Trabajo 2

La base de datos con la que le corresponde trabajar se obtiene como una muestra aleatoria que corresponde a los resultados obtenidos por los jóvenes de todo el país en las pruebas Saber 11 del año 2022. La información aparece agregada por Colegio y solo abarca los colegios de Antioquia. Dicha base contiene las variables: NATURALEZA (Colegio Oficial o No-Oficial), JORNADA (Asociado a los diferentes tipos de jornada de estudio), PROMLECT (Promedio obtenido por el colegio en Lectura Crítica) y PROMMATE (Promedio obtenido por el colegio en Matemáticas).

Pasos previos

Antes de realizar los puntos que se pide, es necesario cargar la base de datos usando alguna herramienta como Python o R. En nuestro caso se usó la primera:

Antes de empezar, se importan las bibliotecas necesarias:

```
# Bibliotecas necesarias
import pandas as pd
import numpy as np
import scipy.stats as stats
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
```

Luego, se obtienen los datos del archivo para realizar el análisis:

```
# Leer archivo csv

raw_data_url = "https://raw.githubusercontent.com/repos-especializacion-
UdeA/estadistica/refs/heads/main/trabajo2/DatosTrabajo2EAE20242.csv"

# Leer el archivo CSV
df = pd.read_csv(raw_data_url)
```

Antes de realizar los cálculos, es bueno conocer si hay valores nulos en cada columna.

```
df.isnull().sum()
```

Como ya se tiene los datos cargados es posible empezar a realizar cada uno de los puntos del taller.

1. **Intervalo de confianza**: Construir en Python un IC para al 95% de confianza necesario para responder:

   ¿Puede afirmarse que el resultado medio obtenido por los colegios en Lectura es superior a 45 puntos? Justifique su respuesta.

   Para construir una estimación del intervalo de confianza para una media poblacional desconocida necesitamos datos de una muestra aleatoria. En este caso, seleccionamos la columna relacionada (**PROMLECT**) con los datos que nos interesan:

   ```
   # Datos del problema
   data_prom_lectura = df.PROMLECT.copy()
   ```

   Procedemos a plantear el problema teniendo en cuenta que cada uno de estos datos es una VA (variable aleatoria) de modo que sea X_i la VA relacionada con el promedio obtenido en la prueba de lectura por el *i-esimo* colegio.

   Las VAs $X_1, X_2, X_i, \dots, X_n$ (con $n = 1807$) constituyen la muestra aleatoria (MA) del promedio obtenido en la prueba.

El primer paso consiste en obtener la media y la desviación estándar de la muestra:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

En Python esto se hace así:

```
# Calculo de la media y la desviacion estandar muestral
n = data_prom_lectura.shape[0]
X = data_prom_lectura.mean()
S = data_prom_lectura.std()

# Despliegue de los resultados
print(f"n = {n}")
print(f"X = {X:.5f}")
print(f"S = {S:.5f}")
```

El resultado mostrado en pantalla fue el siguiente:

```
n = 1807
X = 50.32153
S = 7.02803
```

De este modo tenemos que:

- $n = 1807$
- $X = 50.32153$
- $S = 7.02803$

Como para la muestra aleatoria la distribución no es normal, el número de datos es grande ($n \geq 30$) y la varianza (σ^2) es desconocida, el intervalo de confianza esta dado por:

$$IC = \bar{X} \pm Z_{\alpha/2}\frac{S}{\sqrt{n}}$$

Donde de acuerdo al enunciado, el nivel de confianza es del 95% de modo que:

$$NC = (1-\alpha) * 100 \rightarrow \alpha = 1 - \frac{NC}{100} = 1 - \frac{95}{100} = 0.05$$

De modo que la expresión, con $\alpha = 0.05$, para el IC queda:

$$IC = \bar{X} \pm Z_{\alpha/2}\frac{S}{\sqrt{n}} = 50.32153 \pm Z_{0.025}\frac{7.02803}{\sqrt{1807}}$$

Donde:

$$Z_{0.025} = 1.95996$$

Con lo cual queda que el IC:

$$IC = 50.32153 \pm (1.95996)\frac{7.02803}{\sqrt{1807}} = 50.32153 \pm 0.32404$$

El código Python que hace lo anterior, se muestra a continuación:

```
NC = 95 # Nivel de confianza
alpha = 1 - NC/100
```

```
Z = stats.norm().ppf(1 - alpha/2)
print(f"Z = {Z:.5f}")

EBM = Z*S/np.sqrt(n)
print(f"EBM = {EBM:.5f}")

IC_left = X - EBM
IC_right = X + EBM
print(f"({X:.5f} - {EBM:.5f},{X:.5f} + {EBM:.5f}) = ({IC_left:.5f},{IC_right:.5f})")
```

El resultado a la salida se muestra a continuación:

```
Z = 1.95996
EBM = 0.32404
(50.32153 - 0.32404,50.32153 + 0.32404) = (49.99748,50.64557)
```

Conclusión:

Como el intervalo en el cual se encuentra la media para la prueba de lectura está en el rango $(49.998, 50.646)$ se puede afirmar con un nivel de confianza del 95% que el resultado promedio en la prueba de lectura es superior a 45 puntos. Esto se puede ver por qué el límite inferior del intervalo de confianza es 49.997 (mayor que 45).

Así, en resumen, dado que todo el intervalo está por encima de 45 puntos, podemos concluir que es muy poco probable que el promedio verdadero sea igual o menor a 45 puntos y, por lo tanto, podemos afirmar que el promedio es superior a 45 puntos con un alto nivel de confianza.

2. ¿Se puede afirmar que el resultado medio obtenido por los colegios en Matemáticas es mayor en los Privados que en los Públicos? Justifique su respuesta mediante una prueba de hipótesis con una significancia del 4%.

   De manera similar vamos obtener los datos de la muestra de la base de datos original:

```
# Informacion de la prueba de matematicas

# Instituciones publicas
notas_mat_publico = df[df['NATURALEZA'] == 'OFICIAL']['PROMMATE'].copy()
# Instituciones privadas
notas_mat_privado = df[df['NATURALEZA'] == 'NO OFICIAL']['PROMMATE'].copy()
```

   Tal y como se muestra en el fragmento de código anterior la muestra asociada al promedio obtenido por los colegios en matemáticas para las instituciones públicas se denominó **notas_mat_publico** y la asociada a las instituciones privadas se llamó **notas_mat_publico**.

   El problema por analizar plantea la siguiente pregunta: ¿Se puede afirmar que el resultado medio obtenido por los colegios en Matemáticas es mayor en los Privados que en los Públicos?

   Para resolverlo vamos a definir las siguientes variables aleatorias para el problema:

   - $\boldsymbol{X_i}$: Promedio obtenido por el *i-esimo* colegio privado en la prueba de matemáticas.
   - $\boldsymbol{Y_j}$: Promedio obtenido por el *j-esimo* colegio público en la prueba de matemáticas.

   De modo que, de acuerdo con la pregunta planteada, si μ_1 es el resultado medio obtenido en matemáticas por los colegios privados y μ_2 es la media obtenida en las pruebas de matemáticas de los colegios públicos lo que se nos pregunta es si $\mu_1 > \mu_2$ lo cual llevado a la forma estándar para el planteamiento de hipótesis queda como: $\mu_1 - \mu_2 > 0$. De este modo tenemos el siguiente **planteamiento de hipótesis**:

- $\boldsymbol{H_0}: \mu_1 - \mu_2 \leq 0$
- $\boldsymbol{H_a}: \mu_1 - \mu_2 > 0$

El siguiente paso consiste en determinar la distribución de la prueba. Donde para el caso de cada tenemos:

	$\boldsymbol{X_i}$	$\boldsymbol{Y_j}$
Descripción	Promedio obtenido por el *i-esimo* colegio privado en la prueba de matemáticas	Promedio obtenido por el *j-esimo* colegio público en la prueba de matemáticas.

Numero de datos	$n_x = 341$
	Donde el código python que hace esto es:
	`nx = notas_mat_privado.shape[0]`
	`print(f"nx = {nx}")`
		$n_y = 1466$
		Donde el código python que hace esto es:
		`ny = notas_mat_publico.shape[0]`
		`print(f"ny = {ny}")`

Distribución	No normal con n grande $n \geq 30$ y varianza (σ^2) desconocida.	No normal con n grande $n \geq 30$ y varianza (σ^2) desconocida.

Teniendo en cuenta la información de la tabla anterior, el estadístico de prueba a emplear para este caso es el siguiente, donde para este problema $\delta_0 = 0$:

$$Z_c = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}} \sim N(0,1)$$

Ahora, procedamos a calcular los parámetros asociados a cada distribución antes de hacer el cálculo del estadístico de prueba:

	Parámetros	Código python

$\boldsymbol{X_i}$	$n_x = 341$
	$\bar{X} = \frac{1}{n_x}\sum_{i=1}^{n_x} X_i = \frac{1}{341}\sum_{i=1}^{341} X_i$
	$S_x = \sqrt{\frac{1}{n_x - 1}\sum_{i=1}^{n_x}(X_i - \bar{X})^2}$
		A continuación, se adjunta el código python que realiza los cálculos requeridos:
		`# Colegio privado`
		`nx = notas_mat_privado.shape[0]`
		`X = notas_mat_privado.mean()`
		`Sx = notas_mat_privado.std()`
		`print(f"nx = {nx}")`
		`print(f"X = {X:.5f}")`
		`print(f"Sx = {Sx:.5f}")`
		La salida de este código se muestra a continuación:
		`nx = 341`
		`X = 52.75367`
		`Sx = 9.59462`

$\boldsymbol{Y_j}$	$n_y = 1466$
	$\bar{Y} = \frac{1}{n_y}\sum_{j=1}^{n_y} Y_j = \frac{1}{1466}\sum_{j=1}^{1466} Y_j$
		El código python que realiza los cálculos requeridos se muestra a continuación:
		`# Colegio publico`
		`ny = notas_mat_publico.shape[0]`
		`Y = notas_mat_publico.mean()`
		`Sy = notas_mat_publico.std()`
		`print(f"ny = {ny}")`

	$S_y = \sqrt{\frac{1}{n_y - 1}\sum_{j=1}^{n_x}(Y_j - \bar{Y})^2}$	`print(f"Y = {Y:.5f}")`
		`print(f"Sy = {Sy:.5f}")`
		
		La salida de este código se muestra es:
		
		`ny = 1466`
		`Y = 45.79127`
		`Sy = 6.80895`

La siguiente tabla resume los resultados anteriores antes de sacar el estadístico:

X_i	Y_j
$n_x = 341$ $\bar{X} = 52.75367$ $S_x = 9.59462$	$n_y = 1466$ $\bar{Y} = 45.79127$ $S_y = 6.80895$

Ahora ya es posible obtener el Z_c con $\delta_0 = 0$:

Calculo manual	**Código**

$$Z_c = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$$
$$Z_c = \frac{(52.75367 - 45.79127) - 0}{\sqrt{\frac{9.59462^2}{341} + \frac{6.80895^2}{1466}}}$$
	El código se muestra a continuación:
	
	`d0 = 0`
	`Zc = (X - Y - d0) / ((Sx**2 / nx + Sy**2 / ny)**0.5)`
	`print(f"Zc = {Zc:.5f}")`
	
	De modo que la salida es:
	
	`Zc = 12.67807`

Finalmente, el último paso es comparar el estadístico de prueba con el valor crítico determinado por el nivel de significancia que se requiere para la prueba, siendo este para el caso: $\alpha = 0.04$. Esto es:

$$P(Z < Z_c) \rightarrow VP < \alpha$$

Usando python para calcular el valor P en este caso tenemos que:

```
VP = 1 - stats.norm.cdf(Zc)
print(f"VP = {VP:e}")
```

Que da como salida:

```
VP = 0.000000e+00
```

Conclusión:
Como $VP < \alpha \rightarrow 0.00 < 0.04$ se rechaza la hipótesis nula $\boldsymbol{H_0}$ aceptándose la hipótesis alternativa $\boldsymbol{H_a}$ de modo que podemos decir con un nivel de significancia del 4% que el resultado medio en las pruebas de Matemáticas es mejor en los colegios privados que en los públicos, esto es $\mu_1 - \mu_2 > 0$.

3. Analice si hay una relación lineal entre las variables “Promedio obtenido por el colegio en Lectura Crítica” y “Promedio obtenido por el colegio en Matemáticas”, mediante una regresión lineal con todos sus respectivos componentes.

Obtención y análisis grafico de los datos

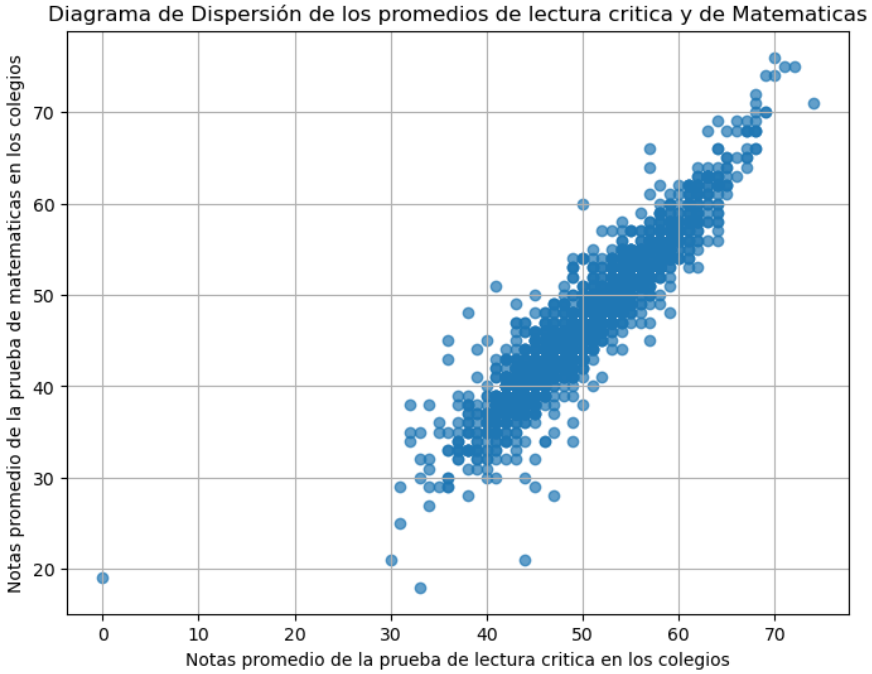
Inicialmente se obtienen los datos y se hace un diagrama de dispersión para visualizar gráficamente la relación entre estos. Para este caso se definieron estos de la siguiente manera:

- **Variable independiente (x)**: Nota promedio de la prueba de lectura crítica para cada colegio.
- **Variable dependiente (y)**: Nota promedio de la prueba de matemáticas para cada colegio.

A continuación se muestra el código python empleado:

```
# Preparar los datos (PROMLECT como variable independiente X y PROMMATE como variable dependiente y)
promedio_notas = df.iloc[:][['PROMLECT','PROMMATE']].copy()
x = promedio_notas.PROMLECT
y = promedio_notas.PROMMATE

# Crear un diagrama de dispersión entre PROMLECT y PROMMATE
plt.figure(figsize=(8, 6))
plt.scatter(x, y, alpha=0.7)

# Añadir etiquetas y título
plt.title('Diagrama de Dispersión de los promedios de lectura critica y de Matematicas')
plt.xlabel('Notas promedio de la prueba de lectura critica en los colegios')
plt.ylabel('Notas promedio de la prueba de matematicas en los colegios')

# Mostrar el gráfico
plt.grid(True)
plt.show()
```

La grafica de dispersión se muestra a continuación:



Obtención del modelo

Ahora vamos a obtener el modelo lineal simple, el cual es de la forma:

$$Y = \beta_0 + \beta_1 X$$

El objetivo es obtener el intercepto y la pendiente que definan la recta que mejor describe la relación entre los datos asociados a x e y.

```
# 1. Preparar los datos
X = x.values.reshape(-1, 1)  # Variable independiente
y = y.values  # Variable dependiente

# 2. Crear el modelo de regresión lineal
modelo = LinearRegression()

# 3. Ajustar el modelo a los datos
modelo.fit(X, y)

# 4. Obtener el coeficiente y la intersección
B1 = modelo.coef_[0]  # Coeficiente de la regresión
B0 = modelo.intercept_  # Intersección (ordenada al origen)

print(f'Pendiente (B1): {B1}')
print(f'Intersepto (B0): {B0}')
```

La salida del código anterior es la siguiente:

```
Pendiente (B1): 1.046945295837803
Intercepto (B0): -5.578739732057144
```

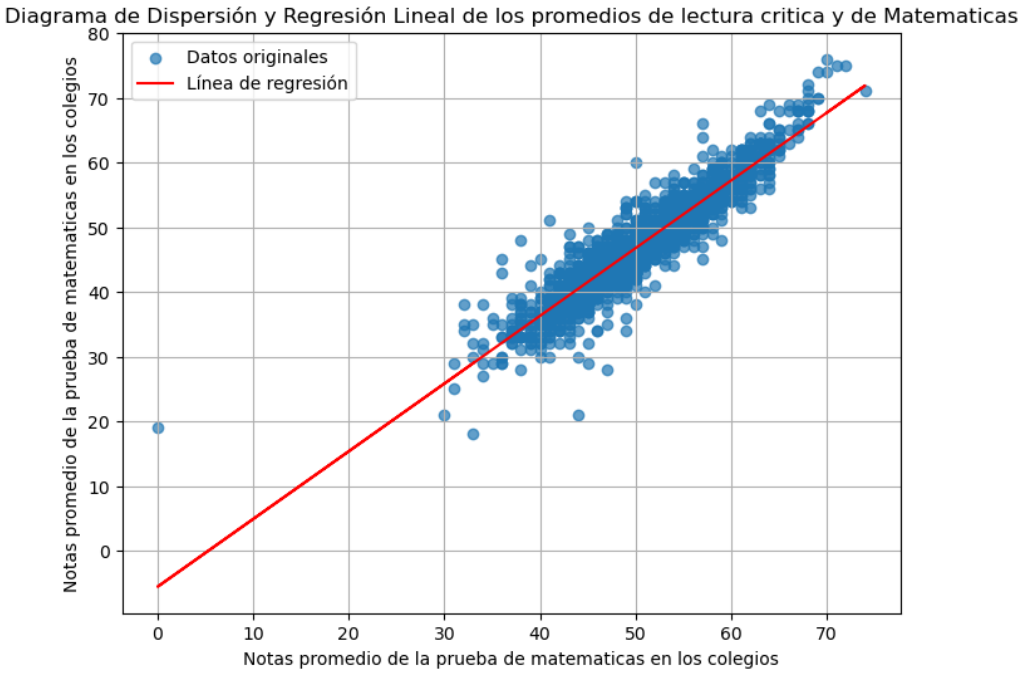
De modo que el modelo lineal queda de la siguiente manera:

$$Y = \beta_0 + \beta_1 X \approx -5.5787 + 1.0469X$$

El siguiente código compara los datos con la aproximación lineal previamente obtenida:

```
# 1. Predecir los valores de y para graficar la línea de regresión
y_pred = modelo.predict(X)

# 2. Graficar el diagrama de dispersión y la línea de regresión
plt.figure(figsize=(8, 6))

# 2.1. Diagrama de dispersión
plt.scatter(X, y, alpha=0.7, label='Datos originales')

# 2.2. Línea de regresión
plt.plot(X, y_pred, color='red', label='Línea de regresión')

# 2.3. Añadir etiquetas y título
plt.title('Diagrama de Dispersión y Regresión Lineal de los promedios de lectura critica y de
Matematicas')
plt.xlabel('Notas promedio de la prueba de matematicas en los colegios')
plt.ylabel('Notas promedio de la prueba de matematicas en los colegios')
plt.legend()
plt.grid(True)
```

```
# 2.4. Mostrar el gráfico
plt.show()
```



Análisis del modelo

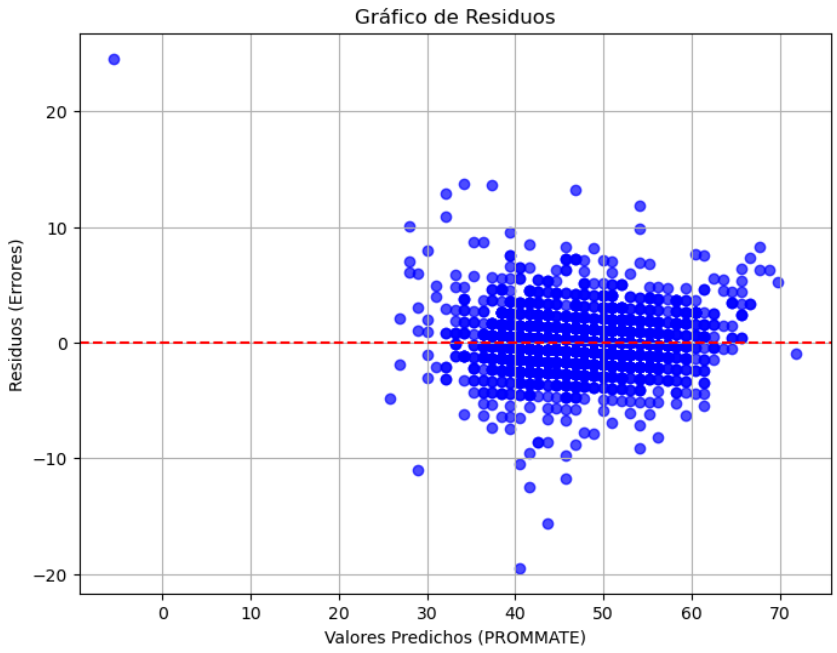
Para mirar que tan bueno es el modelo hay diferentes métricas y gráficas. Una de las gráficas de mayor utilidad tiene que ver con la gráfica de **errores de predicción** (también conocidos como **residuos**) ε_i. La expresión asociada al error se entiende como la diferencia entre el valor observado (y_i) y el valor predicho ($\hat{y}_i$). Esto es:

$$\varepsilon_i = y_i - \hat{y}_i$$

En el siguiente código se realiza la gráfica en cuestión:

```
# 1. Calcular los residuos (errores)
residuos = y - y_pred

# 2. Graficar los residuos (errores) en función de los valores predichos
plt.figure(figsize=(8, 6))
plt.scatter(y_pred, residuos, alpha=0.7, color='blue')

# 2.1. Dibujar una línea horizontal en y=0 para visualizar los errores
plt.axhline(0, color='red', linestyle='--')

# 2.2. Añadir etiquetas y título
plt.title('Gráfico de Residuos')
plt.xlabel('Valores Predichos (PROMMATE)')
plt.ylabel('Residuos (Errores)')
plt.grid(True)

# 2.3. Mostrar el gráfico
plt.show()
```

La grafica se muestra a continuación:



Otras métricas que evalúan el rendimiento del modelo son:

- **Coeficiente de determinación (R^2)**: Indica la proporción de la varianza en la variable dependiente que es explicada por las variables independientes.
- **Error cuadrático medio (MSE)**: Mide la media de los errores al cuadrado.
- **Raíz del error cuadrático medio (RMSE)**: Es la raíz cuadrada del MSE y tiene la misma unidad que la variable dependiente.
- **Error absoluto medio (MAE)**: Mide la media de los errores absolutos, es decir, sin tener en cuenta el signo.

El siguiente fragmento de código muestra cómo realizar estos cálculos:

```
# 1. Calculo de las métricas
r2 = r2_score(y, y_pred)  # Coeficiente de determinación
mse = mean_squared_error(y, y_pred)  # Error cuadrático medio
rmse = np.sqrt(mse)  # Raíz del error cuadrático medio
mae = mean_absolute_error(y, y_pred)  # Error absoluto medio

# 2. Despliegue de las metricas
print(f"R^2 (Coeficiente de determinación): {r2}")
print(f"Error cuadrático medio (MSE): {mse}")
print(f"Raíz del error cuadrático medio (RMSE): {rmse}")
print(f"Error absoluto medio (MAE): {mae}")
```

La salida del fragmento anterior se muestra a continuación:

```
R^2 (Coeficiente de determinación): 0.868118631027635
Error cuadrático medio (MSE): 8.220144718705894
Raíz del error cuadrático medio (RMSE): 2.8670794754777718
Error absoluto medio (MAE): 2.0692130047500217
```

Conclusión:

1. Como en la gráfica de residuos se ve que estos se distribuyen aleatoriamente alrededor de 0 se concluye que el modelo lineal aplicado es apropiado y que por lo tanto si hay una relación lineal entre los promedios de las pruebas de lectura y matemáticas obtenidos por los colegios.

2. Esta misma conclusión se ratifica con el Coeficiente de Determinación (R^2), pues el resultado indica que en un 86% si hay una relación lineal entre los promedios de los resultados.
3. Dado que los valores de los promedios son relativamente altos, mayores a 20, y que el error cuadrático medio es de 8.22, alejado de los promedios, se puede nuevamente afirmar que el modelo de regresión lineal si es válido para concluir que existe una relación lineal entre los promedios de las pruebas obtenidos por los colegios.
4. En conclusión, podemos afirmar que los colegios que tuvieron mejores promedios en la prueba de matemáticas también obtuvieron mejores promedios en la prueba de lectura y que mientras más elevado el puntaje en una prueba más elevado el puntaje en la otra.

Notebook

El notebook se encuentra en el siguiente enlace:

https://github.com/repos-especializacion-UdeA/estadistica/tree/main/trabajo2