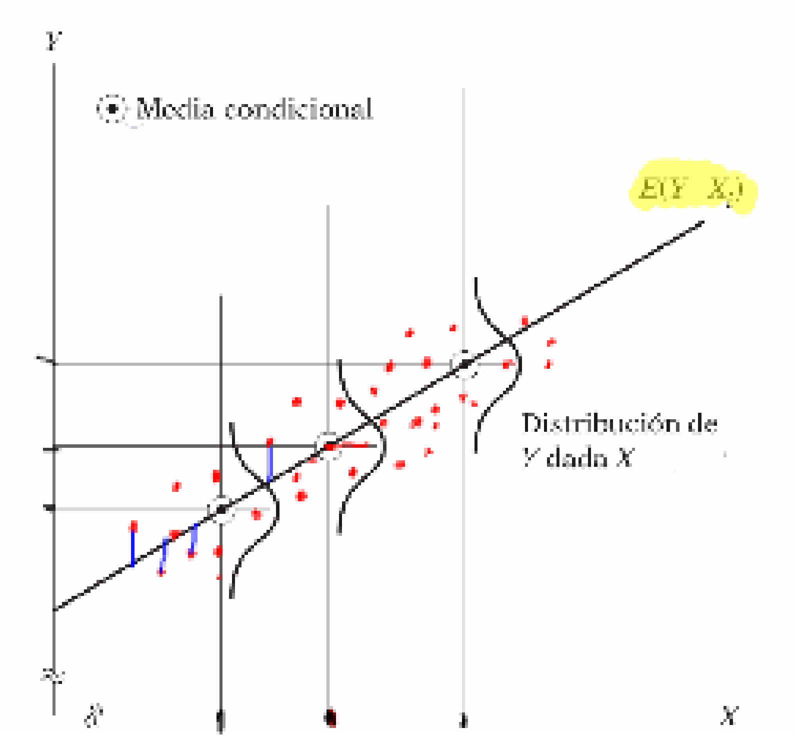
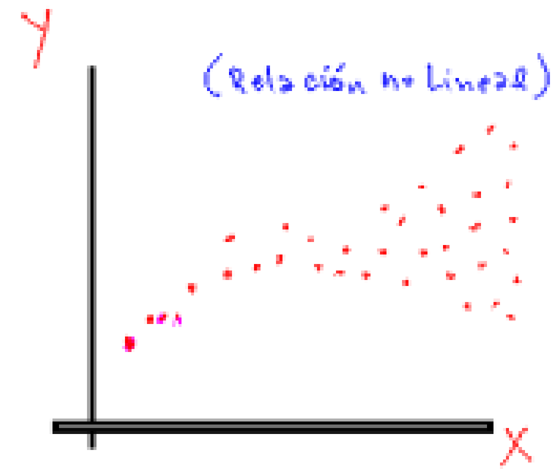
Regresión Lineal

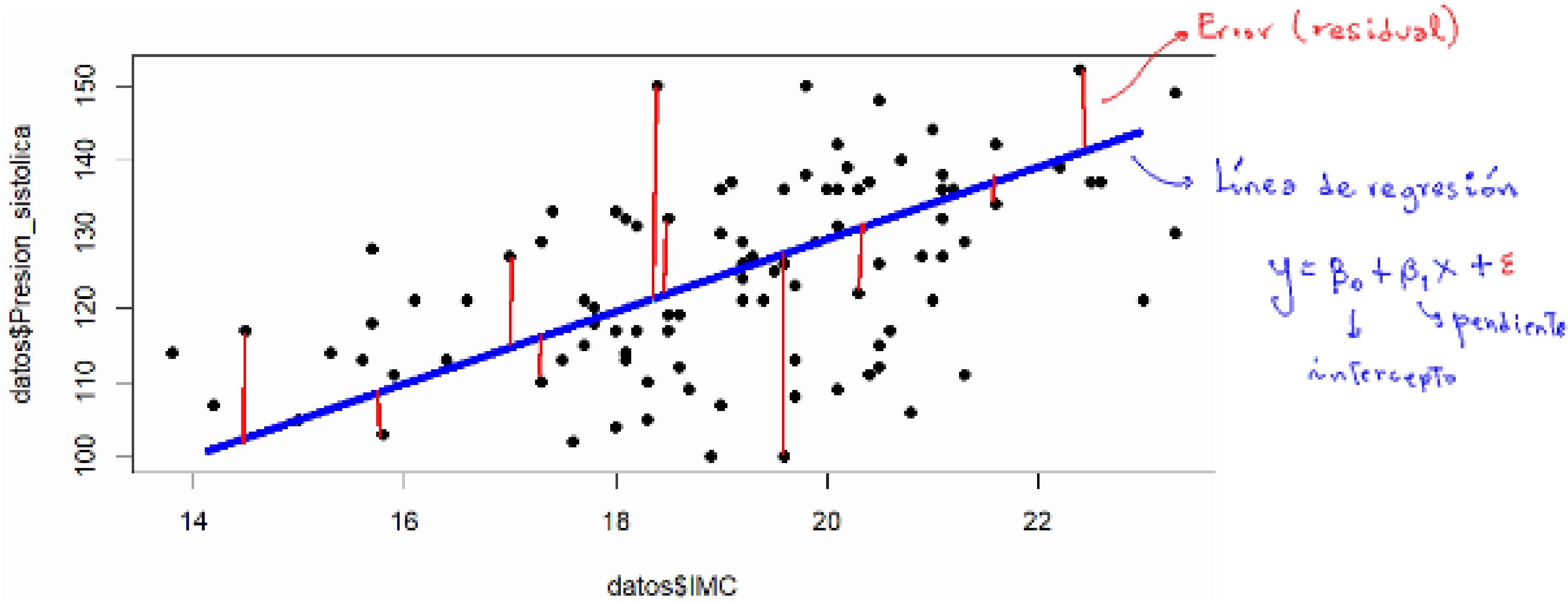
Y = f(X)
Y : Variable respuesta, predicha, dependiente. Target. → cuantitativa
X : Variable explicativa, predictora, independiente. Feature o atributo. → cuantitativa o cualitativa



$\mu_{Y|X_i}$, $\mu_{Y|X=3}$

$Y|X \sim N$





Error (residual)
Línea de regresión
$y = \beta_0 + \beta_1 x + \varepsilon$
intercepto
pendiente
datos$Presion_sistolica
datos$IMC
100
110
120
130
140
150
14
16
18
20
22

Presión sistólica

IMC

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = \mu_i + \epsilon_i \qquad i = 1, ..., 100$$

intercepto

pendiente

$$\mu_i = \beta_0 + \beta_1 X_i$$ → poblacional (β_0, β_1 son parámetros)

$$\hat{\mu}_i = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$ → muestral ($\hat{\beta}_0, \hat{\beta}_1$ son estimadores)

$$\hat{\mu}_i = \hat{Y}_i = 64.5577 + 3.0957 X_i$$

$\hat{\beta}_0$: cuando $X_i = 0 \Rightarrow \hat{\mu}_i = 64.5577 \rightarrow$ La presión sistólica media cuando IMC = 0. **No tiene sentido**

$\hat{\beta}_1$: $\hat{Y}_i = 64.5577 + 3.0957 X_i$

$X_i \Rightarrow X_i + 1$

$\hat{Y}_i^* = 64.5577 + 3.0957(X_i + 1)$

$\hat{Y}_i^* - \hat{Y}_i = 3.0957 = \hat{\beta}_1$

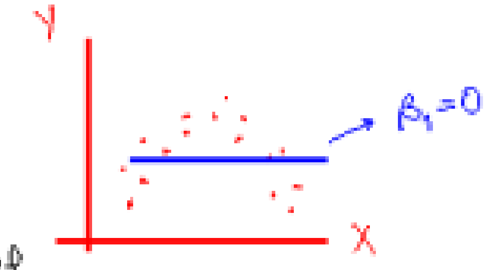
$$\left(\bar{X} - t_{(1-\alpha/2;n-1)}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{(1-\alpha/2;n-1)}\frac{s}{\sqrt{n}}\right)$$

$$IC(\beta_j) = \hat{\beta}_j \mp t_{1-\alpha/2,n-1}\, s_{\hat{\beta}_j}$$ → R

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$\mu = \beta_0 + \beta_1 X$$

- $\beta_1 = 0 \rightarrow$ X e Y <u>no</u> tienen relación <u>lineal</u>
- $\beta_1 \neq 0 \rightarrow$ X e Y sí tienen relación lineal



Regresión GLReg SCReg CMReg Fcalc

```
> modelo |> aov() |> summary()
            Df Sum Sq Mean Sq F value   Pr(>F)    
IMC          1   3924    3924   31.86 1.61e-07 ***
Residuals   98  12068     123                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GLE SCE CME

$\hookrightarrow$ pvalor $= 1.61 \times 10^{-7} = 0.000000161 < \alpha$

$\Downarrow$

Se rechaza H_0

$\Downarrow$

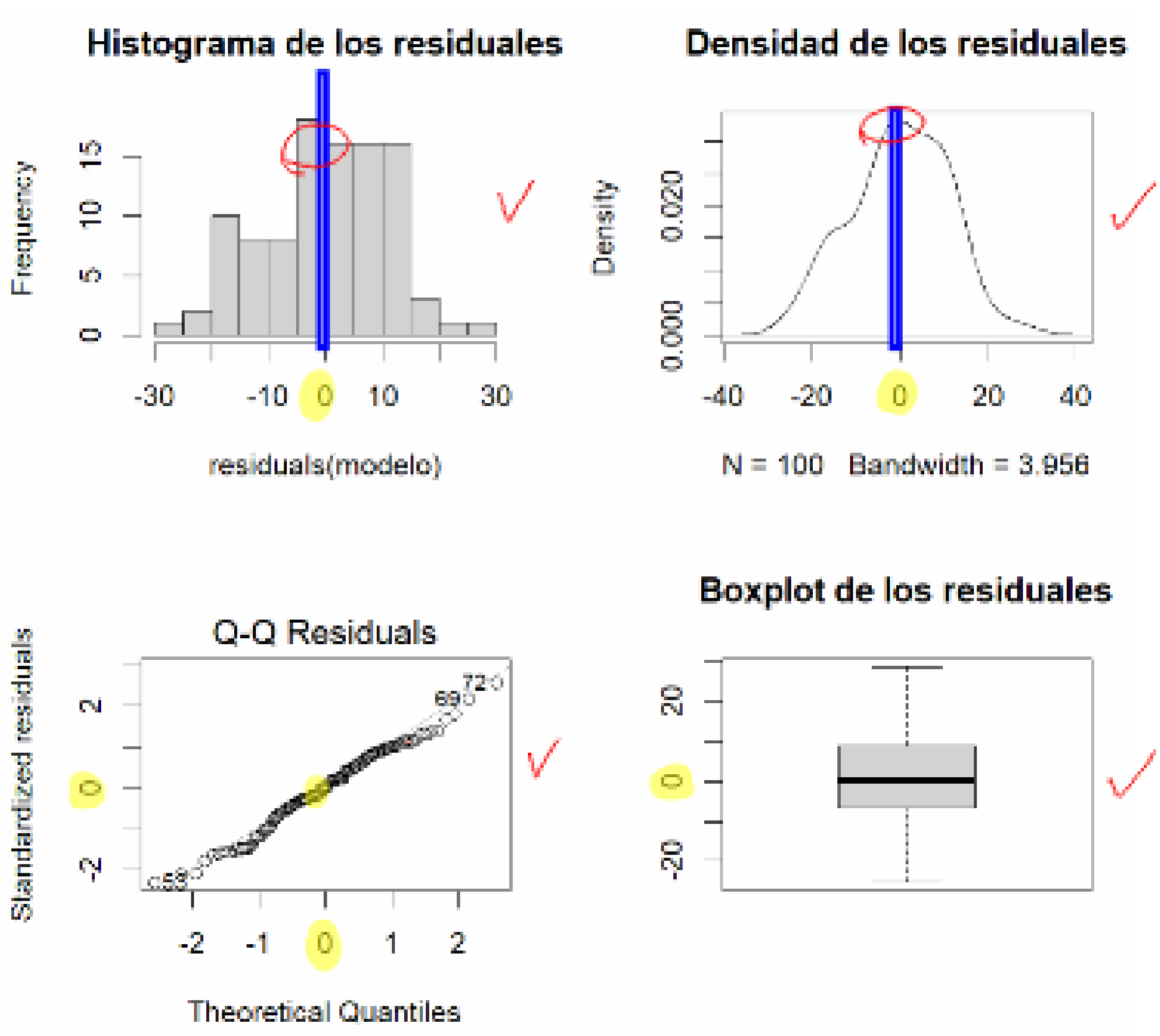
X e Y sí tienen relación lineal

$$Y = \overbrace{\beta_0 + \beta_1 X}^{\text{REGRESIÓN}} + \varepsilon$$

$\to$ Error recoge los efectos de otras variables no observadas

$$E(\varepsilon) = \mu_\varepsilon = 0$$

$$V(\varepsilon) = \sigma_\varepsilon^2 \text{ Constante}$$


Histograma de los residuales
Frequency
0
5
10
15
-30
-10
0
10
30
residuals(modelo)
Densidad de los residuales
Density
0.000
0.020
-40
-20
0
20
40
N = 100 Bandwidth = 3.956
Q-Q Residuals
Standardized residuals
-2
0
2
72
69
53
-2
-1
0
1
2
Theoretical Quantiles
Boxplot de los residuales
-20
0
20

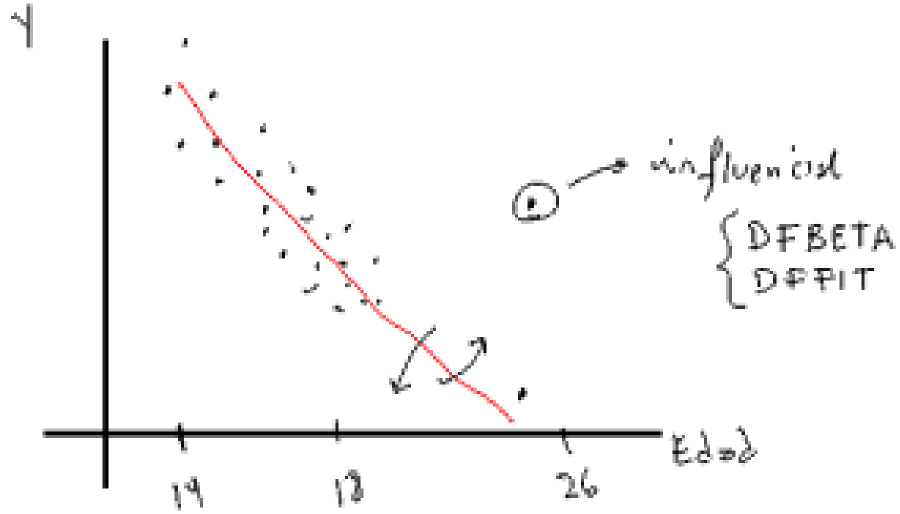
$$E(\varepsilon)=0$$

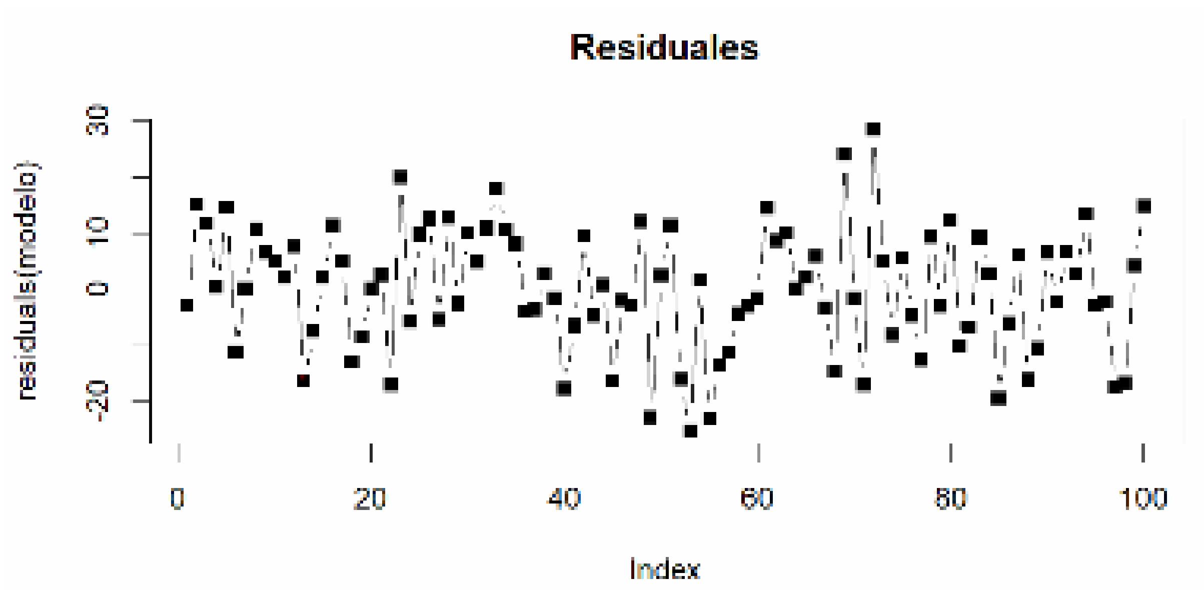
$$E(Y-\hat{Y})=0$$

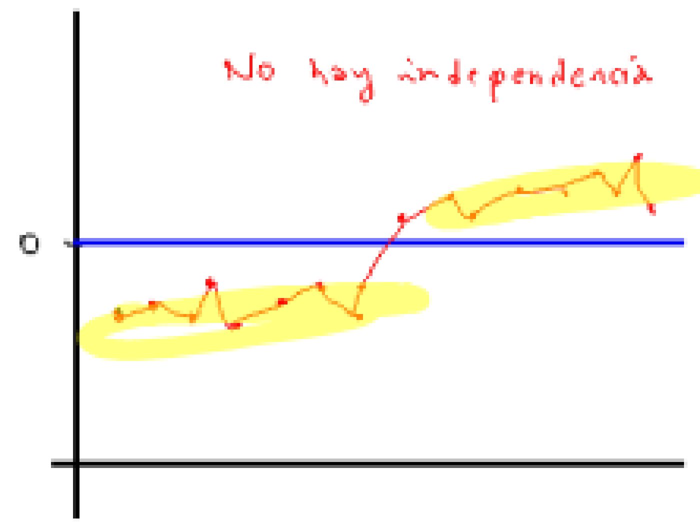
$$E(Y-\hat{\beta}_0-\hat{\beta}_1X)=0$$

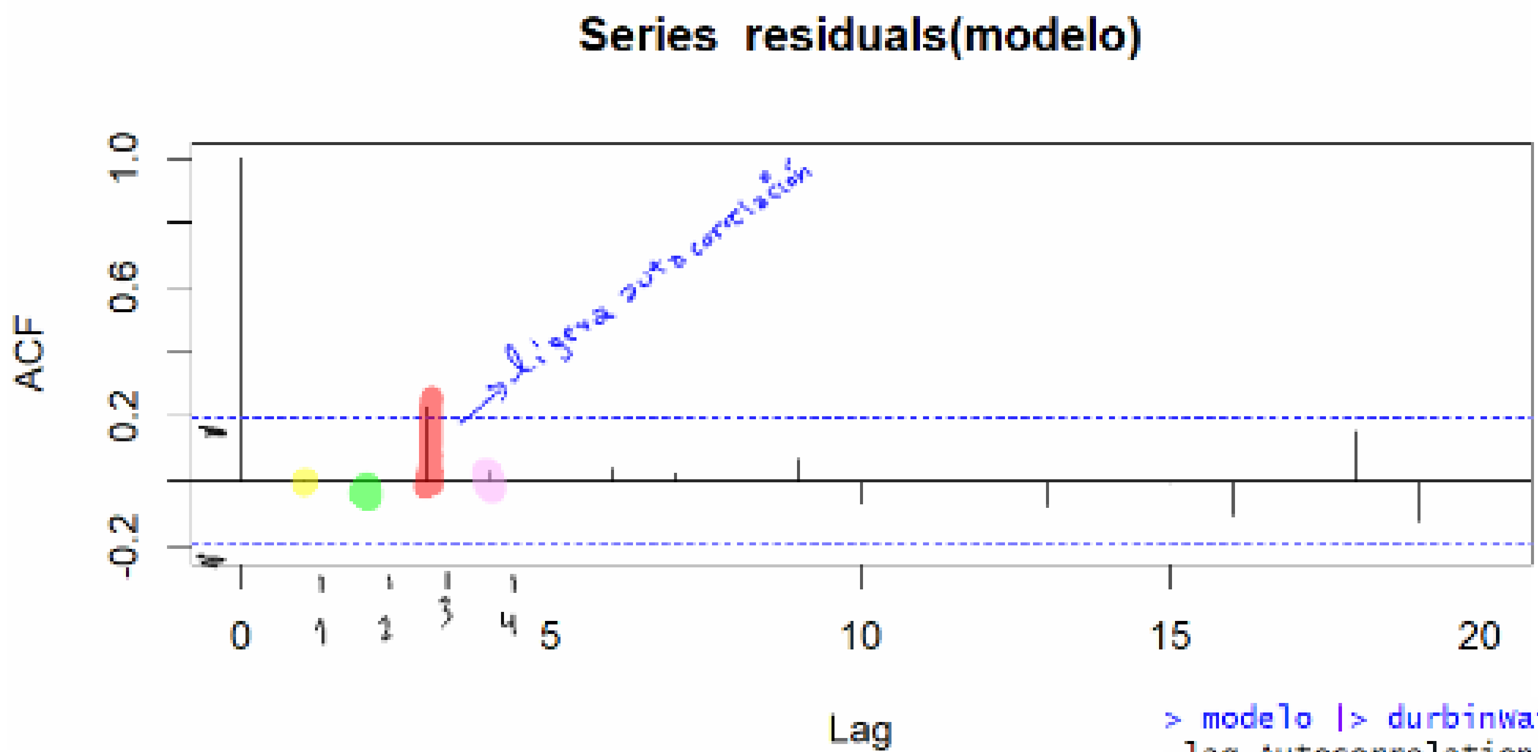
$$\downarrow \qquad \downarrow$$

$$E(\cdot\ \cdot\cdot\ - \qquad\qquad)=0$$




Residuales
residuals(modelo)
-20
0
10
30
0
20
40
60
80
100
Index


No hay independencia
0



$\to \hat{\rho}_1 = Cor(\varepsilon_t, \varepsilon_{t-1})$

$\to \hat{\rho}_2 = Cov(\varepsilon_t, \varepsilon_{t-2})$

$\to \hat{\rho}_3 = Cor(\varepsilon_t, \varepsilon_{t-3})$

```
> modelo |> durbinWatsonTest(max.lag = 5)
 lag Autocorrelation D-W Statistic p-value
   1     0.005751365      1.969609   0.934
   2    -0.036473991      2.033392   0.790
   3     0.228924274      1.467707   0.014
   4     0.026582705      1.846881   0.648
   5    -0.086195271      2.054406   0.506
 Alternative hypothesis: rho[lag] != 0
```

¿?

Estimación de la media
Predicción individual

Puntual — Intervalar

$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 X \longrightarrow$ Intervalo de confianza : $IC(\mu)$

$\tilde{y} = \hat{\beta}_0 + \hat{\beta}_1 X \longrightarrow$ Intervalo de predicción : $IP(y) \longrightarrow$ más ancho

```
> modelo |> predict(data.frame(IMC = c(18,20,23)))
       1        2        3 
120.2800 126.4714 135.7584 
> modelo |> predict(data.frame(IMC = c(18,20,23)),
+                   level = 0.95, interval = "confidence")
       fit      lwr      upr
1 120.2800 117.7909 122.7691
2 126.4714 124.0459 128.8968
3 135.7584 130.9439 140.5729
```

$IC(\mu)$

```
> modelo |> predict(data.frame(IMC = c(18,20,23)))
       1        2        3 
120.2800 126.4714 135.7584 
> modelo |> predict(data.frame(IMC = c(18,20,23)),
+                   level = 0.95, interval = "prediction")
       fit       lwr      upr
1 120.2800  98.11787 142.4421
2 126.4714 104.31629 148.6264
3 135.7584 113.21636 158.3005
```

MÁS ANCHO

$IP(y)$

$$\hat{V}(\hat{\mu}) = \hat{V}(\hat{y}) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x - \overline{x})^2}{SCX} \right)$$

$$\hat{V}(\hat{y}_0) = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{SCX} \right)$$

$$V(\hat{\mu}) < V(\hat{y}_0)$$