# Taller Práctico Regresión Lineal Simple (1) *

**Estadística II** *Universidad Nacional de Colombia, Sede Medellín*

---

Este documento corresponde al primer taller práctico del curso de **Estadística II** para la *Universidad Nacional de Colombia*, Sede Medellín, en el periodo 2024 - 2. Se brinda una introducción al análisis de regresión. El enfoque de este taller está sobre las componentes asociadas al modelo de regresión lineal simple -especialmente los parámetros-. **Monitor:** *Santiago Carmona Hincapié.*

*Keywords*: regresión, parámetros

---

## Información general

Con el propósito de profundizar en los conceptos del modelo de regresión lineal simple vistos en clase, se propone afrontar dos problemas prácticos. El segundo consta de una simulación que ahonda en las propiedades de los parámetros del modelo de regresión. El primero, es una aproximación al uso del modelo en situaciones de la vida real.

**La solución para cada uno de los problemas se efectúa a partir del software estadístico R.**

*Ejercicio con datos reales*

El índice de libertad económica es una medida que *evalúa el grado de libertad económica* en diferentes países. Se presentan diversos atributos, cuya descripción puede encontrarse **en el siguiente enlace:** https://www.kaggle.com/datasets/mlippo/freedom-economic-index/data

Table 1: Información en análisis

| OS | PR | GI | JE | TB | GS | FH | BF | LF | MF | TF | IF | FF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **83.5** | 94.2 | 88.3 | 58.3 | 90.7 | 89.2 | 76.0 | 86.9 | 77.3 | 76.3 | 95.0 | 90 | 80 |
| **83.0** | 94.2 | 91.3 | 98.1 | 70.4 | 64.6 | 95.7 | 89.3 | 60.7 | 80.8 | 86.4 | 85 | 80 |
| **82.6** | 93.5 | 83.4 | 94.3 | 78.0 | 82.4 | 91.7 | 91.3 | 62.8 | 74.5 | 79.2 | 90 | 70 |
| **80.0** | 82.2 | 73.4 | 94.0 | 79.2 | 90.5 | 90.3 | 84.9 | 69.1 | 80.1 | 86.4 | 70 | 60 |
| **79.2** | 96.9 | 84.9 | 95.8 | 64.6 | 40.6 | 97.6 | 89.5 | 57.7 | 69.1 | 79.2 | 95 | 80 |

Considere a *'Overall Score'* como la variable respuesta. *Escoja una covariable y de respuesta a los siguientes planteamientos:*

1. Realice un **breve análisis descriptivo**. ¿Un modelo de regresión lineal simple podría ser adecuado en este caso? ¿Por qué?

*El material asociado a este taller puede encontrarse en el repositorio del curso, **(https://github.com/Itsssach/Estadistica-II)**

2. Escriba la ecuación del modelo de regresión lineal, considerando los supuestos asociados. Obtenga los valores calculados de los parámetros, $\overline{y}, \hat{y}, \hat{\varepsilon_i}$ y analícelos.

3. Determine si los parámetros del modelo $\beta_0, \beta_1$ son significativos, considerando $\alpha = 0.05$. Realice una interpretación en relación al problema. **¿Estos parámetros tienen sentido?**

4. Calcule un intervalo de confianza -considerando $\alpha = 0.05$- para ambos parámetros. ¿Puede concluir a partir de este resultado si los parámetros son significativos?

*Ejercicio de simulación*

A partir de una simulación se pretende ilustrar algunos de los principios teóricos que comprenden el análisis de regresión lineal simple. Así, se plantea un modelo $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i, \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, tal que $\beta_0 = 50, \beta_1 = 10, \sigma^2 = 16$. Suponga que se emplean $n = 20$ observaciones para ajustar el modelo. **Genere 500 muestras de 20 observaciones de tal manera que** $x = 1, 1.5, 2, ..., 10$ **para cada muestra.**

1. Para cada muestra calcule $\hat{\beta}_0, \hat{\beta}_1$. Construya un histograma para cada parámetro estimado y concluya.

2. Para cada muestra, calcule un intervalo de confianza al 90% para $\beta_1$. *¿Cuántos de estos intervalos contienen al verdadero valor del parámetro? ¿Se corresponde la teoría con la práctica?*

**Solución**

Se brinda la solución para los problemas planteados. Se brinda el código requerido para la generación de las gráficas (sin las opciones de personalización). **Solo se muestra el código esencial (para código completo ver el archivo 01Solucion.Rmd).** Sólo se brindan los análisis para la simulación. **Los análisis restantes son labor del estudiante.**

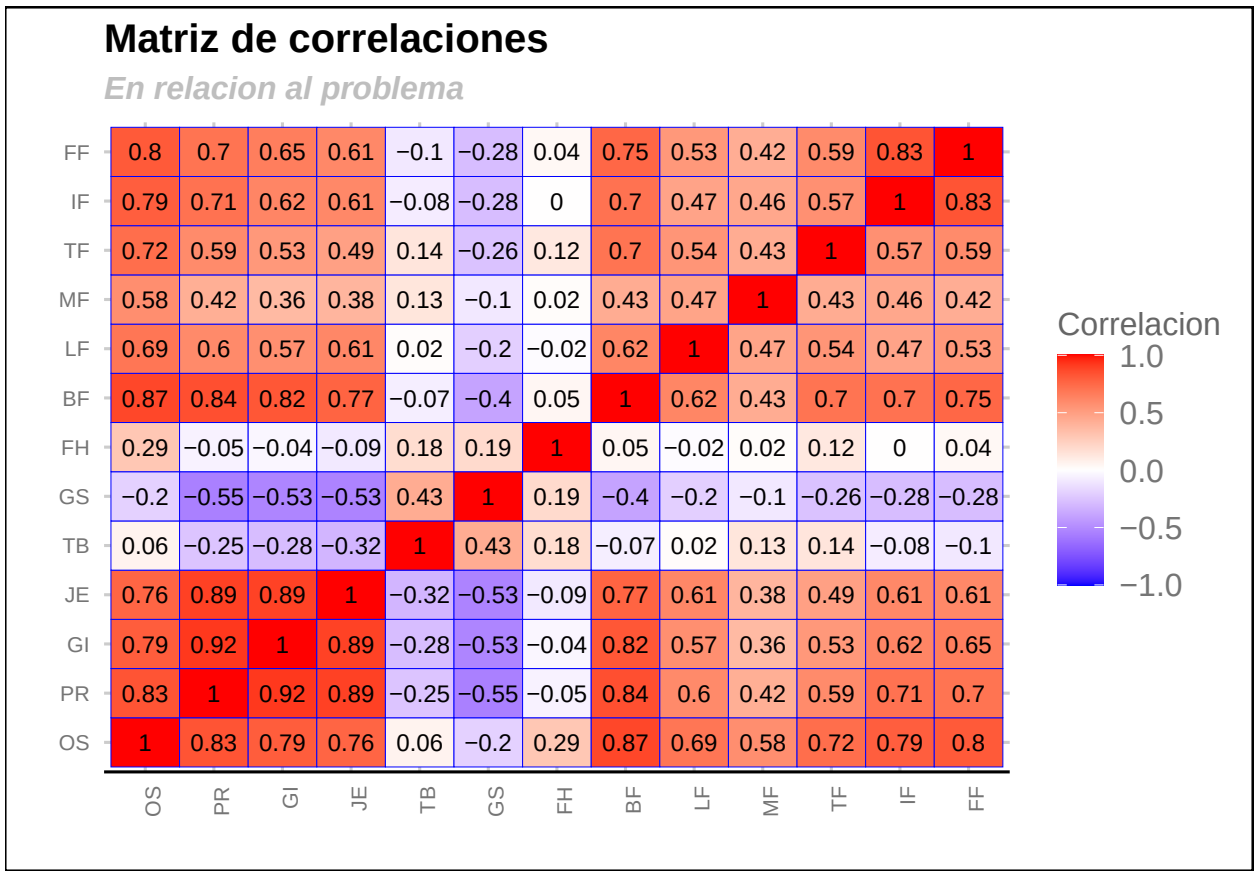
*Parte práctica*

*Primer punto*

Realice un **breve análisis descriptivo**. ¿Un modelo de regresión lineal simple podría ser adecuado en este caso? ¿Por qué?

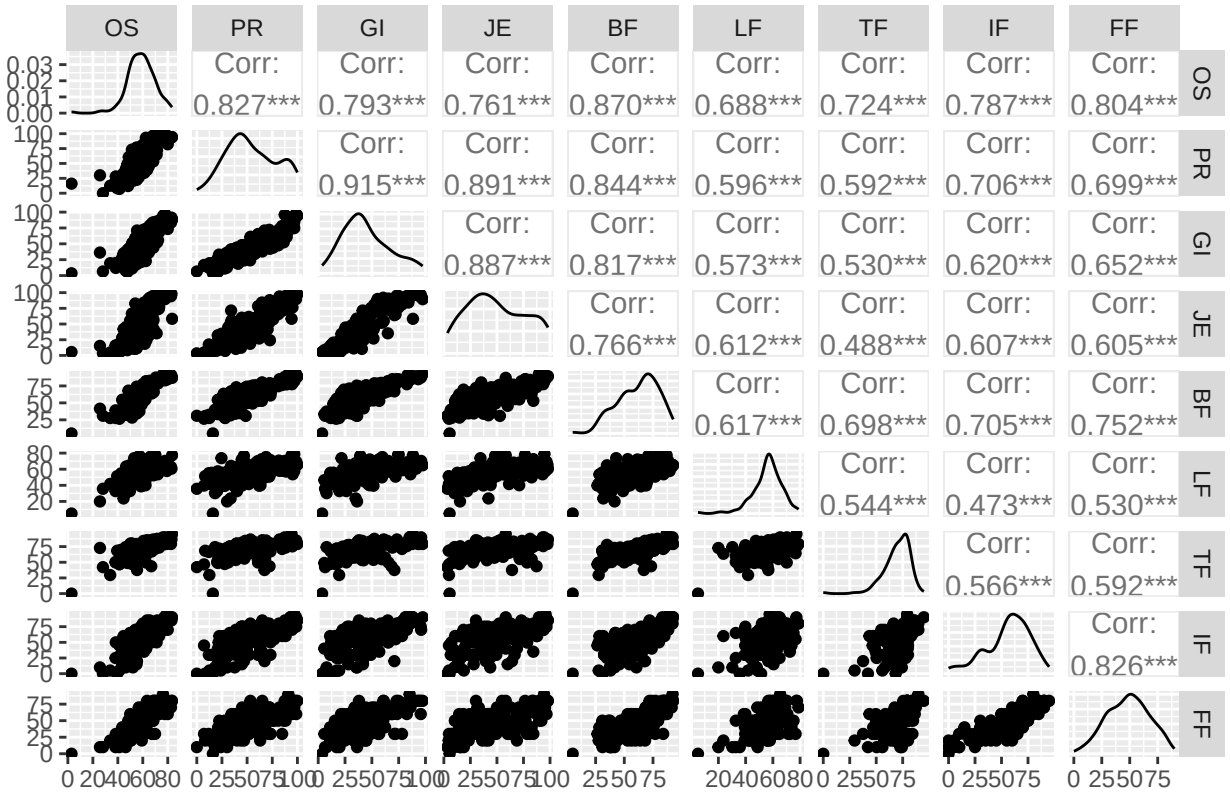
```
# ----------------------------
#     Análisis descriptivo
# ----------------------------
cor_matriz <- cor(data) |>
  reshape2::melt() # Formato largo

# ----------------------------
correlation_plot <- ggplot(cor_matriz, aes(Var1, Var2, fill= value)) +
  geom_tile(color= "blue") +
  scale_fill_gradient2(low= "blue", high= "red",
                       midpoint= 0, limit= c(-1, 1), space= "Lab",
                       name= "Correlacion") +
  geom_text(aes(label = round(value, 2)), color = "black", size = 3) +
  # coord_fixed() +
  labs(title= "Matriz de correlaciones", x= "",
       y= "", subtitle= "En relacion al problema")
```

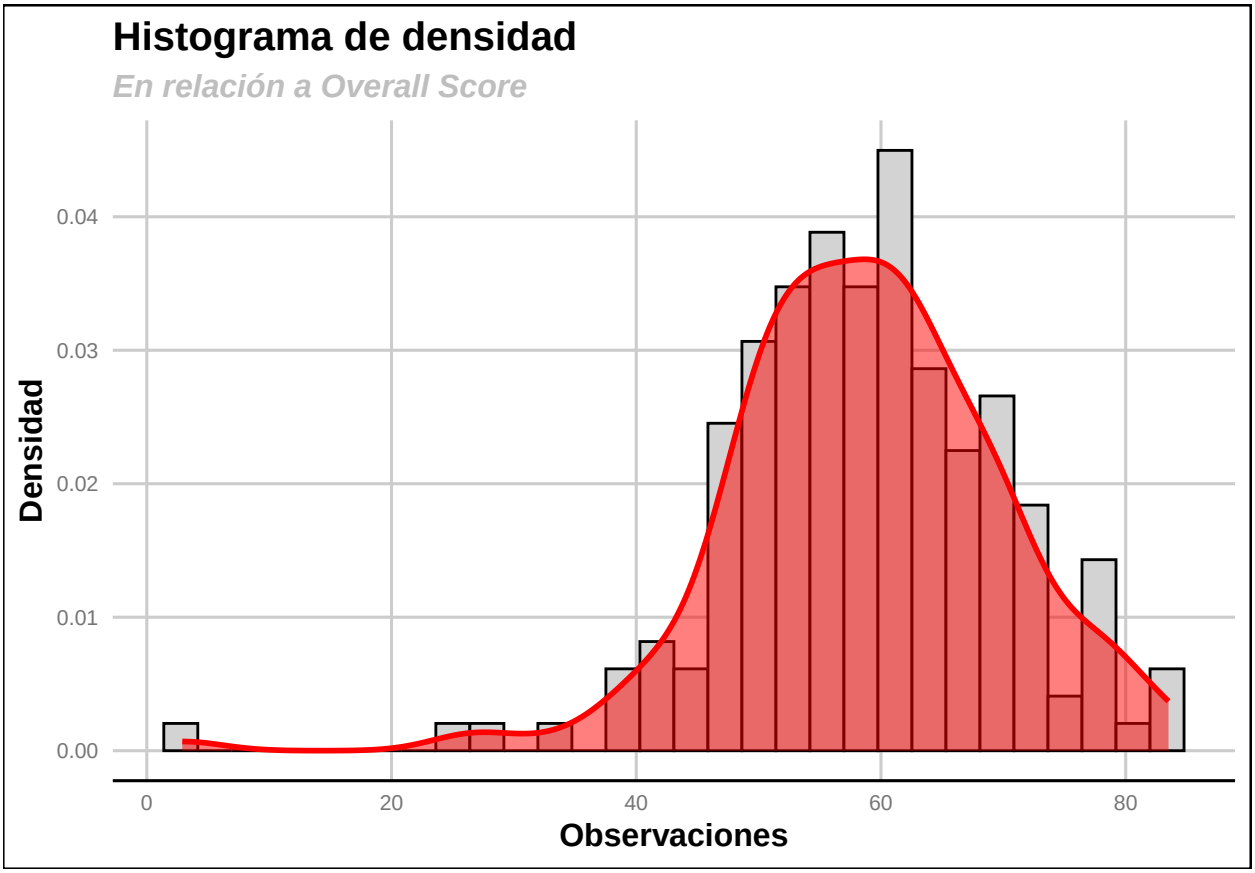



```r
# ------------------------------
#    Gráfico de correlación
# ------------------------------
highly_correlated <- data |>
  dplyr::select(-TB, -GS, -FH, -MF)
GGally::ggpairs(highly_correlated, title = "Gráfico de correlación")
```

Gráfico de correlación




```
# ------------------------------
#    Gráfico de densidad (OS)
# ------------------------------
density_plot <- ggplot(data, aes(OS)) +
  geom_histogram(aes(y= after_stat(density)), fill= "lightgray",
                 color= "black", position= "identity") +
  geom_density(color= "red", lwd= 1, fill= "red", alpha= 0.5) +
  labs(title= "Histograma de densidad", x= "Observaciones",
       y= "Densidad", subtitle= "En relación a Overall Score" )
```

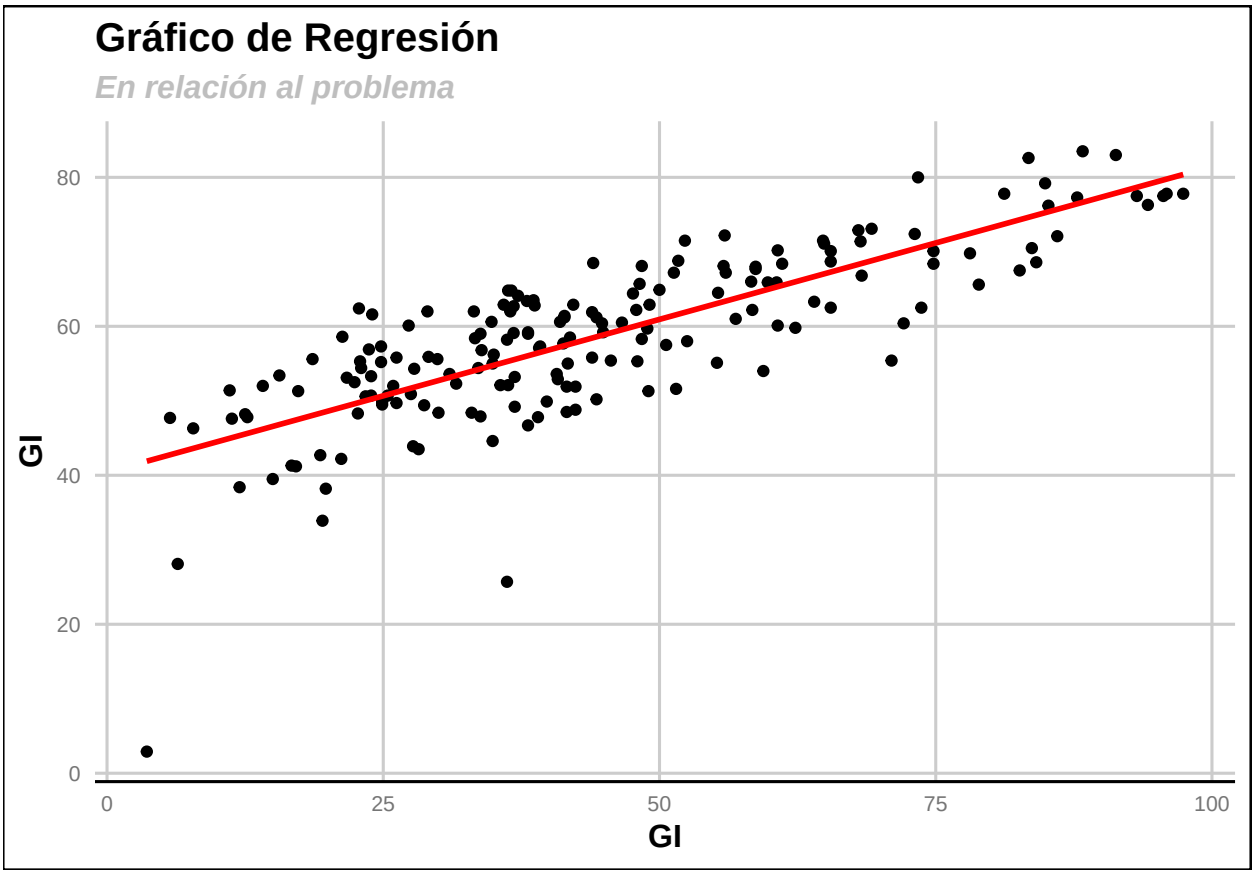



*Segundo punto*

```r
# ----------------------------
#      Modelo de regresión
# ----------------------------
model <- stats::lm(OS ~ GI, data = data)
# Ajustando el modelo de regresión
# Encontrar la información
beta_0 <- coef(model)[1] # Beta_0
# Same as model$coefficients[1]
beta_1 <- coef(model)[2] # Beta_1
y_bar <- mean(data$OS) # Mean Y
y_hat <- fitted(model) # Ajustados
residuals <- model$residuals
sigma_2 <- sigma(model)^2 # Sigma^2
# ----------------------------
# Hallando los valores a mano
x <- data$GI; y <- data$OS; n <- length(data)
x_bar <- mean(x); y_bar <- mean(y)
Sxx <- sum((x- x_bar)^2); Sxy <- sum((x- x_bar)*y)
```

```r
beta_1 <- Sxy/Sxx # Beta_0
beta_0 <- y_bar - (beta_1 * x_bar) # Beta_1
residuals <- y - y_hat # Residuales
sigma_2 <- sum(residuals^2)/(n -2)
# ----------------------------
```

```r
# -------------------------
#    Gráfico de regresión
# -------------------------
regression <- ggplot(data, aes(x = GI, y = OS)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE,color = "red") +
  labs(title = "Gráfico de Regresión", x = "GI", y = "GI",
       subtitle= "En relación al problema" )
```




*Tercer punto*

```r
# -------------------------
#         Parámetros
```

```r
# -------------------------
model_summary <- summary(model)
```

Table 2: Significancia parámetros (Prueba t)

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 40.4090683 | 1.1791559 | 34.26949 | 0 |
| GI | 0.4104104 | 0.0238915 | 17.17808 | 0 |

*Cuarto punto*

```r
# -------------------------
#   Intervalo de confianza
# -------------------------
confint_beta0 <- confint(model, "(Intercept)", level = 0.95)
confint_beta1 <- confint(model, "GI", level = 0.95)
# -------------------------
alpha <- 0.05 # Nivel de significancia
beta0_lower <- beta_0 - qt(1 - alpha/2, n- 2)*sqrt(sigma_2*sum(x^2)/(n*Sxx))
beta0_upper <- beta_0 + qt(1 - alpha/2, n-2)*sqrt(sigma_2*sum(x^2)/(n*Sxx))
# -------------------------
beta1_lower <- beta_1 - qt(1- alpha/2, n-2)*sqrt(sigma_2/Sxx)
beta1_upper <- beta_1 + qt(1- alpha/2, n-2)*sqrt(sigma_2/Sxx)
```

Table 3: Significancia parámetros (Prueba t)

| Inferior | Superior | Parametro |
|---|---|---|
| 2.4293910 | 78.3887456 | beta[0] |
| 0.2012694 | 0.6195513 | beta[1] |

*Simulación*

*Primer punto*

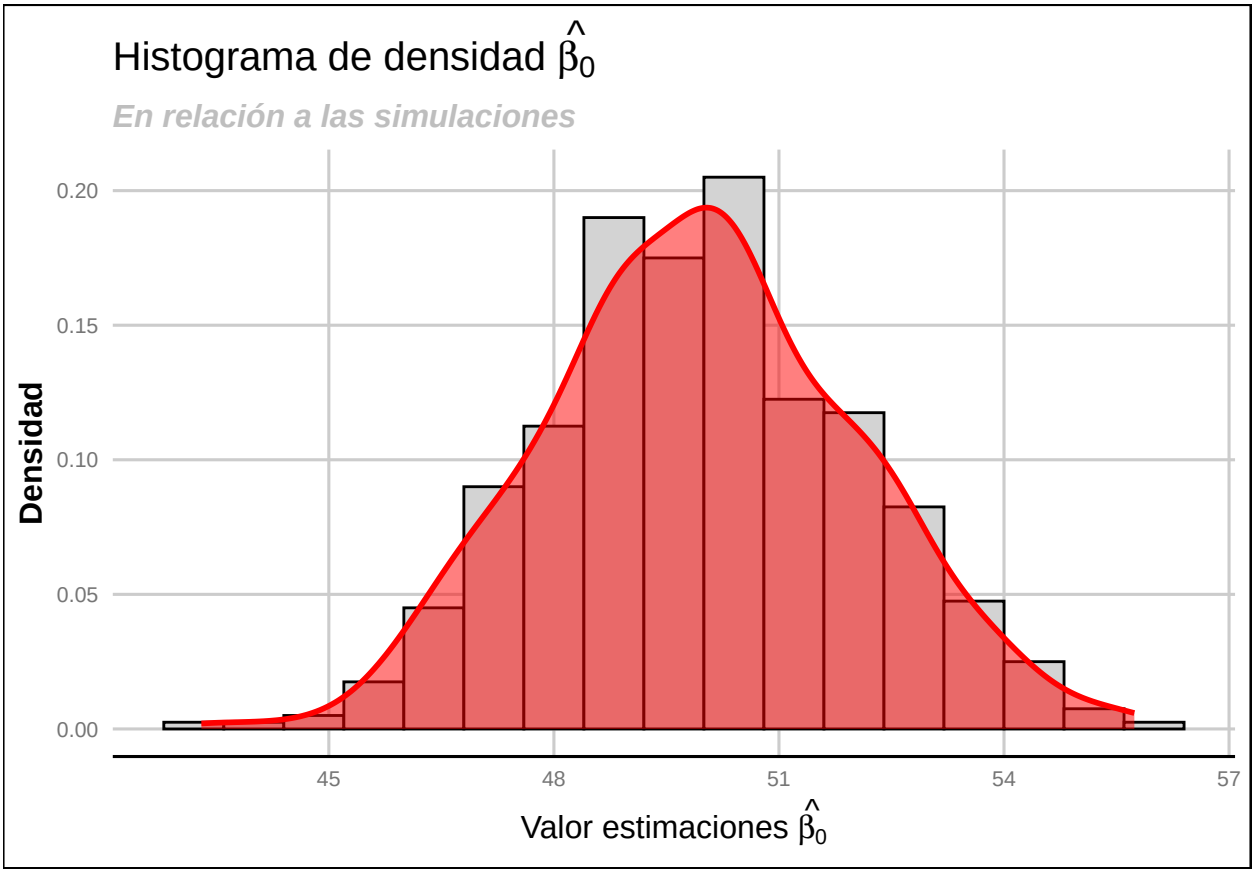
Para cada muestra calcule $\hat{\beta}_0, \hat{\beta}_1$. Construya un histograma para cada parámetro estimado y concluya.

Al llevar a cabo las simulaciones requeridas en esta sección, se tiene que para el modelo de regresión lineal simple planteado como $Y_i = 50 + 10X_i + \varepsilon_i, \varepsilon_i \sim N(0, 16)$, se definió diferentes muestras -en específico, 500 muestras- a las que se les calculó los parámetros estimados $\hat{\beta}_0$ y $\hat{\beta}_1$. Se muestra a continuación algunos de los parámetros calculados:

Table 4: Valores estimados simulados

| | (Intercept) | x |
|---|---|---|
| | 49.1763933612139 | 10.2207051660037 |
| | 49.6963132237989 | 10.2706605043094 |
| | 48.5195982339312 | 10.482203529646 |
| | 49.8972306577429 | 9.99049769764923 |
| | 52.5240732938405 | 9.69780960008486 |
| | . | .vphantom2 . |
| | . | . |
| | . | . |
| **[496,]** | 50.8301645192476 | 10.1284790586654 |
| **[497,]** | 51.41542508623 | 9.75385580694254 |
| **[498,]** | 48.7789084490136 | 9.92244402901368 |
| **[499,]** | 46.0433500432867 | 10.7441317784026 |
| **[500,]** | 51.5845178373664 | 9.70337987700957 |

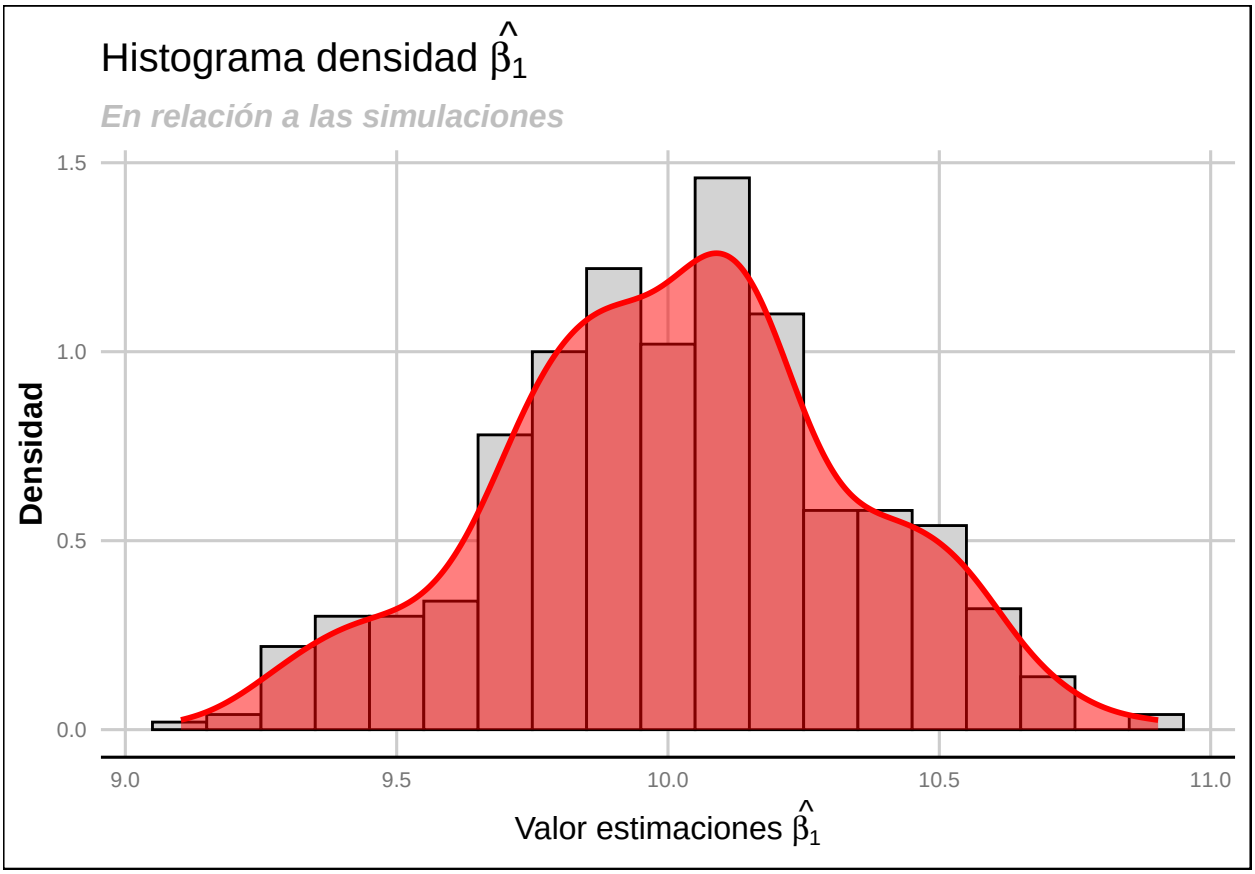
Para la información anterior, se muestra a continuación, los histogramas correspondientes. Se muestra en primer lugar el histograma para $\hat{\beta}_0$, luego, se mostrará el correspondiente histograma para $\hat{\beta}_1$:




Obsérvese que este histograma sigue una curva simétrica, asemejándose a una distribución normal. De hecho, su media está dada por $\overline{\beta_0} = 49.9523895$, mientras que su mediana está

dada por 49.9655355 -al menos el 50% de los estimadores son menores o iguales al valor de la mediana especificado-, lo que indica que existe una diferencia de 0.013146, que es una diferencia mínima, indicando la gran simetría de la distribución en cuestión. De hecho, tomando en consideración $\overline{\beta_0} = 49.9523895$, puede verse que los estimadores simulados, en promedio, se acercan en gran medida al valor real del parámetro $\beta_0 = 50$.

Se muestra a continuación el respectivo histograma para las simulaciones de los parámetros estimados $\hat{\beta}_1$:




Obsérvese que este histograma sigue una curva simétrica, asemejándose a una distribución normal. De hecho, su media está dada por $\overline{\beta_1} = 10.0149896$, mientras que su mediana está dada por 10.0256334 -al menos el 50% de los estimadores son menores o iguales al valor de la mediana especificado-, lo que indica que existe una diferencia de 0.0106439, que es una diferencia mínima, indicando la gran simetría de la distribución en cuestión. De hecho, tomando en consideración $\overline{\beta_1} = 10.0149896$, puede verse que los estimadores simulados, en promedio, se acercan en gran medida al valor real del parámetro $\beta_1 = 10$.

En general, se podría generar un correcto ajuste al modelo de regresión planteado $Y_i = 50 + 10X_i + \varepsilon_i, \varepsilon_i \sim N(0, 16)$.

*Segundo punto*

Para cada muestra, calcule un intervalo de confianza al 90% para $\beta_1$. *¿Cuántos de estos intervalos contienen al verdadero valor del parámetro? ¿Se corresponde la teoría con la práctica?*

En esta sección, para cada muestra, se pretende calcular un intervalo de confianza al 90% para $\beta_1$, de manera que sea posible determinar el número de intervalos de confianza acertados para cada una de las simulaciones. Es así que, se muestran los resultados obtenidos en la tabla a continuación:

Table 5: Intervalos de confianza

| | Límite inferior | Límite superior |
|---|---|---|
| | 9.7856016257952 | 10.6558087062122 |
| | 9.61693693044987 | 10.9243840781689 |
| | 9.99322907575766 | 10.9711779835344 |
| | 9.24604825457134 | 10.7349471407271 |
| | 9.11610947332314 | 10.2795097268466 |
| | . | . |
| | . | . |
| | . | . |
| [496,] | 9.65764950477764 | 10.5993086125531 |
| [497,] | 9.19906314438144 | 10.3086484695036 |
| [498,] | 9.2670053927737 | 10.5778826652537 |
| [499,] | 10.3728250087829 | 11.1154385480222 |
| [500,] | 9.07860308030324 | 10.3281566737159 |

De los cuales, fue posible determinar que exactamente 452 intervalos contienen el verdadero valor del parámetro $\beta_1 = 10$, que equivale a 0.904, como era de esperarse, dado el nivel de significancia especificado inicialmente de un 90%. **Esto permite reflejar la manera en que se acerca la teoría a la práctica.**