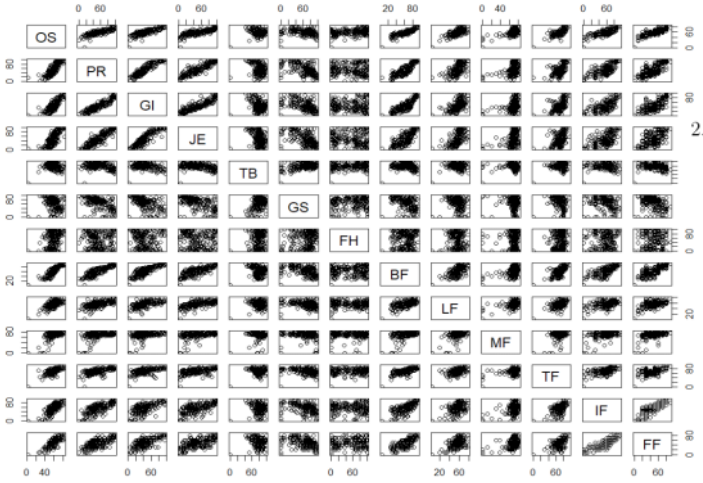
Segundo Taller

sábado, 6 de septiembre de 2025 12:07 p. m.

OS: Overall score ; GI: Government Integrity; PR: Property rights

1. Realice un breve análisis descriptivo, ¿Qué covariables podría escoger para un análisis de regresión lineal simple? *Seleccione dos.*

(1). Realizar una matriz de correlación



. Se seleccionarán GI y PR debido a su alta correlación con la respuesta OS.

2. Escriba la ecuación del modelo de regresión lineal, considerando los supuestos asociados. Ajuste el modelo lineal para cada una de las covariables seleccionadas. ¿Son comparables ambos modelos? ¿Tienen las mismas unidades?

$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$; $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$; $i = 1, \dots, n$

$n = 176$

(1). $OS_i = \beta_0 + \beta_1 GI_i + \varepsilon_i$; $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

(2). $OS_i = \beta_0 + \beta_1 PR_i + \varepsilon_i$; $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

Versión genérica ajustada:

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1}$

Resumen primer modelo

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.40907    1.17916   34.27   <2e-16 ***
GI           0.41041    0.02389   17.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.812 on 174 degrees of freedom
Multiple R-squared:  0.6291,    Adjusted R-squared:  0.6269
F-statistic: 295.1 on 1 and 174 DF,  p-value: < 2.2e-16
```

(1). $\hat{OS}_i = 40{,}40907 + 0{,}41041\, GI_i$

Resumen segundo modelo

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.90303    1.16936   32.41   <2e-16 ***
PR           0.37993    0.01958   19.40   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.289 on 174 degrees of freedom
Multiple R-squared:  0.6839,    Adjusted R-squared:  0.6821
F-statistic: 376.4 on 1 and 174 DF,  p-value: < 2.2e-16
```

(2). $\hat{OS}_i = 37{,}90303 + 0{,}37993\, PR_i$

. Significancia de los coeficientes: Pruebas de hipótesis $\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$ $j = 0, 1$

. Estadísticas de prueba: Significancia individual (distribución T)

$$T_i = \frac{\hat{\beta}_j - 0}{\sqrt{Var(\hat{\beta}_j)}} \sim t_{n-2}$$

$$T_0 = \frac{0{,}41041}{0{,}02389} = 17{,}18$$

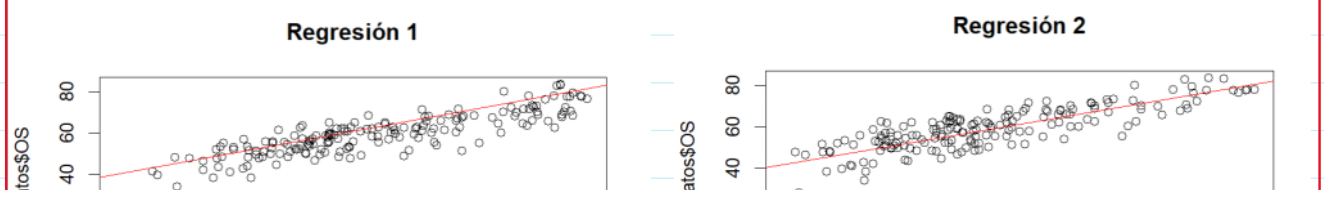
(1). Con el valor P

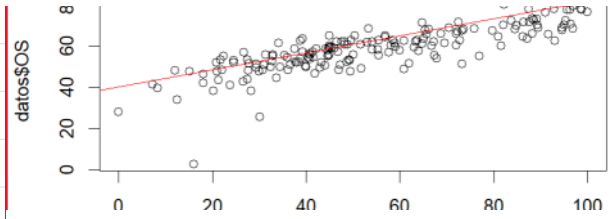
$P(|t_{n-2}| > |T_0|) < \alpha$

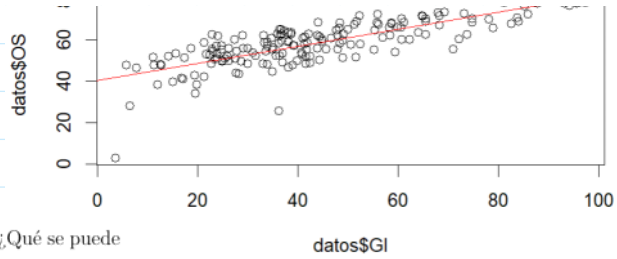
$\alpha = 0{,}05$

. Pval ≈ 0; $\alpha = 0{,}05$: Dado que Pval < 0,05: A un nivel de significancia del 5%, existe evidencia suficiente para rechazar H_0, de forma que β_1 es significativo, así, por un aumento unitario del valor de GI, el puntaje general (OS) aumenta en promedio 0,41041 unidades

. ¿Cómo sé yo que β_0 es significativo? β_0 es solo interpretable si el cero está contenido en mi conjunto de datos







4. Calcule la métrica R^2 para ambos modelos. Realice una interpretación. ¿Qué se puede concluir de una comparación del R^2?

¿Cómo se comparan los modelos?

R^2: Cuantificar el grado de variabilidad explicada por la relación lineal entre la respuesta y la covariable

- R^2: El valor más alto posible, sujeto al principio de parsimonia
- La variabilidad explicada por el modelo es superior en el segundo caso

(1). 0,6291 ; (2). 0,6839

```
       OS              PR              GI       
 Min.   : 2.90   Min.   :  0.00   Min.   :  3.60  
 1st Qu.:51.98   1st Qu.: 37.27   1st Qu.: 28.10  
 Median :58.80   Median : 49.50   Median : 40.90  
 Mean   :58.64   Mean   : 54.59   Mean   : 44.43  
 3rd Qu.:65.75   3rd Qu.: 72.55   3rd Qu.: 58.48  
 Max.   :83.50   Max.   :100.00   Max.   : 97.40  
```

¿Tienen las mismas unidades?

- Las dos están acotadas entre 0 y 100
- Las variables están estandarizadas

3. Realice la prueba de significancia para *la pendiente* en alguno de los modelos, luego una prueba de significancia para la regresión mediante el análisis de varianza. ¿Existe alguna relación entre ambas pruebas? Brinde una interpretación para los parámetros (β_0, β_1) de uno de los modelos considerados.

- Probar significancia de β_1 en el primer modelo

- Significancia de los coeficientes: Prueba de hipotesis $\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$ $j = 0, 1$
- Estadísticos de prueba: Significancia individual (distribución T)

$$T_i = \frac{\hat{\beta}_j - 0}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} \sim t_{n-2}$$

$$T_0 = \frac{0,41041}{0,02389} = 17,18$$

(1). Con el valor P

$P(|t_{n-2}| > |T_0|) < \alpha$

$\alpha = 0,05$

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

$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ $gl(SSE) = n-2$

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

- Análisis de Varianza (ANOVA): $Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \varepsilon_i$

$Y_i = \beta_0 + \beta_1 X_{i1}$

$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$: $F = \frac{SSR/1}{SSE/n-2} = \frac{MSR}{MSE} \sim f_{1,n-2}$

- El MSE es la estimación de la varianza

$\hat{\sigma}^2 = MSE$ $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Resumen primer modelo

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 40.40907    1.17916   34.27   <2e-16 ***
GI           0.41041    0.02389   17.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.812 on 174 degrees of freedom
Multiple R-squared:  0.6291,    Adjusted R-squared:  0.6269 
```

$F = 295,1$; $Pval < \alpha$

- La regresión es significativa

```
> anova(modelo1)
Analysis of Variance Table

Response: OS
          Df  Sum Sq Mean Sq F value    Pr(>F)    
GI         1 13694.3 13694.3  295.09 < 2.2e-16 ***
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.812 on 174 degrees of freedom
Multiple R-squared:  0.6291,    Adjusted R-squared:  0.6269
F-statistic: 295.1 on 1 and 174 DF,  p-value: < 2.2e-16
```

(1). $\hat{OS}_i = 40,40907 + 0,41041\, GI_i$

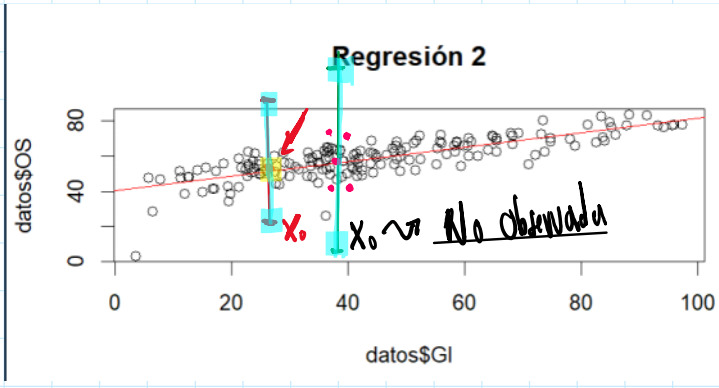
```
Analysis of Variance Table

Response: OS
           Df  Sum Sq Mean Sq F value    Pr(>F)
GI          1 13694.2 13694.2  295.09 < 2.2e-16 ***
Residuals 174  8074.9    46.4
---
```

$F = \frac{13694,2}{46,4} = 295,09$ → MSR / MSE

(1). Propiedad: $P(|t_{n-2}| > |t_0|) = P(f_{1,n-2} > F_{cal})$

(1). Intervalo confianza para la respuesta media.
(2). Intervalo predicción

5. Seleccione uno de los modelos planteados. Realice una predicción puntual para la respuesta en un valor apropiado $X = x_0$. Brinde el correspondiente intervalo de confianza al 95% para la respuesta media y el intervalo de predicción para el valor futuro de la respuesta. ¿Qué diferencias existen entre ambos?



- El IP es para una observación puntual Y_0 y IC para la media

IC: $\hat{\theta} \pm t_{n-2} \cdot \sqrt{Var(\hat{\theta})}$

$\hat{Y}_0 \pm t_{n-2} \cdot \sqrt{Var(\hat{Y}_0)}$.

IP: $\hat{Y}_0 \pm t_{n-2} \cdot \sqrt{Var(Y_0 - \hat{Y}_0)}$

$Y_0 \sim N(E(Y|X_0), \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right])$

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

MSE → σ^2

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

↳ El intervalo es más ancho (mayor variabilidad)

X_0: 9

```
> predict(modelo1, newdata = data.frame(GI = 9),
+         interval = "confidence",
+         level = 0.95)
       fit      lwr      upr
1 44.10276 42.14875 46.05678
```

```
> predict(modelo1, newdata = data.frame(GI = 9),
+         interval = "prediction",
+         level = 0.95)
       fit      lwr      upr
1 44.10276 30.51615 57.68937
```