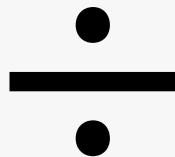
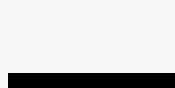
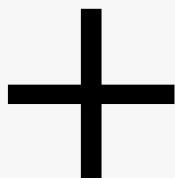


ITAM
Departamento de Estadística

Inferencia Estadística– Laboratorio #10
Modelo de Regresión Lineal Simple



Recordar:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad (1)$$

$$S_{XY} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (2)$$

$$S_{JJ} = \sum_{i=1}^n (j_i - \bar{j})^2 \text{ para } J = x, y \quad (3)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (4)$$

Y la recta de mínimos cuadrados es:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (5)$$

1. Conteste cada una de las siguientes preguntas:

- (a) ¿Qué es un modelo de regresión lineal simple?
(b) ¿Cuáles son los objetivos específicos que persigue este modelo?

(a) Es un modelo de probabilidad condicional que describe a una variable aleatoria Y dada una variable explicativa (X). El modelo supone que: $Y|X=x \sim N(\beta_0 + \beta_1 x, \sigma^2)$

(b) Los objetivos específicos del modelo de regresión es usar datos muestrales que permitan estimar los parámetros β_0, β_1 para ajustar una recta a los datos y con ellos explicar el comportamiento de Y .

Análisis estructural del modelo (estimadores puntuales, IC's, interpretación e inferencia, pronósticos)

2. Si $\hat{\beta}_0$ y $\hat{\beta}_1$ son las estimaciones de MCO para un modelo de regresión lineal simple. Demuestre que la recta de MCO siempre pasa por el punto (\bar{x}, \bar{y}) .

Sabemos que la recta de MCO está dada por $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$y \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Queremos minimizar la suma de los cuadrados del error que es lo mismo que suma de los cuadrados residuales

$$\text{i.e. } SCE = SCR = \sum_1^n (y_i - \hat{y}_i)^2 = \sum_1^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

$$\bullet \frac{\partial SCR}{\partial \hat{\beta}_0} = -2 \sum_1^n [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i] = -2 \left[\sum_1^n y_i - n \hat{\beta}_0 - \hat{\beta}_1 \sum_1^n x_i \right] = 0 \rightarrow (1)$$

$$\bullet \frac{\partial SCR}{\partial \hat{\beta}_1} = -2 \left[\sum_1^n [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i] x_i \right] = -2 \left[\sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2 \right] = 0 \rightarrow (2)$$

De (2)

$$\sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 \stackrel{\downarrow}{\Rightarrow} \frac{1}{n} \left[\sum y_i - \hat{\beta}_1 \sum x_i \right] \sum x_i + \hat{\beta}_1 \sum x_i^2$$

$$\Rightarrow \frac{1}{n} \left[\sum x_i \sum y_i \right] - \frac{\hat{\beta}_1}{n} \left(\sum x_i \right)^2 + \hat{\beta}_1 \sum x_i^2$$

$$\Rightarrow \frac{1}{n} \left[\sum x_i \sum y_i \right] + \hat{\beta}_1 \left[\sum x_i^2 - \frac{1}{n} \left(\sum x_i \right)^2 \right]$$

$$\therefore \hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n} \left[\sum x_i \sum y_i \right]}{\sum x_i^2 - \frac{1}{n} \left(\sum x_i \right)^2} = \frac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\text{Desarrollando (1)} \quad \sum y_i = n \hat{\beta}_0 + \hat{\beta}_1 \sum x_i \Rightarrow \hat{\beta}_0 = \frac{1}{n} \left[\sum y_i - \hat{\beta}_1 \sum x_i \right] \\ = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

\therefore La recta $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ Siempre pasa por el punto (\bar{x}, \bar{y})

3. Ajuste la recta de MCO. Para los siguientes datos:

y	3.0	2.0	1.0	1.0	0.5
x	-2.0	-1.0	0.0	1.0	2.0

Recordar: $\hat{\beta}_1 = \frac{\sum xy}{\sum x^2} = \frac{\sum x_i y_i - \frac{1}{n} (\sum x_i \sum y_i)}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2}$; $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

La recta de MCO $\Rightarrow \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Obs	x	y	$x_i y_i$	x_i^2
1	-2	3	-6	4
2	-1	2	-2	1
3	0	1	0	0
4	1	1	1	1
5	2	0.5	1	4

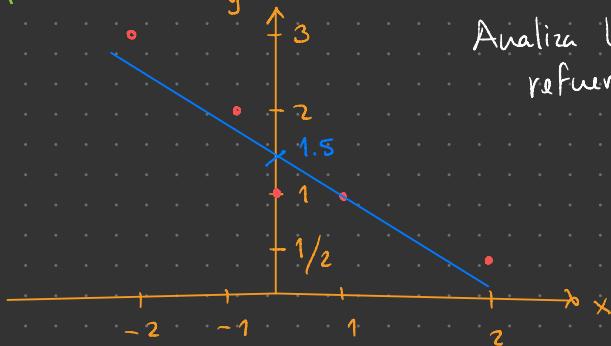
$$\begin{aligned} \bar{x} &= 0 \\ \bar{y} &= 1.5 \end{aligned}$$

$$\sum x_i = 0 \quad \sum y_i = 7.5 \quad \sum x_i y_i = -6 \quad \sum x_i^2 = 10$$

$$\Rightarrow \hat{\beta}_1 = \frac{-6 - (0 \times 1.5)}{10 - \frac{1}{5}(0)^2} = \frac{-6}{10} = -0.6$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1.5 + (-0.6)(0) = 1.5 \quad \therefore \quad \hat{y} = 1.5 - 0.6x$$

GRAFICAMOS y



Análiza la gráfica y
refuerza el resultado del
ejercicio 2

4. Es frecuente que a los auditores se les exija comparar el valor auditado (o de lista) de un artículo de inventario contra el valor en libros. Si una empresa está llevando su inventario y libros actualizados, debería haber una fuerte relación lineal entre los valores auditados y en libros. Una empresa muestreó diez artículos de inventario y obtuvo los valores auditado y en libros que se dan en la tabla siguiente. Ajuste el modelo a estos datos.

Artículo	Valor auditado (y_i)	Valor en libros (x_i)
1	9	10
2	14	12
3	7	9
4	29	27
5	45	47
6	109	112
7	40	36
8	238	241
9	60	59
10	170	167

Datos :

- $S_{xy} = 54243, S_{xx} = 54714$
- $\sum_1^n x_i = 720, \sum_1^n y_i = 721$

- (a) ¿Cuál es su estimación para el cambio esperado en valor auditado para un cambio de unidad en el valor de libros?
- (b) Si el valor en libros es $x = 100$, ¿qué usaría para estimar el valor auditado?

$$(1) \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{54243}{54714} \approx .9914$$

$$\bar{x} = \frac{1}{10}(720) = 72 ; \bar{y} = \frac{1}{10}(721) = 72.1$$

$$\hat{\beta}_0 = 72.1 - .9914(72) \approx .72 \therefore \hat{y} = .72 + .9914x$$

(a) "Cambio esperado en \hat{y} para un cambio en x "

$$\underline{\hat{\beta}_1 \approx .99}$$

$$(b) \text{ Si } x=100 \Rightarrow \hat{y} = .72 + .9914(100) \approx 99.72$$

5. Suponga que postulamos el siguiente modelo

$$y_i = \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n.$$

Encuentre el estimador de MCO de β_1 . [Hint: SCR = $\sum_1^n [y_i - \hat{y}_i]^2$]

$$\text{Caso } \epsilon_i \sim N(0, \sigma^2) \Rightarrow \begin{cases} \mathbb{E}[\epsilon_i] = 0 & \forall i \\ \text{Var}(\epsilon_i) = \sigma^2 & \forall i \\ \text{Cov}(\epsilon_i, \epsilon_j) = 0 & \forall i \neq j \end{cases}$$

$$\Rightarrow \hat{y}_i = \hat{\beta}_1 x_i \Rightarrow \text{SCR} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial \text{SCR}}{\partial \hat{\beta}_1} = -2 \left[\sum (y_i - \hat{\beta}_1 x_i) x_i \right] = 0$$

$$\Leftrightarrow \sum x_i y_i - \hat{\beta}_1 \sum x_i^2 = 0 \Leftrightarrow \hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

6. Encuentre:

- (a) $E[y_i], V[y_i], \text{Cov}(y_i, y_j)$
- (b) $E[\hat{\beta}_1], V[\hat{\beta}_1]$
- (c) $E[\hat{\beta}_0], V[\hat{\beta}_0]$
- (d) $\text{Cov}(\hat{\beta}_1, \hat{\beta}_0)$

Recordar

$$\epsilon_i \sim N(0, \sigma^2) \Rightarrow \begin{cases} \mathbb{E}[\epsilon_i] = 0 & \forall i \\ \text{Var}(\epsilon_i) = \sigma^2 & \forall i \\ \text{Cov}(\epsilon_i, \epsilon_j) = 0 & \forall i \neq j \end{cases}$$

$$(a) \quad \mathbb{E}[y_i] = \mathbb{E}[\beta_0 + \beta_1 x_i + \epsilon_i] = \mathbb{E}[\beta_0 + \beta_1 x_i] + \mathbb{E}[\epsilon_i] \xrightarrow{0} \beta_0 + \beta_1 x_i \quad \forall i$$

$$V[y_i] = V[\beta_0 + \beta_1 x_i + \epsilon_i] = V[\beta_0 + \beta_1 x_i] + V[\epsilon_i] = \sigma^2 \quad \text{Recordar si } c \in \mathbb{R}$$

$$\begin{aligned} \text{Cov}(y_i, y_j) &= \text{Cov}[(\beta_0 + \beta_1 x_i + \epsilon_i), (\beta_0 + \beta_1 x_j + \epsilon_j)] \quad \text{Cov}(c, Y) = 0 \\ &= \text{Cov}(\beta_0, \beta_0) + \text{Cov}(\beta_0, \beta_1 x_j) + \text{Cov}(\beta_0, \epsilon_j) + \text{Cov}(\beta_1 x_i, \beta_0) \\ &\quad + \text{Cov}(\beta_1 x_i, \beta_1 x_j) + \text{Cov}(\beta_1 x_i, \epsilon_j) + \text{Cov}(\epsilon_i, \beta_0) + \text{Cov}(\epsilon_i, \beta_1 x_i) + \text{Cov}(\epsilon_i, \epsilon_j) \\ &= 0 \end{aligned}$$

$$\text{Definición } \text{Cov}(X, Y) = \mathbb{E}[(X - E(X))(Y - E(Y))]$$

$$(b) \text{ Notemos que podemos reescrever } \hat{\beta}_1 = \frac{\sum x_i y_i - \bar{x}\bar{y}}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) y_i}{S_{xx}}$$

$$\mathbb{E}\{\hat{\beta}_1\} = \mathbb{E}\left[\frac{\sum (x_i - \bar{x}) y_i}{S_{xx}}\right] = \frac{1}{S_{xx}} \sum (x_i - \bar{x}) \mathbb{E}[y_i] = \frac{1}{S_{xx}} \sum (x_i - \bar{x}) \underbrace{\mathbb{E}[\beta_0 + \beta_1 x_i]}_{(\beta_0 + \beta_1 \bar{x})}$$

$$= \frac{\beta_0 \sum (x_i - \bar{x})}{S_{xx}} + \beta_1 \frac{\sum (x_i - \bar{x}) x_i}{S_{xx}} = \beta_1$$

$$\text{V}\{\hat{\beta}_1\} = \text{V}\left[\frac{\sum (x_i - \bar{x}) y_i}{S_{xx}}\right] = \frac{1}{S_{xx}^2} \sum \text{V}[(x_i - \bar{x}) y_i]$$

$$= \frac{1}{S_{xx}^2} \sum (x_i - \bar{x})^2 \text{V}(y_i) = \sigma^2 / S_{xx}$$

$$(c) \mathbb{E}\{\hat{\beta}_0\} = \mathbb{E}\{\bar{y} - \hat{\beta}_1 \bar{x}\} = \mathbb{E}\{\bar{y}\} - \mathbb{E}[\hat{\beta}_1] \bar{x} = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x}$$

$$= \beta_0$$

$$\text{V}\{\hat{\beta}_0\} = \text{V}(\bar{y} - b_1 \bar{x}) = \text{V}(\bar{y}) + \text{V}(b_1 \bar{x}) - 2 \text{Cov}(\bar{y}, b_1 \bar{x})$$

$$= \frac{\sigma^2}{n} + \bar{x}^2 \text{V}[b_1] - 2 \bar{x} \underbrace{\text{Cov}(\bar{y}, b_1)}_{=0} = \frac{\sigma^2}{n} + \bar{x}^2 \left[\frac{\sigma^2}{\sum x_i^2} \right]$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2} \right] = \sigma^2 \left[\frac{\sum x_i^2 + n \bar{x}^2}{n \sum x_i^2} \right] = \frac{\sigma^2}{n S_{xx}} \sum x_i^2$$

$$(d) \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{Cov}(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) = \text{Cov}(\bar{y}, \hat{\beta}_1) - \text{Cov}(\hat{\beta}_1 \bar{x}, \hat{\beta}_1)$$

$$= -\bar{x} \text{Cov}(\hat{\beta}_1, \hat{\beta}_1) = -\bar{x} \text{V}(\hat{\beta}_1) = -\frac{\bar{x} \sigma^2}{S_{xx}}$$

7. (a) Demuestre que $SCR = S_{yy} - \hat{\beta}_1 S_{xy}$

(b) Use el inciso anterior para demostrar que $SCR \leq S_{yy}$

$$\begin{aligned}(a) \quad SCR &= \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum (y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i)^2 \\&= \sum [(y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})]^2 = \sum [(y_i - \bar{y})^2 - 2\hat{\beta}_1(y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}_1^2(x_i - \bar{x})^2] \\&= (y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 \\&= (y_i - \bar{y})^2 - \hat{\beta}_1 \sum (x_i - \bar{x})(y_i - \bar{y}) \quad \leftarrow \because \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\&= S_{yy} - \hat{\beta}_1 S_{xy}\end{aligned}$$

$$(b) \quad \text{Sabemos que } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\Rightarrow SCR = S_{yy} - \frac{S_{xy}}{S_{xx}} S_{xy} = S_{yy} - \underbrace{\frac{S_{xy}^2}{S_{xx}}}_{\geq 0} \leq S_{yy}$$

8. Suponga que $Y_1, Y_2, \dots, Y_n \stackrel{\text{v.a.i.d}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2) \quad \forall i$. Demuestre que los estimadores de máxima probabilidad (MLE) de β_0, β_1 son iguales a los estimadores de MCO.

Las ecuaciones a las que llegamos por MCO son

$$[\hat{\beta}_0]: -2 \left[\sum_i y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_i x_i \right] = 0 \Leftrightarrow \sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i$$

$$[\hat{\beta}_1]: -2 \left[\sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2 \right] = 0 \Leftrightarrow \sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2$$

Ahora

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \bar{Y})^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2}$$

$$L(\beta_0, \beta_1) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2} = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i)^2}$$

$$\ell(\beta_0, \beta_1) = n \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial \ell}{\partial \beta_0} = -\frac{1}{2\sigma^2} \left[-2 \sum (y_i - \beta_0 - \beta_1 x_i) \right] = 0 \Leftrightarrow \sum y_i = \beta_0 + \beta_1 x_i \circ \sigma^2 > 0$$

$$\frac{\partial \ell}{\partial \beta_1} = -\frac{1}{2\sigma^2} \left\{ 2 \sum (y_i - \beta_0 - \beta_1 x_i) x_i \right\} = \frac{1}{\sigma^2} \left[\sum x_i y_i - \beta_0 \sum x_i - \beta_1 \sum x_i^2 \right] = 0$$

$$\Rightarrow \sum x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

Que es el mismo sistema de ecuaciones que el de MCO

∴ Concluimos que los estimadores $\hat{\beta}_0, \hat{\beta}_1$ de MCO SON MELÍS

laboratorio 10 Inferencia Estad Parte R

Luis Martinez

Otoño 20201

Vamos a correr un modelo de regresión lineal simple en R.

```
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.0 --
## v ggplot2 3.3.3     v purrr    0.3.4
## v tibble   3.1.0     v dplyr    1.0.4
## v tidyverse 1.1.1     v stringr  1.4.0
## v readr    1.3.1     v forcats 0.5.0

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()

library(ggplot2)
library(ggcormrplot)
library(dplyr)
library(ggpubr)
```

En este caso vamos a usar los datos *mtcars*, ya cargados a R y realizaremos inferencia sobre los mismos.

```
#Cargamos los datos
data("mtcars")
datos.coches<- mtcars
#Glimpse al data frame
head(mtcars)

##          mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4      21.0   6 160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21.0   6 160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710     22.8   4 108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive 21.4   6 258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360 175 3.15 3.440 17.02  0  0    3    2
## Valiant        18.1   6 225 105 2.76 3.460 20.22  1  0    3    1

# Manipulación de variables
x<- datos.coches$wt
y<- datos.coches$mpg

#Correlaciones
(cor(x, y, method="pearson"))

## [1] -0.8676594
```

```
(cor(x,y, method = "spearman"))

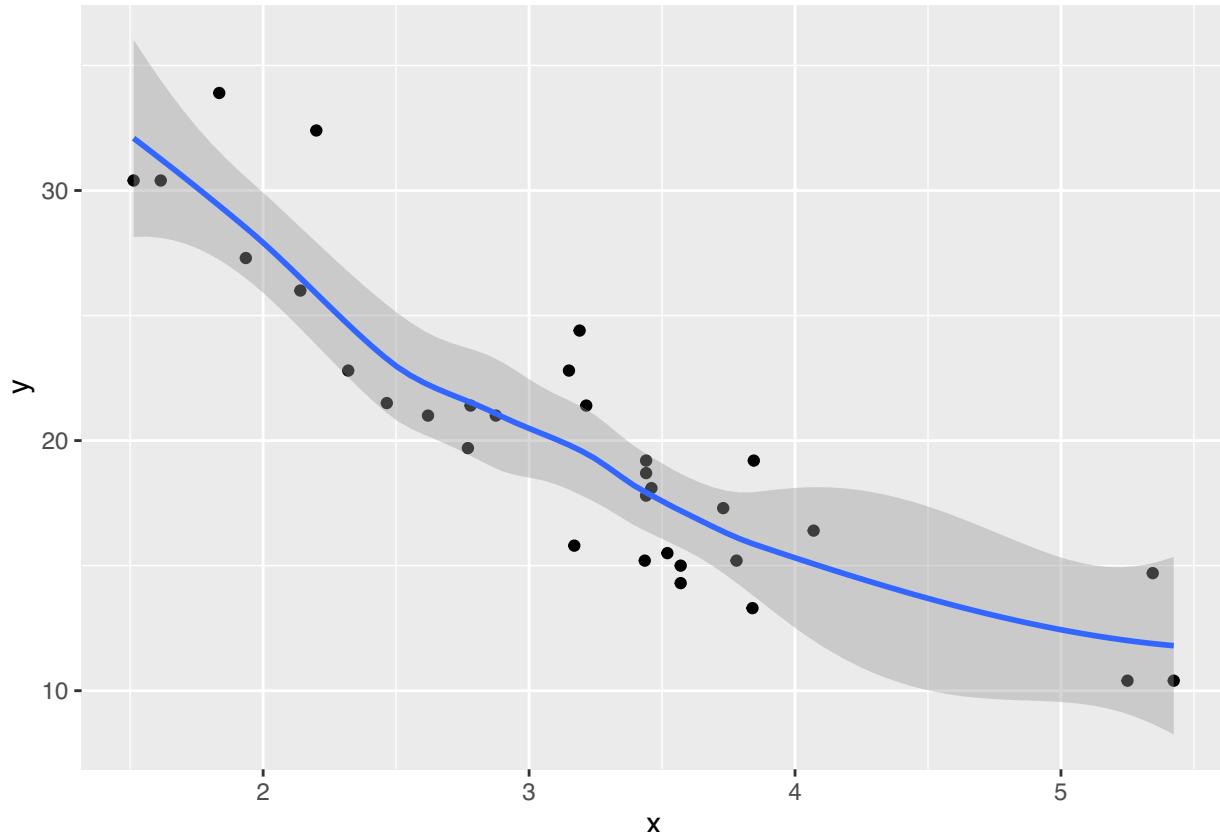
## [1] -0.886422

(cor(x,y, method = "kendall"))

## [1] -0.7278321

#Sugerencia de una relacion negativa entre miles per galon y weight
ggplot(datos.coches, aes(x,y))+geom_point()+stat_smooth()

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
#Construimos el modelo de regresion lineal simple
(modelo<- lm(y~x, data=datos.coches))
```

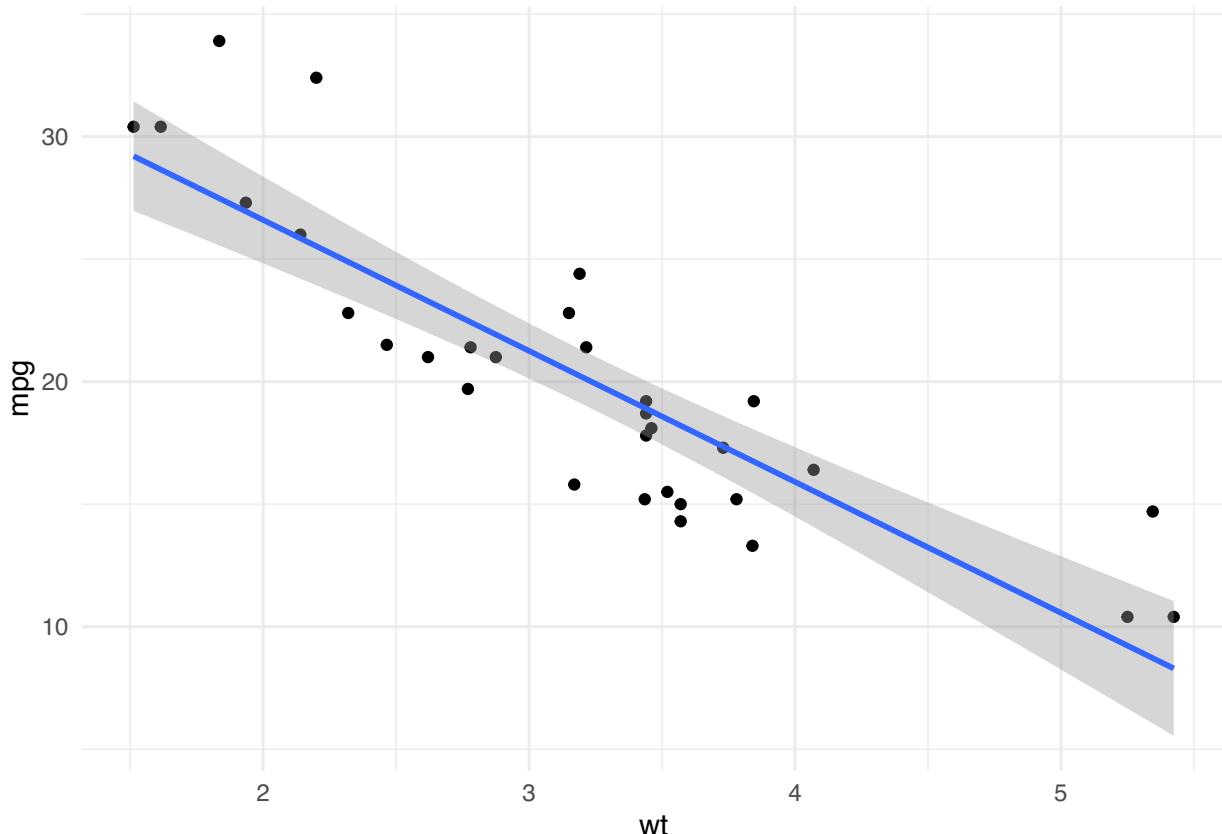
```
##
## Call:
## lm(formula = y ~ x, data = datos.coches)
##
## Coefficients:
## (Intercept)          x
##       37.285        -5.344
summary(modelo)
```

```
##
## Call:
## lm(formula = y ~ x, data = datos.coches)
##
```

```

## Residuals:
##      Min      1Q Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.2851    1.8776  19.858 < 2e-16 ***
## x           -5.3445    0.5591 -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
#Corremos la Regresion
ggplot(datos.coches)+ geom_point(aes(x= wt, y= mpg))+ stat_smooth(aes(x= wt, y= mpg), method= "lm", formula= y ~ x, se=TRUE)+theme_minimal()

```



```

#Intervalos de confianza para estimadores del modelo al 97.5%
confint(modelo)

```

```

##              2.5 %    97.5 %
## (Intercept) 33.450500 41.119753
## x          -6.486308 -4.202635

```

```

#Coeficiente de correlación R^2
#Este coeficiente mide cuanta proporcion del modelo es explicada por la

```

```
#regresion  
(R.cuadrada<- (cor(x, y, method="pearson"))^2)  
## [1] 0.7528328
```