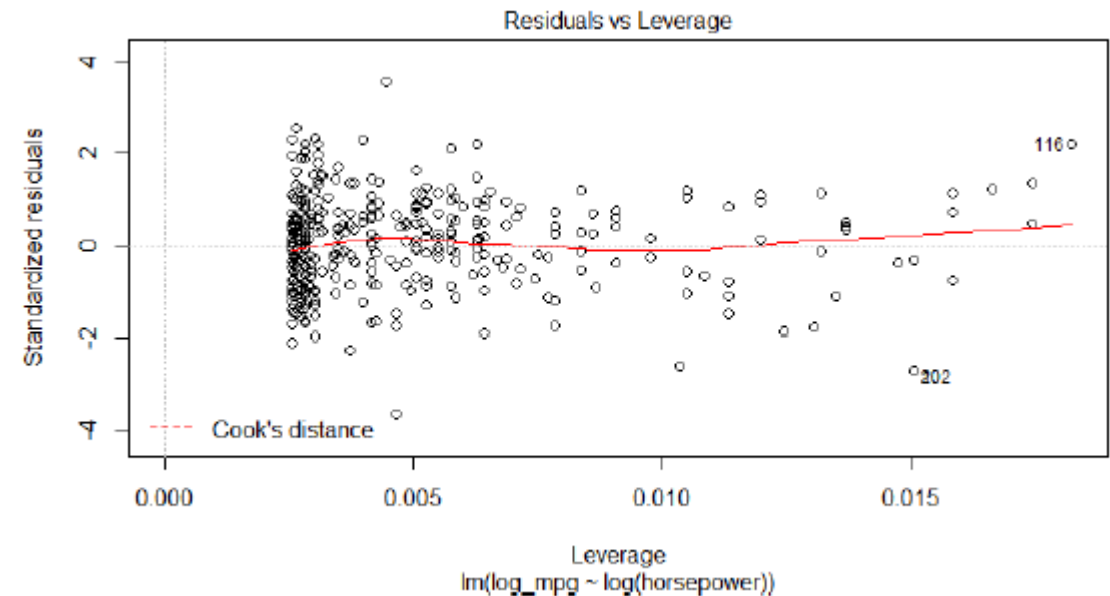
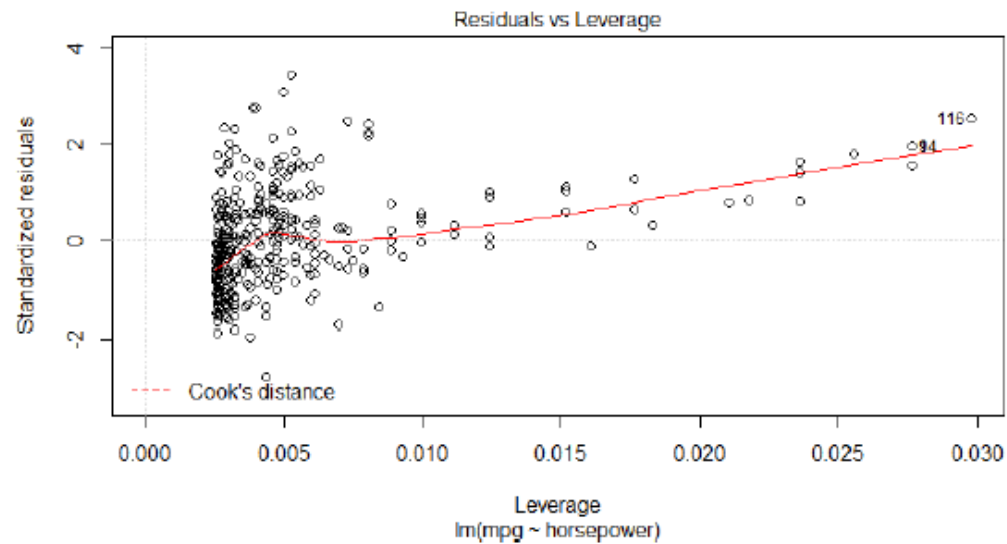
```
##
## One Sample t-test
##
## data: mpg_fit_who$residuals
## t = 2.8221e-16, df = 391, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -0.3297026 0.3297026
## sample estimates:
## mean of x
## 4.732653e-17
```

Es la única función que recibe como input los residuos, en lugar del modelo.

(Para que se cumplan la hipótesis de errores con media de un modelo de regresión lineal debemos obtener un p-valor  $> 0.05$ , así rechazamos la hipótesis nula)

## 4. Métodos de diagnóstico

# Residuos: varianza no homogénea



Estos gráficos se obtienen al ejecutar `plot(linear_model)` en R.

## 4. Métodos de diagnóstico

# Residuos: varianza no homogénea

bptest: **Breusch-Pagan Test** (lmtest)

Test de hipótesis:  $H_0$ : existe homocedasticidad

```
> bptest(lm_fit_mpg)

      studentized Breusch-Pagan test

data:  lm_fit_mpg
BP = 8.7535, df = 1, p-value = 0.00309

> bptest(lm_fit_log_mpg)

      studentized Breusch-Pagan test

data:  lm_fit_log_mpg
BP = 4.1216, df = 1, p-value = 0.04234
```

- En el primer caso rechazamos la hipótesis nula
  - podemos decir entonces que existe heterocedasticidad
- En el segundo caso no podemos rechazar la hipótesis nula.
  - Podemos decir entonces que existe **homocedasticidad**.

## 4. Métodos de diagnóstico

# Residuos: autocorrelación nula (independencia)

dwtest: **Durbin-Watson test**(lmtest)

Se usa para detectar existencia de autocorrelación.

Test de hipótesis:  $H_0$ : autocorrelación nula

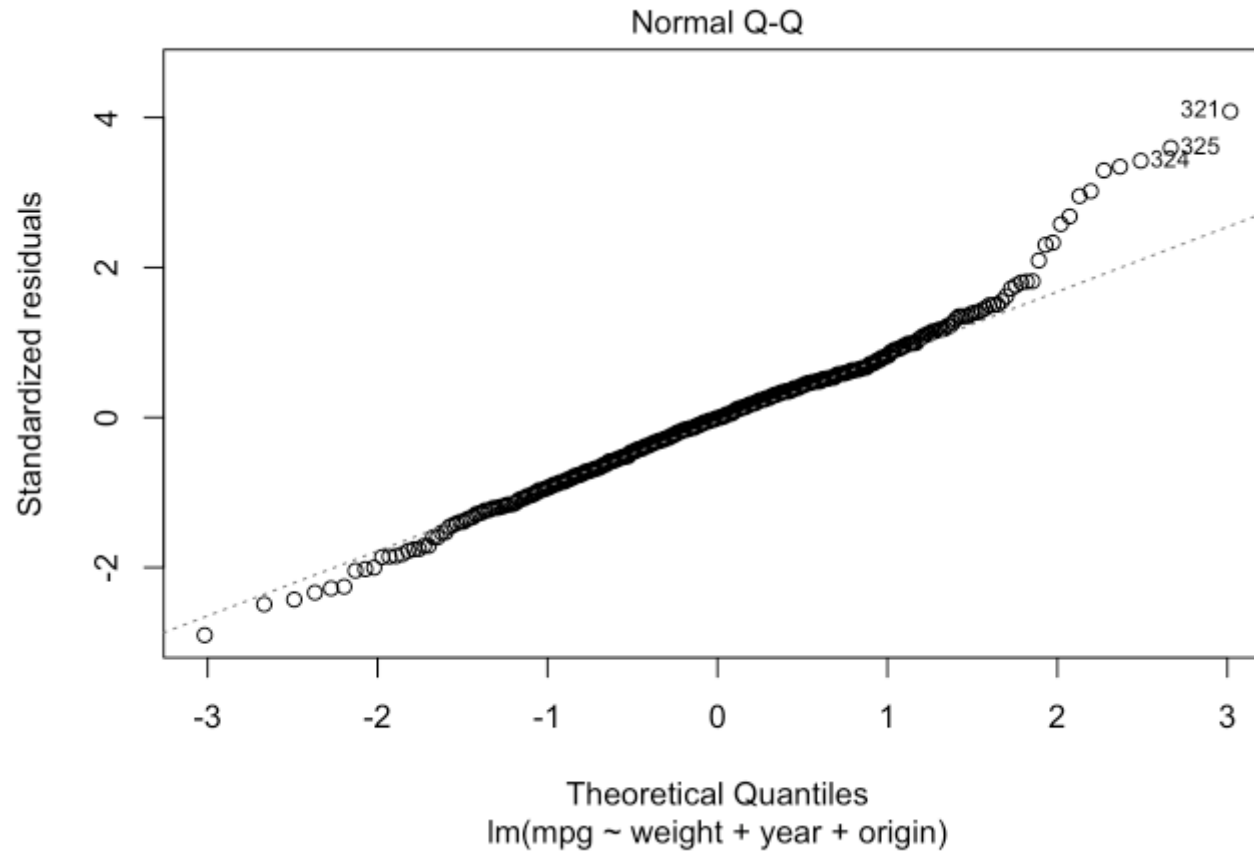
```
dwtest(modelo_lineal, alternative = "two.sided", iterations = 1000)

##
##  Durbin-Watson test
##
## data:  modelo_lineal
## DW = 1.9483, p-value = 0.5649
## alternative hypothesis: true autocorrelation is not 0
```

El p-valor es muy alto, por lo que no podemos rechazar la hipótesis nula que afirma que la autocorrelación es nula.

## 4. Métodos de diagnóstico

# Errores normales



## 4. Métodos de diagnóstico

# Errores normales

Shapiro.test: **Shapiro-Wilk Test**(lmtest)

La prueba de Shapiro-Wilk es un test estadístico empleado para contrastar la normalidad de un conjunto de datos. El test utiliza el contraste de hipótesis para rechazar la normalidad de la muestra.

Test de hipótesis:  $H_0$ : la distribución es normal. ( $X \sim N(\mu, \sigma^2)$ )

```
> shapiro.test(lm_fit_mpg$residuals)

      Shapiro-Wilk normality test

data:  lm_fit_mpg$residuals
W = 0.98207, p-value = 8.734e-05

> shapiro.test(lm_fit_log_mpg$residuals)

      Shapiro-Wilk normality test

data:  lm_fit_log_mpg$residuals
W = 0.99354, p-value = 0.09273
```

Cuando el p-valor es menor que 0.05 se rechaza la hipótesis nula, es decir, los residuos no se distribuyen según una normal.



## 4. Métodos de diagnóstico

# Errores normales

ad.test: **Anderson-Darling Test**(lmtest)

Cuando el **tamaño muestral es reducido**, el test Anderson-Darling posee mayor de detección de la no normalidad, mientras que se suele usar el test de Shapiro-Wilk cuando se analizan muestras de tamaños superiores.

```
> ad.test(lm_fit_mpg$residuals)

Anderson-Darling normality test

data:  lm_fit_mpg$residuals
A = 1.9162, p-value = 6.785e-05

> ad.test(lm_fit_log_mpg$residuals)

Anderson-Darling normality test

data:  lm_fit_log_mpg$residuals
A = 0.38119, p-value = 0.3994
```

## 4. Métodos de diagnóstico

# Summary

Para comprobar las hipótesis de un modelo de regresión lineal debemos realizar los siguientes test de hipótesis que nos sirven como *checks*:

	Suposición	Test	$H_0, H_1$	R	p-valor
1	Relación lineal entre la variable target y las explicativas	Rainbow	$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ $H_1: \text{existe linealidad}$	raintest	<
2	Residuos: media nula	T-test (univariante)	$H_0: \mu = 0$	t.test	>
3	Residuos: homocedasticidad	Breusch-Pagan	$H_0: \text{existe homocedasticidad}$	bp.test	>
4	Residuos: autocorrelación nula	Durbin-Watson	$H_0: \text{autocorrelación nula}$	dwtest	>
5	Errores normales	Shapiro-Wilk	$H_0: \text{distribución es normal}$	shapiro.test	>
6	Errores normales (tamaño muestral reducido)	Anderson-Darling	$H_0: \text{distribución es normal}$	ad.test	>

## 4. Métodos de diagnóstico

# Lab + Ejercicio

- **Lab:** (*Lab3.Rmd*) método de diagnóstico.
- **Ejercicio:** (*Ejercicio2.Rmd*) utiliza el dataset Auto, que tienes disponible en el fichero *auto.csv* o utilizando la librería ISLR. Resuelve las siguientes cuestiones:
  - Realiza un análisis exploratorio sencillo.
  - Representa diagramas de dispersión de todos los pares de variables numéricas. Calcula la matriz de correlaciones. A priori, ¿piensas que tiene sentido ajustar un modelo lineal de mpg frente al resto de variables, o frente a alguna de ellas?
  - Para confirmar o desmentir tu sospecha, ajusta tal modelo y trata de optimizarlo.
    - ☐ ¿Existe relación entre los predictores y la respuesta?
    - ☐ ¿Cómo varía la respuesta frente a cada uno de los predictores?
  - Realiza el diagnóstico del modelo

(Envía tu resolución a [rsanchez@afi.es](mailto:rsanchez@afi.es) antes de que lo resolvamos en clase.)



# 5. Evaluación del modelo

# Estimación del error de predicción

Asumiendo que existe relación lineal entre  $X$  e  $Y$  y que  $n \gg p$ , las predicciones obtenidas por el modelo de regresión lineal serán buenas. Si  $n \simeq p$  o  $n < p$  esto es falso. Esto último nos indica que es mejor construir modelos con **pocas variables**.

Para comparar la bondad de diferentes modelos compararemos el error de ambos. Se consideran dos alternativas:

- Ajustar el error de entrenamiento para estimar el verdadero error de predicción.
- Emplear un conjunto de validación, que no intervenga en el proceso de entrenamiento. (Técnicas de remuestreo).

# Estimación del error de predicción

Naturalmente, si incluimos más variables en el ajuste, el  $RSME$  y  $R^2$  mejorarán, o, en el peor de los casos, se mantendrán. Es importante notar que esto no hace que el modelo sea mejor.

Se define el **residual sum of squares** como sigue:

$$RSS = nMSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

## 5. Evaluación del modelo

# Estimación del error de predicción

Para estimar el error de predicción, se puede utilizar los siguientes estadísticos:

- Criterio de información de Akaike (AIC):

$$AIC = \frac{1}{n\hat{\sigma}^2} (RSS + 2p\hat{\sigma}^2) + C$$

- Criterio de información Bayesiana (BIC):

$$BIC = \frac{1}{n} (RSS + p\hat{\sigma}^2 \log(n))$$

- El estadístico  $C_p$  de Mallows:

$$C_p = \frac{1}{n} (RSS + 2p\hat{\sigma}^2)$$

- $R^2$  *ajusted*

El objetivo es minimizar los tres primeros o maximizar el último.

# Criterio de información de Akaike

$$AIC = \frac{1}{n\hat{\sigma}} (RSS + 2p\hat{\sigma}) + C$$

- $C$  es una constante que depende de la muestra.
- Es un estadístico que no está acotado
- Cuanto menor es este valor, mejor es el ajuste.
- Al aumentar el número de predictores el AIC aumenta. Por tanto, para que compense aumentar los predictores, se debe reducir el error. En resumen: **hacer un modelo más complejo solo compensa si el modelo es sustancialmente mejor.**



# Criterio de información Bayesiana

$$BIC = \frac{1}{n} (RSS + p\hat{\sigma} \log(n))$$

- Estadístico no acotado.
- Cuanto menor es este valor, mejor es el ajuste
- Este criterio **penaliza aún más que el anterior la inclusión de nuevos predictores.**

## 5. Evaluación del modelo

# Estadístico $C_p$ de Mallows

$$C_p = \frac{1}{n} (RSS + 2p\hat{\sigma}^2)$$

Compara la precisión y el sesgo del modelo completo con modelos que incluyen un subconjunto de predictores.

Similar al AIC. Cuanto menor es el valor, más preciso (tiene menos varianza) es el modelo en la estimación de los coeficientes de regresión verdaderos y en sus predicciones.



## 6. Selección de características

# Selección de variables

Idealmente podríamos considerar todos los modelos lineales posibles de  $Y$  en función de los conjuntos de  $X_1, X_2, \dots, X_p$  para elegir el mejor modelo posible dentro de la información con la que contamos.

En la práctica esto puede ser costoso, por lo que contamos con diferentes técnicas de **selección de variables**.

- **Best subset selection**
- **Forward stepwise selection**
- **Backward stepwise selection**

# Best subset selection

1. Sea  $\mathcal{M}_0$  el modelo nulo, es decir, sin predictores.
2. Para  $k = 1, 2, \dots, p$ :
  1. Ajustar todos los modelos que contienen exactamente  $k$  predictores.
  2. Elegimos como  $\mathcal{M}_k$  el mejor de todos los anteriores ( $RSS$  o  $R^2$ ).  
Hay un total de  $\binom{p}{k}$  posibilidades.
3. Elegimos el mejor de los modelos entre  $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$  según algún criterio ( $AIC, BIC, C_p, R_a^2$ ) o error de validación

Se prueban en total  $2^p$  modelos.

# Forward stepwise selection

1. Sea  $\mathcal{M}_0$  el modelo nulo, es decir, sin predictores.
2. Para  $k = 0, 1, 2, \dots, p - 1$ :
  1. Ajustar todos los  $p - k$  modelos que se obtienen al **añadir** un predictor a  $\mathcal{M}_k$ .
  2. Elegimos como  $\mathcal{M}_{k+1}$  el mejor de todos los anteriores ( $RSS$  o  $R^2$ ).
3. Elegimos el mejor de los modelos entre  $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$  según algún criterio ( $AIC, BIC, C_p, R_a^2$ ) o error de validación

Se prueban en total  $\frac{p^2+p}{2}$  modelos.

# Backward stepwise selection

1. Sea  $\mathcal{M}_p$  el modelo completo, es decir, sin predictores.
2. Para  $k = p, p - 1, \dots, 1$ :
  1. Ajustar todos los  $k$  modelos que se obtienen al **eliminar** un predictor a  $\mathcal{M}_k$ .
  2. Elegimos como  $\mathcal{M}_{k-1}$  el mejor de todos los anteriores ( $RSS$  o  $R^2$ ).
3. Elegimos el mejor de los modelos entre  $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$  según algún criterio ( $AIC, BIC, C_p, R_a^2$ ) o error de validación

Se prueban en total  $\frac{p^2+p}{2}$  modelos.

# Pasos en la construcción del modelo

Podemos dividir la construcción de un modelo de regresión en los siguientes tres pasos:

1. **Seleccionar** cuáles serán los predictores de entre todas las variables posibles: elegir  $X_1, X_2, \dots, X_n$ .
2. **Ajustar** el modelo. En este caso buscamos determinar los coeficientes  $\beta_0, \beta_1, \dots, \beta_n$ .
3. **Evaluar** la calidad del modelo obtenido.



## 5. Evaluación del modelo

# Ajuste del modelo en R

*lm(target ~ predictores, datos)*

```
Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min       1Q   Median       3Q      Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.30337    18.71788   0.657   0.5181
cyl          -0.11144     1.04502  -0.107   0.9161
disp         0.01334     0.01786   0.747   0.4635
hp           -0.02148     0.02177  -0.987   0.3350
drat         0.78711     1.63537   0.481   0.6353
wt          -3.71530     1.89441  -1.961   0.0633
qsec         0.82104     0.73084   1.123   0.2739
vs           0.31776     2.10451   0.151   0.8814
am           2.52023     2.05665   1.225   0.2340
gear         0.65541     1.49326   0.439   0.6652
carb        -0.19942     0.82875  -0.241   0.8122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,    Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```



# 7. Problemas de un modelo de regresión

## 7. Problemas de un modelo de regresión

# Uso de predictores categóricos


Se ha definido el modelo de regresión lineal como:

$$Y \approx \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

¿Deben ser entonces todos los predictores variables numéricas?

Podemos **dummificar** las variables categóricas

Row	Animal
1	Cat
2	Dog
3	Bird
4	Dog
5	Dog
6	Cat



Row	Cat	Dog	Bird
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0
5	0	1	0
6	1	0	0

## 7. Problemas de un modelo de regresión

# Uso de predictores categóricos

Si una variable de nuestro conjunto es categórica la transformaremos en variable numérica para poder utilizarla en nuestro modelo de regresión lineal.

*Dummificar* una variable categórica  $X$  que toma  $v$  valores diferentes  $(x_1, \dots, x_v)$  consiste en crear  $v$  variables nuevas

$$X_{(k)} = \begin{cases} 1 & \text{si } X = x_k \\ 0 & \text{c. c.} \end{cases}$$

**Nota:** alternativamente se pueden construir tan solo  $v - 1$  variables.

## 7. Problemas de un modelo de regresión

# Relaciones no lineales

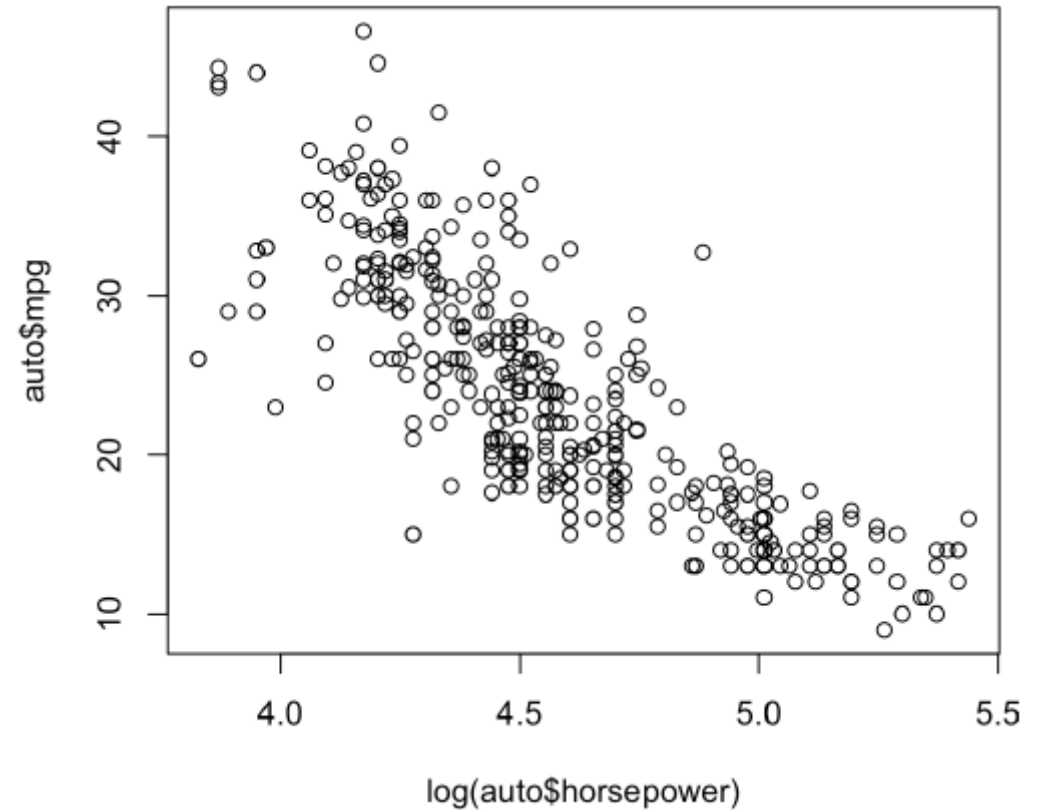
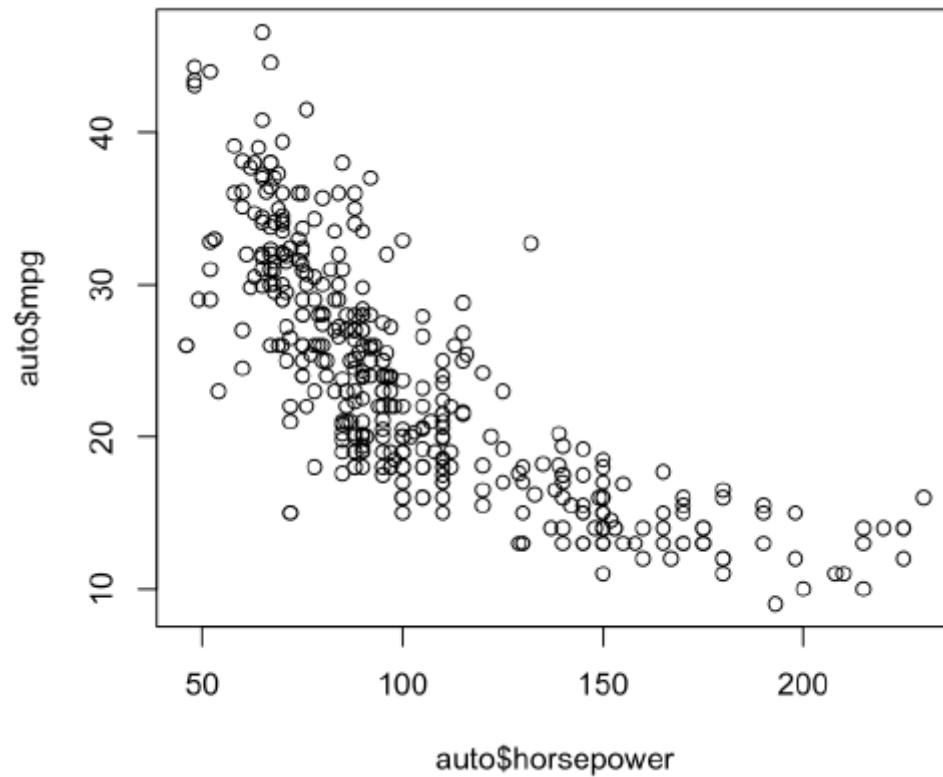
El modelo de regresión lineal se construye bajo la hipótesis de existencia de relaciones lineales entre los predictores y la variable respuesta.

¿Y si esta relación no existe?

En muchas ocasiones encontraremos otro tipo de relaciones entre las variables. Una simple **transformación sobre los predictores** puede ayudara conseguir relaciones lineales.

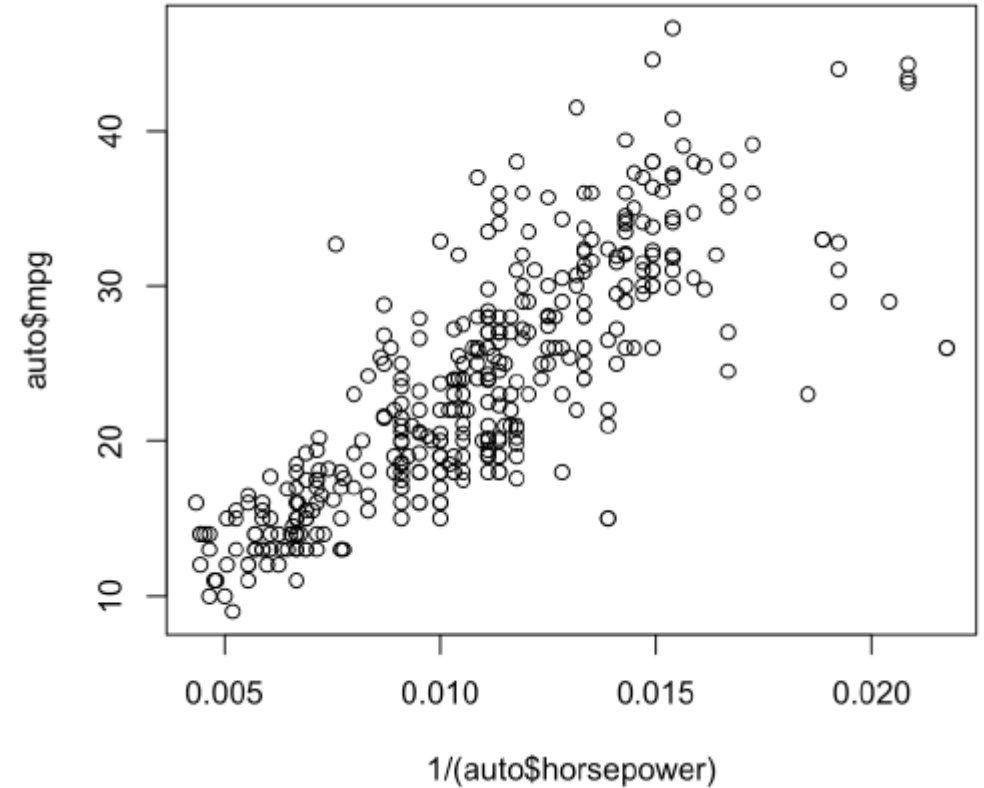
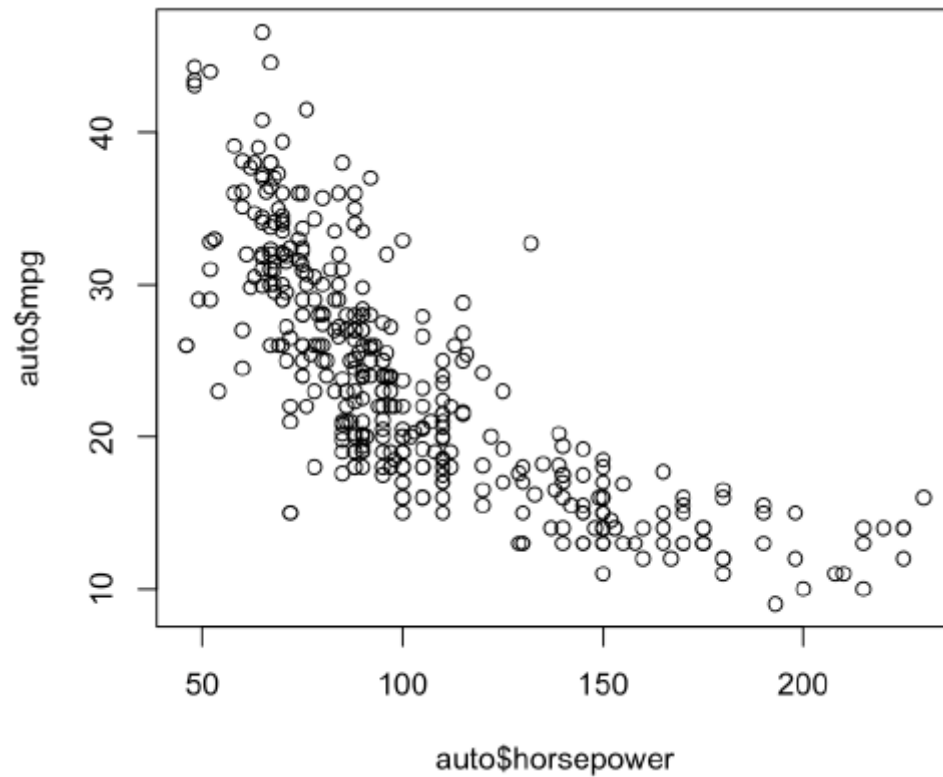
## 7. Problemas de un modelo de regresión

# Relaciones no lineales



## 7. Problemas de un modelo de regresión

# Relaciones no lineales



## 7. Problemas de un modelo de regresión

# Relaciones no lineales

Algunas transformaciones habituales son:

- Polinomios (y sus transformaciones inversas):  $f(x) = x^P$ ;  $f(x) = \sqrt[p]{x}$ .
- Inversa:  $f(x) = \frac{1}{x}$ .
- Logaritmos o exponenciales:  $\log(x)$ ,  $\log(x + 1)$ ,  $e^x$

Siempre hay que asegurarse que las transformaciones están bien definidas.

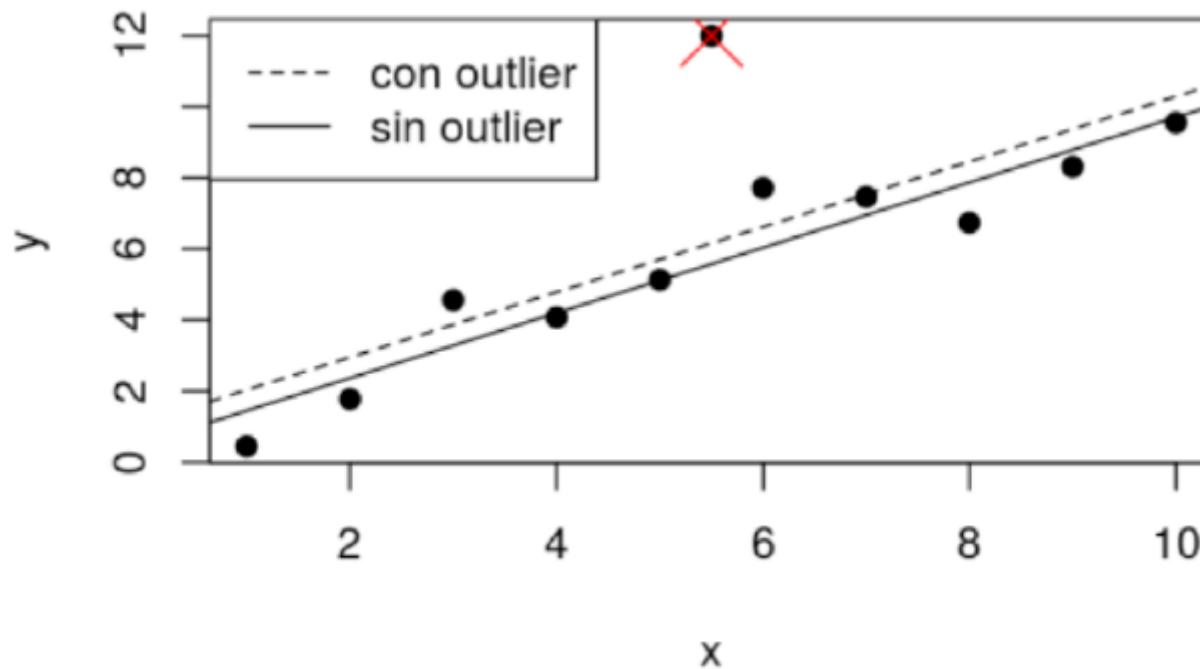
(Atención al calcular errores)



## 7. Problemas de un modelo de regresión

# Identificación de outliers

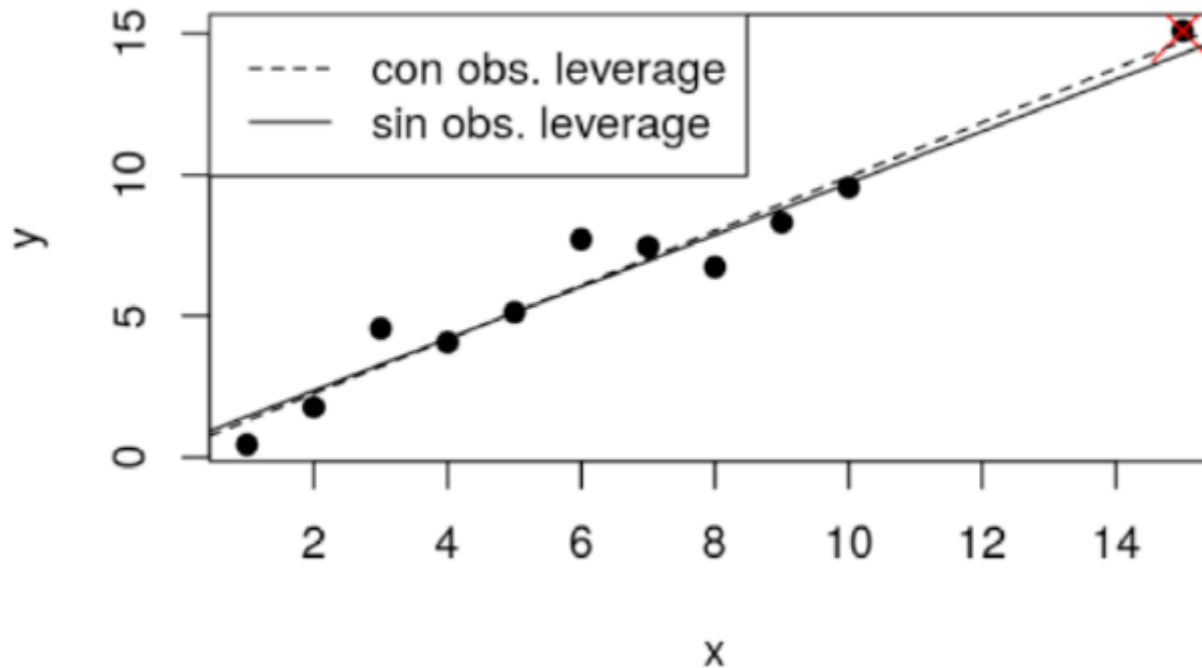
**Outlier u observación atípica:** observaciones que no se ajustan bien al modelo. El valor real de la variable respuesta se aleja mucho del valor predicho, por lo que su residuo es excesivamente grande.



## 7. Problemas de un modelo de regresión

# Identificación de outliers

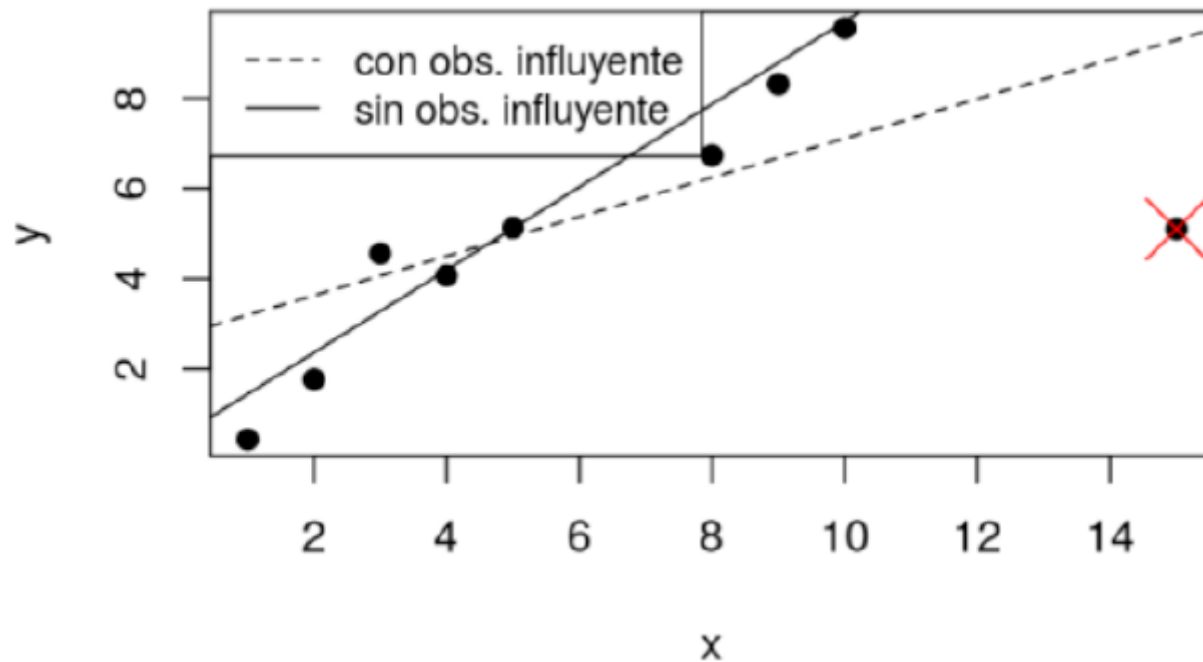
**Observación con alto leverage:** observación con un valor extremo para alguno de los predictores.



## 7. Problemas de un modelo de regresión

# Identificación de outliers

**Observación influyente:** observación que influye notablemente en el modelo, haciendo que su exclusión cambie notablemente el resultado final. No todos los outliers ni puntos con high leverage tienen por qué ser influyentes.



## 7. Problemas de un modelo de regresión

# Puntos influyentes

Para determinar si una cierta observación  $j$  es de *high leverage* para la variable  $X$ , se calcula:

$$h_j = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Donde  $\frac{1}{n} \leq h_j$  y la media de los  $h_j = \frac{2}{n}$ . Por tanto, cuando  $h_j \gg \frac{2}{n}$ , se considera que  $j$  es una observación con *high leverage*.

El problema es que estamos detectando puntos influyentes para un único predictor, pero podría serlo de forma multivariante.

## 7. Problemas de un modelo de regresión

# Identificación de outliers

En R, se puede usar la función **hatvalues** para calcular el leverage de cada observación.

Para estimar la influencia de cada observación en el modelo final se puede utilizar la **distancia de Cook**, que combina la magnitud del residuo y el grado de leverage.

En R se calcula con la función `cook.distance(model)`.

Valores de Cook mayores a 1 suelen considerarse puntos influyentes.

# ¿Qué hacer con outliers o influyentes?

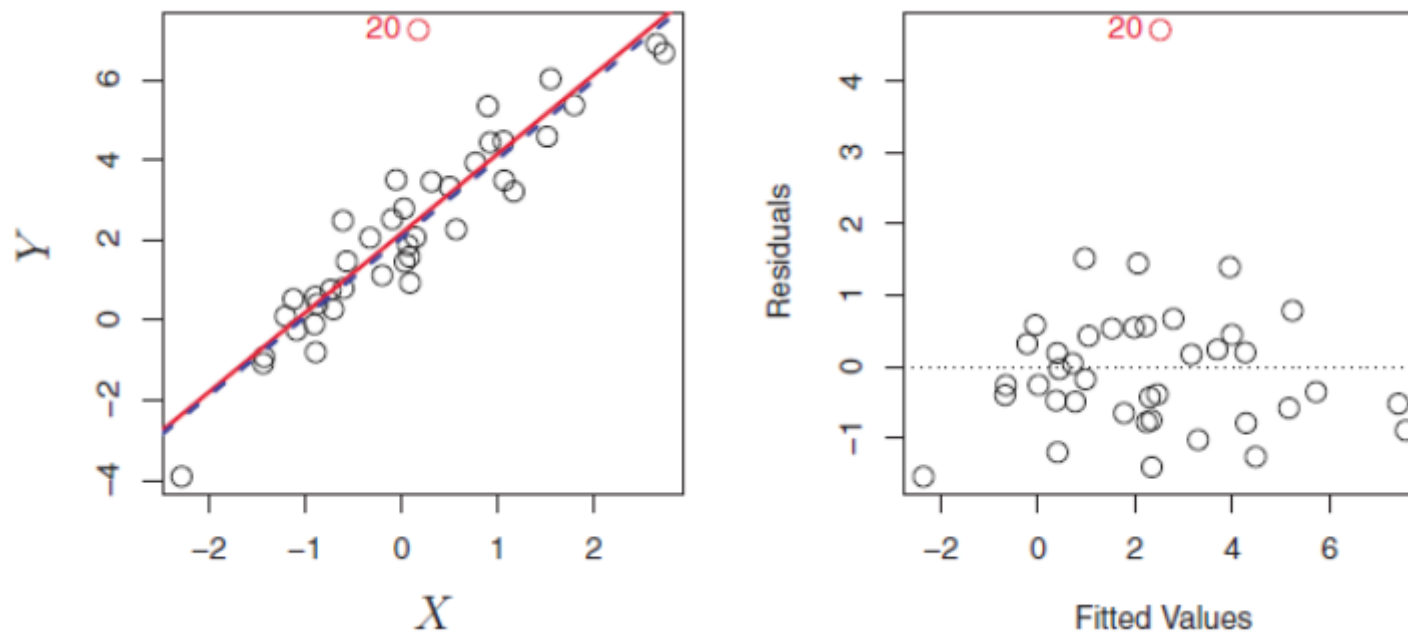
- Examinar el caso para descartar que sean un error o pertenezcan a una población diferente. Si es así, eliminarlos de nuestro estudio.
- Si el datos es correcto, debemos considerar si el modelo de regresión lineal es adecuado.

Una alternativa antes de descartar *outliers* es usar métodos alternativos, que ponderen las diferentes observaciones, dando menos importancia a los outliers.  
(Modelos de regresión robusta).

## 7. Problemas de un modelo de regresión

# ¿Qué hacer con outliers o influyentes?

Si un outlier no es un punto influyente, eliminarlo modifica mínimamente el resultado final, aunque mejora drásticamente la bondad del ajuste. En el siguiente ejemplo, se pasa de un valor de  $R^2 = 0.805$  a uno de  $R^2 = 0.892$  tan solo por eliminar el outlier.



## 7. Problemas de un modelo de regresión

# Colinealidad

Se dice que en un problema de regresión existe **colinealidad** cuando dos o más predictores son casi linealmente dependientes.

La existencia de colinealidad entre predictores produce problemas, ya que puede llevarnos a **introducir variables irrelevantes en nuestro modelo**. Para identificarlas se puede:

- Buscar pares de variables  $(X_i, X_j)$  tales que  $\rho(X_i, X_j)^2 \simeq 1$ .
- Hacer una regresión de cada variable explicativa sobre el resto y analizar el resultado de  $R^2$ . Si alguno es alto, podríamos estar ante **multicolinealidad**.
- Calcular el factor de inflación de la varianza:

$$VIF(X_k) = \frac{1}{1 - R_{X_j|X_{-j}}^2}$$

Usualmente, un  $VIF > 5$  indicad posible colinealidad. Un  $VIF > 10$  representa una muy alta colinealidad de  $X_j$  con el resto de variables.



## 7. Problemas de un modelo de regresión

### Ejercicio Final (+1pto)

- **Lab:** (Lab3.Rmd) con el dataset Auto realizamos un breve análisis exploratorio y ajustamos un modelo de regresión lineal múltiple sobre la variable mpg utilizando todas las variables.
- **Ejercicio:** Realiza un breve análisis exploratorio, ajusta un modelo de regresión lineal múltiple, comprueba si hay colinealidad, diagnóstícalo y evalúalo. El dataset será asignado durante la clase.

(Envía tu resolución a [rsanchez@afi.es](mailto:rsanchez@afi.es) antes de que lo resolvamos en clase.)



Afi Escuela

---

© 2021 Afi Escuela. Todos los derechos reservados.