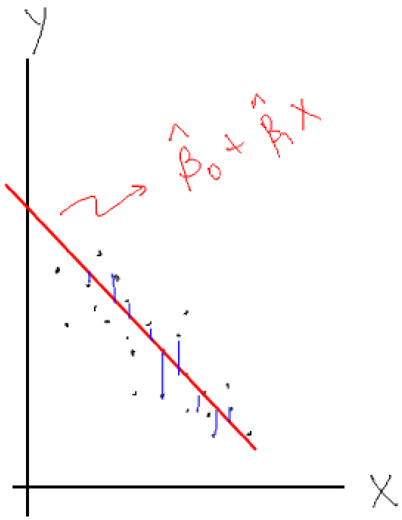


$H_0: \beta_1 = 0 \rightarrow$ no existe relación lineal entre X e Y

$H_1: \beta_1 \neq 0 \rightarrow$ existe relación lineal entre X e Y (Variab Regres > Variab Error)

α

F_{calc} : estadístico de prueba

Variabilidad
- Regresión ✓
- Error ✓

	Fuente	GL	SC	CM	F_{cal}
✓	Regresión	1	SC_{Reg}	CM_{Reg}	CMReg / CMError
✓	Error	$n-2$	SC_{Error}	CM_{Error}	
	Total	$n-1$	SC_{Total}		

$CMError = \hat{\sigma}^2$

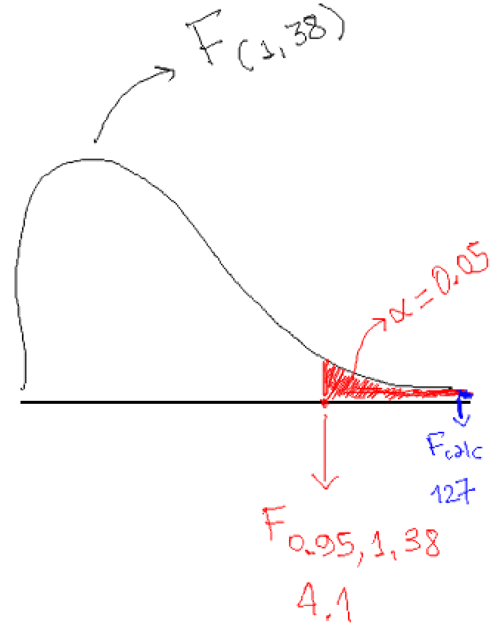
(Varianza estimada del modelo)

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0 \qquad \alpha = 0.05$$

```r
anova = aov(y ~ x) |> tidy()
anova
```

Fcalc

```
# A tibble: 2 x 6
  term         df sumsq meansq statistic   p.value
  <chr>     <dbl> <dbl>  <dbl>     <dbl>     <dbl>
1 x             1 267.   267.       127.  1.16e-13
2 Residuals    38  80.2   2.11       NA  NA
```



→ α

→ pvalor

$pV < \alpha \Rightarrow$ Rech. Ho

$H_0: \beta_0 \geq 4$

$H_1: \beta_0 < 4$

$\alpha = 0.05$

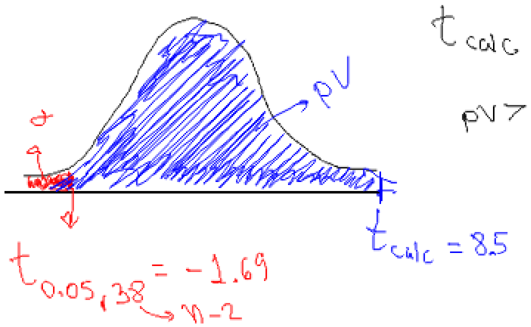
```r
library(broom)
modelo |> tidy()
```

```
# A tibble: 2 x 5
  term         estimate std.error statistic  p.value
  <chr>           <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     8.99     0.588       15.3 8.13e-18
2 horas_estudio   0.641    0.0570      11.3 1.16e-13
```

estimate → $\hat{\beta}$; std.error → $s_{\hat{\beta}}$

$$t_{calc} = \frac{\hat{\beta}_0 - b}{s_{\hat{\beta}_0}} \sim t_{n-2}$$

$$t_{calc} = \frac{8.99 - 4}{0.588} = 8.5$$



t_{calc} cae en la zona de No rechazo de H_0

$pV > \alpha \Rightarrow$ No se rechaza H_0

$H_0: \beta_1 \leq 0.7$

$H_1: \beta_1 > 0.7$

$$t_{calc} = \frac{\hat{\beta}_1 - b}{s_{\hat{\beta}_1}} \sim t_{n-2}$$

$H_0: \beta_0 = 10$

$H_1: \beta_0 \neq 10$

$$t_{calc} = \frac{\hat{\beta}_0 - b}{s_{\hat{\beta}_0}} \sim t_{n-2}$$

Coeficiente de determinación

- El coeficiente de determinación se define como:

$$R^2 = \frac{SC_{\text{Reg}}}{SC_{\text{Total}}}$$

$SCTotal = SCReg + SCError$

Coeficiente de no determinación : $1 - R^2$: % de variabilidad de Y explicada por otros factores distintos de X

```
rsq    = anova$sumsq[1]/(anova$sumsq[1] + anova$sumsq[2])
adjrsq = 1 - (1-rsq)*(n-1)/(n-1-1)
summary(modelo)$r.squared
```

→ R^2 ✓

```
[1] 0.7691646
```

```
summary(modelo)$adj.r.squared
```

→ R^2_{aj} ✓

```
[1] 0.76309
```

$R^2 = 0.769 = 76.9\%$

$R^2_{aj} = 0.763 = 76.3\%$

Estimación de la media y **predicción de un valor individual**

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

$\bar{y} \sim N(\mu, \sigma^2/n)$

Estimación

La media estimada de Y para un valor dado de $X = x$ es:

$$\hat{\mu} = \hat{y} = \boxed{\hat{\beta}_0 + \hat{\beta}_1 x}$$

Su varianza estimada es:

$$\widehat{Var}(\hat{\mu}) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{SC_X} \right) = \frac{\hat{\sigma}^2}{n} + \frac{\hat{\sigma}^2 (x-\bar{x})^2}{SC_X}$$

Predicción

La predicción para una nueva observación Y_0 viene dada por:

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

$Var(\hat{y}_0) > Var(\hat{\mu})$

con varianza estimada:

$$\widehat{Var}(\hat{y}_0) = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SC_X} \right)$$

```r
modelo |> predict(newdata = data.frame(1, horas_estudio = x0),
                  interval = "confidence")
```

```
       fit      lwr      upr
1 14.12054 13.62429 14.61678
```

$$IC(\mu|x = 8) = (13.62, 14.62)$$

```r
modelo |> predict(newdata = data.frame(1, horas_estudio = x0),
                  interval = "prediction")
```

```
       fit      lwr     upr
1 14.12054 11.13748 17.1036
```

$$IP(Y|x = 8) = (11.14, 17.10) \Rightarrow \text{intervalo } \underbrace{\text{más ancho}}_{\text{más variable}}$$

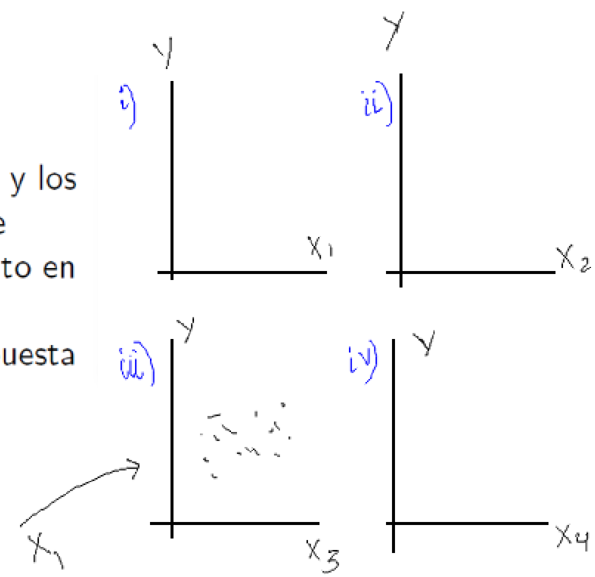
Regresión Lineal Simple → Regresión Lineal Múltiple

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Ejemplo aplicado

Se busca estudiar el efecto lineal de la edad (años), el sexo (femenino / masculino) y los años de educación sobre el sueldo mensual (en miles de soles). Adicionalmente, se incorpora una variable ficticia X_4, generada al azar, para evaluar su comportamiento en el modelo. Se dispone de una muestra de $n = 24$ observaciones.
Una primera etapa del análisis consiste en explorar la relación entre la variable respuesta y cada variable explicativa mediante diagramas de dispersión.



Regresión Lineal Múltiple

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\beta_{(p+1)\times 1} + \varepsilon_{n\times 1}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \underbrace{x_{n1}}_{X_1} & \underbrace{x_{n2}}_{X_2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$\mathbf{X}$: matriz del modelo

$$y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} + \varepsilon_1$$

$$y_1 = \beta_0 + \sum_{j=1}^{p} \beta_j X_{1j} + \varepsilon_1$$

$$\hat\beta = (\underbrace{\underbrace{\underbrace{\mathbf{X}'}_{(p+1)\times n}\underbrace{\mathbf{X}}_{n\times(p+1)}}_{(p+1)\times(p+1)})^{-1}\underbrace{\underbrace{\mathbf{X}'}_{(p+1)\times n}\underbrace{\mathbf{y}}_{n\times 1}}_{(p+1)\times 1}}_{(p+1)\times 1}$$

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

$$\hat\beta = \begin{pmatrix} \hat\beta_0 \\ \hat\beta_1 \\ \vdots \\ \hat\beta_p \end{pmatrix}$$

$$Var(\hat\beta) = \underbrace{\sigma^2}_{\text{número}}\underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{(p+1)\times(p+1)} = \sigma^2 \begin{pmatrix} & \\ & \end{pmatrix}_{(p+1)\times(p+1)}$$

Ejm: $\hat\beta = \begin{pmatrix} \hat\beta_0 \\ \hat\beta_1 \\ \hat\beta_2 \end{pmatrix}$, $Var(\hat\beta) = \sigma^2 \begin{pmatrix} & \\ & \end{pmatrix}_{3\times 3} = \begin{pmatrix} Var(\hat\beta_0) & Cov(\hat\beta_0,\hat\beta_1) & Cov(\hat\beta_0,\hat\beta_2) \\ \cdot & Var(\hat\beta_1) & Cov(\hat\beta_1,\hat\beta_2) \\ \cdot & \cdot & Var(\hat\beta_2) \end{pmatrix}$

simétrica

```r
library(dplyr)
y = datos$Sueldo
X = model.matrix(Sueldo~Educacion+Sexo+Edad+X4,data=datos)
solve(t(X)%*%X)%*%t(X)%*%y
```

modelo

→ genera la matiz X

inversa

transpuesta

$(X'X)^{-1}X'y$

	[,1]	
(Intercept)	-0.73053615	→ $\hat{\beta}_0$
Educacion	0.15695057	→ $\hat{\beta}_1$
SexoM	0.82211261	→ $\hat{\beta}_2$
Edad	0.10462774	→ $\hat{\beta}_3$
X4	-0.04537505	→ $\hat{\beta}_4$

incluir todas las variables

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

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

Ejm: $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$ $\quad Var(\hat{\beta}) = \begin{pmatrix} Var(\hat{\beta}_0) & cov(\hat{\beta}_0, \hat{\beta}_1) \\ cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{pmatrix}$