```r
modelo = lm(Sueldo ~ ., datos)
(beta  = coef(modelo))
```

F = 0

M = 1

```
(Intercept)   Educacion       SexoM        Edad          X4
-0.73053615  0.15695057  0.82211261  0.10462774 -0.04537505
```

miles S/   miles S/ / años   años   miles S/   miles S/ / año   años   miles S/

$$\hat{Y} = -0.731 + 0.157\,\text{Educación} + 0.822\,\text{SexoM} + 0.105\,\text{Edad} - 0.045\,X_4$$

$\hat{\beta}_0 = -0.731$: No tiene interpretación lógica porque Edad $\neq 0$ y el sueldo no puede ser negativo

$\hat{\beta}_1 = 0.157$: por cada año adicional de educación, el sueldo promedio se incrementa en 0.157 miles de soles, manteniendo las demás características en valores fijos

$$\hat{Y} = -0.73 + 0.157\,\text{Educ} + 0.822\,\text{SexoM} + 0.105\,\text{Edad} - 0.045\,X_4$$

$$-\quad \hat{Y}^* = -0.73 + 0.157\,(\text{Educ}+1) + 0.822\,\text{SexoM} + 0.105\,\text{Edad} - 0.045\,X_4$$

$$\hat{Y}^* - \hat{Y} = 0.157$$

$\hat{\beta}_3 = 0.105$: por cada año adicional de edad, el sueldo promedio aumenta en 0.105 miles de soles, manteniendo constantes la educación, el sexo y X_4.

$\hat{\beta}_2 = 0.822$: el sueldo promedio de los hombres (Sexo = 1) supera en 0.822 miles de soles al de las mujeres (Sexo = 0), manteniendo ctes la edad, la educación y X_4.

Estimación intervalar

Si g_{ii} es el elemento i-ésimo de la diagonal de $(X'X)^{-1}$, entonces:

K = # coeficientes

p = # variables

$$IC(\beta_i) = \hat{\beta}_i \pm t_{1-\alpha/2,n-k}\hat{\sigma}\sqrt{g_{ii}}$$

```
G    = solve(t(X)%*%X)
g    = G |> diag()
n    = datos |> nrow()
k    = beta |> length()   # coef
valt = qt(0.975, 24-5)
beta - valt*sigma*sqrt(g)
```

```
(Intercept)   Educacion       SexoM        Edad          X4
-2.58809073  0.02638407 -0.18289392  0.04174457 -0.12999986
```

```
beta + valt*sigma*sqrt(g)
```

```
(Intercept)   Educacion       SexoM        Edad          X4
 1.12701843  0.28751706  1.82711914  0.16751091  0.03924976
```

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i} + \varepsilon_i, \quad i = 1, \ldots, n$$

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \longrightarrow$ $Y_i = \beta_0 + \varepsilon$: las X no influyen linealmente en Y

H_1 : Al menos un $\beta_j \neq 0$

Variabilidad		GL	SC	CM	Fcalc
	Regresión	$K-1$	SCReg	CMReg	CMReg/CME
	Error	$n-K$	SCE	CME	

vector de datos

vector de 1s

$$SC_{Total} = (\mathbf{y} - \overline{y}\mathbf{1})'(\mathbf{y} - \overline{y}\mathbf{1}) \qquad \overline{y} = \frac{1}{n}\mathbf{1}'\mathbf{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\mathbf{1}' = (1 \ 1 \ \ldots \ 1)$$

$$(1 \ 1 \ \ldots \ 1)\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^{n} y_i$$

vector de valores estimados

$$SC_{Reg} = (\hat{\mathbf{y}} - \overline{y}\mathbf{1})'(\hat{\mathbf{y}} - \overline{y}\mathbf{1})$$

$$SC_{Error} = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$$

$$SC_{Total} = (\mathbf{y} - \bar{y}\mathbf{1})'(\mathbf{y} - \bar{y}\mathbf{1}) \qquad \bar{y} = \frac{1}{n}\mathbf{1}'\mathbf{y}$$

$$\boxed{SC_{Reg}} = (\hat{\mathbf{y}} - \bar{y}\mathbf{1})'(\hat{\mathbf{y}} - \bar{y}\mathbf{1})$$

$$SC_{Error} = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$$

Sumas de cuadrados

```
uno = rep(1,n)
ybarra  = as.numeric(1/n*(uno)%*%y)
yhat    = predict(modelo)
(SCTotal = t(y - ybarra*uno)%*%(y - ybarra*uno))
```

```
         [,1]
[1,] 103.0379
```

```
(SCReg   = t(yhat - ybarra*uno)%*%(yhat - ybarra*uno))
```

```
         [,1]
[1,] 78.57194
```

```
(SCError = t(y - yhat)%*%(y-yhat))
```

```
         [,1]
[1,] 24.46592
```


Estadístico de prueba

```
(Fcalc = CMReg / CMError)
```

```
         [,1]
[1,] 15.25455
```

Valor crítico:

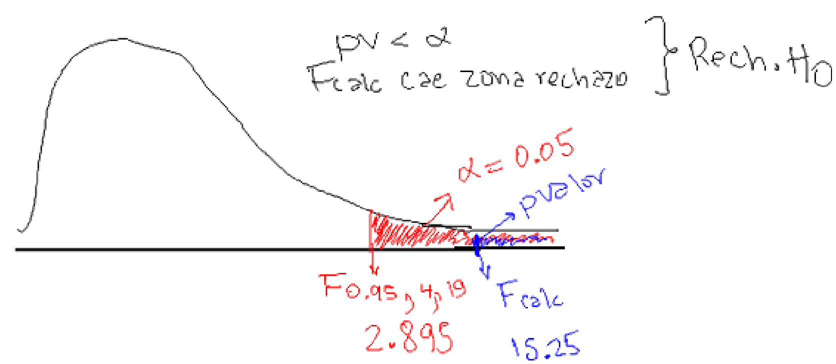
```
qf(0.95,4,19)
```

```
[1] 2.895107
```

Pvalor:

```
pf(Fcalc,4,19, lower.tail = F)
```

```
             [,1]
[1,] 9.638667e-06
```



Prueba de hipótesis individual

$H_0 : \beta_i = 0$
$H_1 : \beta_i \neq 0$

```r
library(broom)
modelo |> tidy()
```

```
# A tibble: 5 x 5
```

		term	estimate	std.error	statistic	p.value
		<chr>	<dbl>	<dbl>	<dbl>	<dbl>
$\hat{\beta}_0$	1	(Intercept)	-0.731	0.887	-0.823	0.421
$\hat{\beta}_1$	2	Educacion	0.157	0.0624	2.52	0.0210
$\hat{\beta}_2$	3	SexoM	0.822	0.480	1.71	0.103
$\hat{\beta}_3$	4	Edad	0.105	0.0300	3.48	0.00249
$\hat{\beta}_4$	5	X4	-0.0454	0.0404	-1.12	0.276

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$\alpha = 0.05$

pv = 0.021

Se rechaza H_0

Los años de educación tienen influencia lineal sobre el sueldo

$H_0 : \beta_4 = 0$

$H_1 : \beta_4 \neq 0$

$\alpha = 0.05$

pv = 0.276

No se rechaza H_0

X_4 no tiene influencia lineal sobre el sueldo

```r
library(broom)
modelo |> tidy()
```

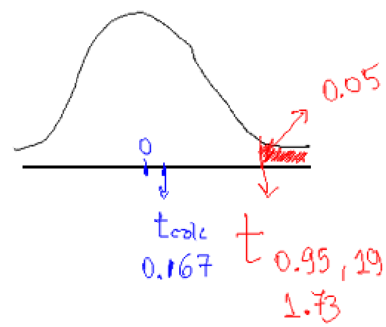
```
# A tibble: 5 x 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)  -0.731     0.887     -0.823 0.421
2 Educacion     0.157     0.0624     2.52  0.0210
3 SexoM         0.822     0.480      1.71  0.103
4 Edad          0.105     0.0300     3.48  0.00249
5 X4           -0.0454    0.0404    -1.12  0.276
```

$H_0: \beta_3 \leq 0.1$

$H_1: \beta_3 > 0.1$

$\alpha = 0.05$

$$t_{calc} = \frac{\hat{\beta}_3 - \beta_3}{S_{\hat{\beta}_3}} = \frac{0.105 - 0.1}{0.03} = 0.167$$



No se rechaza H_0

No existe evidencia estadística para afirmar que por cada año adicional de edad, el sueldo promedio se incremente en más de 0.1 miles de soles

	R^2	R^2_{aj}
	siempre sube	puede disminuir si la variable que se añade no contribuye al modelo
$Y \sim X_1$	0.47	0.465
$Y \sim X_1 + X_2$	0.76	0.75
$Y \sim X_1 + X_2 + X_3$	0.77	0.74

$$AIC = -2ln(L) + 2k$$

k → # coeficientes

L → Verosimilitud (Likelihood)

Busca modelos parsimoniosos

	Nº Variables	L	AIC
i)	2	30	$-2\ln(30) + 2\times2 = -6.8 + 4 = -2.8$
ii)	5	45	$-2\ln(45) + 2\times5 = -7.6 + 10 = 2.4$
iii)	9	48	$-2\ln(48) + 2\times9 = -7.7 + 18 = 10.3$

Estimación de la media de la respuesta

Para un vector de valores explicativos $\mathbf{x} = (1 \quad x_1 \quad x_2 \quad \dots \quad x_p)$

$$\hat{\mu} = \hat{y} = \mathbf{x}'\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_p X_p$$

Estimación del sueldo medio para una mujer de 35 años de edad, con 12 años de educación, y tomando $X_4 = 1$.

Puntual:

Educ sexo Edad X4

```
x0 = c(1, 12, 0, 35, 1)
yest = x0%*%beta
yest
```

el orden de los datos debe ser el mismo que se definió en el modelo

Todas las variables

```
modelo = lm(Sueldo ~ ., datos)
(beta  = coef(modelo))
```

```
        [,1]
[1,] 4.769467
```

Intervalo de confianza:

vector de predicción (x_0)

$$IC(\mu|\mathbf{x}) = \hat{\mu} \pm t_{1-\alpha/2,n-k}\sqrt{\hat{\sigma}^2\mathbf{x}'(X'X)^{-1}\mathbf{x}}$$

Matriz del modelo

```
predict(modelo,
        data.frame(Educacion = 12,
                   Sexo = "F",
                   Edad = 35,
                   X4 = 1),
        interval = "confidence")
```

```
       fit      lwr      upr
1 4.769467 3.944129 5.594804
```

```
predict(modelo,
        data.frame(Educacion = c(12,7),
                   Sexo = c("F","M"),
                   Edad = c(35,28),
                   X4 = c(1,-2)),
        interval = "confidence")
```

```
       fit      lwr      upr
1 4.769467 3.944129 5.594804
2 4.210557 3.177470 5.243645
```

Intervalo de confianza:

$$IC(\mu|\mathbf{x}) = \hat{\mu} \pm t_{1-\alpha/2,n-k}\sqrt{\hat{\sigma}^2\mathbf{x}'(X'X)^{-1}\mathbf{x}}$$

Intervalo de predicción:

$$IP(y|\mathbf{x}) = \hat{y}_0 \pm t_{1-\alpha/2,n-k}\sqrt{\hat{\sigma}^2\left(1 + \mathbf{x}'(X'X)^{-1}\mathbf{x}\right)}$$

$$IC(\mu|\mathbf{x}) = (3.94, 5.59)$$

$$IP(y|\mathbf{x}) = (2.255, 7.284)$$

Residuales

El vector de residuales es:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\overset{\text{matriz identidad}}{I} - \overset{\text{matriz hat}}{H})\mathbf{y} = I\mathbf{y} - H\mathbf{y}$$

donde:

$$H = \underset{n\times(p+1)}{X}(\underbrace{\underset{(p+1)\times n}{X'}\,\underset{n\times(p+1)}{X}}_{(p+1)\times(p+1)})^{-1}\underset{(p+1)\times n}{X'}$$

$X_{n\times(p+1)}$

$H_{n\times n}$