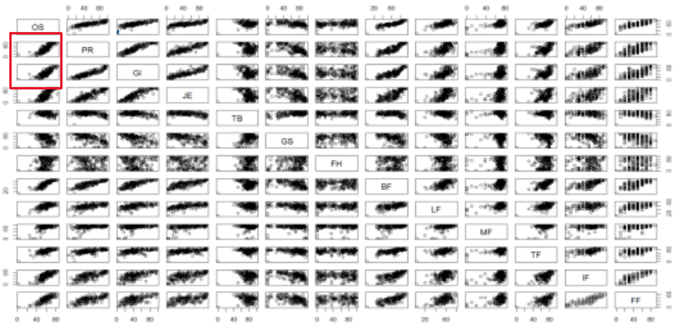
# Segundo taller

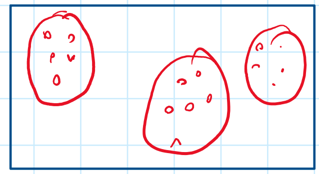
miércoles, 13 de noviembre de 2024 1:57 p. m.

$\boxed{Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i}$ ; $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

• $\boxed{Y_i\text{: Overall score; } X_1\text{: GI; } X_2\text{: PR}}$ ; $\boxed{OS_i = \beta_0 + \beta_1 GI_i + \varepsilon_i}$ $\boxed{OS_i = \beta_0 + \beta_1 PR_i + \varepsilon_i}$




Función seed:



N

(3). tercer punto: Recta ajustada: $\boxed{\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i}}$

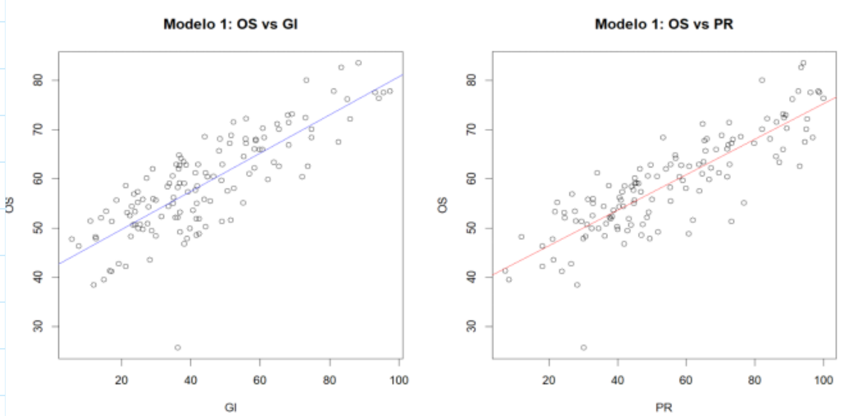
```
> summary_1$coefficients
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 41.8968861 1.20241065 34.84407 8.393346e-68
GI           0.3888582 0.02475893 15.70578 7.999741e-32
>
```

$\hat{\beta}_1$

$\boxed{\hat{OS}_i = 41,896 + 0,388\, GI_i}$ Modelo 1

```
> summary_2$coefficients
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 39.1856735 1.19554767 32.77634 1.065072e-64
PR           0.3606188 0.02004955 17.98637 4.409235e-37
```

$\boxed{\hat{OS}_i = 39,185 + 0,360\, PR_i}$ Modelo 2.




Con los datos que estamos trabajando, se tiene que todas las variables varían desde cero hasta cien.

(4). Probar significancia de los parámetros y regresión ; $\boxed{n = 132}$

• Significancia parámetros: $\beta_0$: $\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases}$ : $T = \dfrac{\hat{\beta}_0 - 0}{\sqrt{Var[\hat{\beta}_0]}} \sim t_{n-2} \longrightarrow T_{cal}$ (Al evaluar)

• (1). $P(|t_{n-2}| > |T_{cal}|) < \alpha$ : Rechazar $H_0$.

(2). $|T_{cal}| > t_{\alpha/2, n-2}$

$T_{cal}$: $\boxed{32,77634}$


```
> summary_2$coefficients
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 39.1856735 1.19554767 32.77634 1.065072e-64
PR           0.3606188 0.02004955 17.98637 4.409235e-37
```


$P(|t_{130}| > |32,77634|) \approx 0 < \alpha$

• Rechazar $H_0$. $\alpha = 0,05$


$\boxed{\hat{OS}_i = 39,185 + 0,360\, PR_i}$ Modelo 2.


El parámetro B0 es significativamente distinto de cero.

En el caso de $\beta_0$: $\boxed{PR_i \text{ contener al cero}}$

Como el rango de la covariable PR no contiene al cero, no se puede realizar una interpretación.

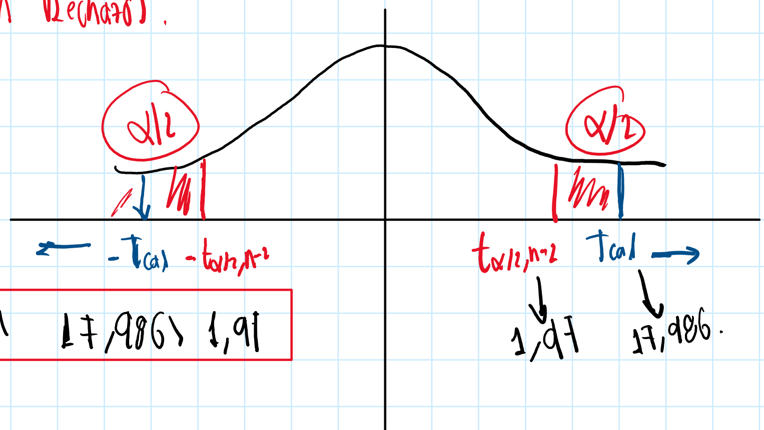
$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\boxed{T = 17{,}986;\ P(|t_{130}| > 17{,}986) \approx 0 < \alpha.}$

(2). $|t_{cal}| > t_{\alpha/2, n-2}$ (Región Rechazo).

$t_{\alpha/2, n-2} \rightarrow 1{,}97$




$|17{,}986| > 1{,}97$

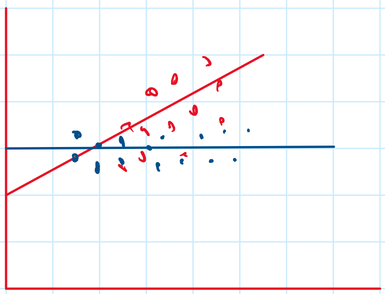
$\boxed{-17{,}986 < -1{,}97 \ \wedge \ 17{,}986 > 1{,}97}$

∘ Se rechaza $H_0$

B1 (pendiente) es significativamente diferente de cero. Existe un efecto de PR sobre el promedio del OS. Por un cambio unitario en el PR, el OS cambia en promedio 0,360.

Significancia Regresión

Significancia regresión (Análisis de Varianza – ANOVA).



$SST = S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad gl(SST) = n-1$

$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} \qquad gl(SSE) = n-2$

$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \rightarrow gl(SSR) = 1$

$\boxed{SST = SSE + SSR}$ $\quad$ $\boxed{MSR = SSR}$

| Fuente | | gl | Medias | |
|---|---|---|---|---|
| Regresión | SSR | 1 | SSR/1 = | MSR |
| Error | SSE | n-2 | SSE/n-2 = | MSE |
| Total | SST | n-1 | | |

Estimador Varianza

$\boxed{\hat{\sigma}^2 = MSE}$

$\boxed{y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i}$ ; $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Estadístico: $F = \dfrac{MSR}{MSE} \sim f_{1, n-2}$

$\beta_1$ : $\boxed{\text{Prueba T}}$

$\boxed{P(|t_{n-2}| > |T_{cal}|) = P(f_{1,n-2} > F_{cal})}$ ; $\boxed{T_{cal}^2 = F_{cal}}$

MSR

$\boxed{F: \dfrac{9390/1}{3745{,}2/130} = \dfrac{9390}{28{,}8} = 323{,}51}$

```
> anova_2
Analysis of Variance Table

Response: OS
           Df Sum Sq Mean Sq F value    Pr(>F)
PR          1 9320.0  9320.0  323.51 < 2.2e-16 ***
Residuals 130 3745.2    28.8
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSR

SSE

∘ $P(f_{1,130} > 323{,}51) \approx 0 < \alpha$

EL MODELO DE REGRESIÓN DE OS EN FUNCIÓN DE PR, ES SIGNIFICATIVO PARA EXPLICAR LA VARIABILIDAD OBSERVADA EN OS.

## (5). Comparación $R^2$ : Proporción (Variabilidad) Total observada en Y dada su relación lineal con X.

- El modelo es adecuado (ERRÓNEO)
- El modelo presenta buen ajuste (ERRÓNEO)

```
> summary_1$r.squared
[1] 0.6548717
> summary_2$r.squared
[1] 0.7133468
```

M1: GI
M2: PR.

MODELO 2: Existe mayor proporción de la variabilidad total observada en OS que explicada por el modelo, es decir, de su relación lineal con PR.

$$R^2 = \frac{SSR}{SST}$$

## (6). Respuesta Media y predicción

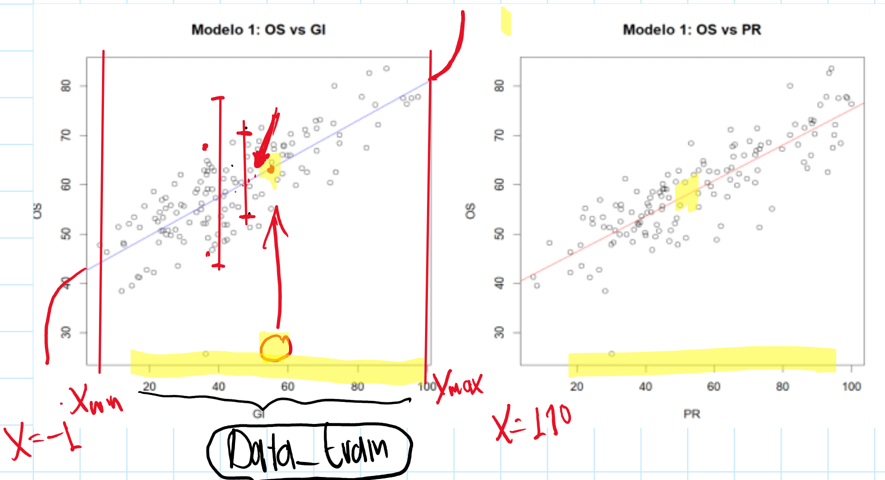
Recta ajustada: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i}$

$$\boxed{\widehat{OS}_i = 41,896 + 0,308\, GI_i}$$

M1

$$\boxed{\widehat{OS}_i = 39,185 + 0,360\, PR_i}$$

M2.




Respuesta media:

$$E[Y|X_0] = \beta_0 + \beta_1 X_0$$

$$\hat{E}[Y|X_0] = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

$$\hat{Y}_0 \sim N\left(E[Y|X_0], \sigma^2\left[\frac{1}{n} + \frac{(X_0-\bar{X})^2}{S_{xx}}\right]\right)$$

MSE

→ Estimador puntual $\mu_{Y|X_0}$

→ Interesado para obs. futuras.

S.E($\hat{Y}_0$).

- Intervalo conf. $\mu_{Y|X_0}$: $\hat{Y}_0 \pm t_{\alpha/2,n-2}\sqrt{Var[\hat{Y}_0]}$

$$\hat{Y}_0 \pm t_{\alpha/2,n-2}\sqrt{\widehat{Var}[\hat{Y}_0]}$$

- Intervalo predicción: $\hat{Y}_0 \pm t_{\alpha/2,n-2}\sqrt{MSE\left[1 + \frac{1}{n} + \frac{(X_0-\bar{X})^2}{S_{xx}}\right]}$

```
fit      lwr      upr
```

| | fit | lwr | upr |
|---|---|---|---|
| 42 | 64.56732 | 63.33285 | 65.80179 |
| 159 | 44.11338 | 41.98378 | 46.24298 |
| 103 | 53.21266 | 51.96519 | 54.46013 |

Conjunto de entrenamiento.

| | fit | lwr | upr |
|---|---|---|---|
| 2 | 77.39964 | 65.47604 | 89.32324 |
| 5 | 74.91095 | 63.04437 | 86.77753 |
| 7 | 79.18839 | 67.21890 | 91.15788 |

Int Predicción (Observaciones no empleadas en ajuste)

$\boxed{X_0 = 9}$

| | fit | lwr | upr |
|---|---|---|---|
| 1 | 45.39661 | 33.57646 | 57.21676 |