Estadística

Proyecto Final

Díaz Sanchéz David
Morales Ramírez Ángel Francisco
Soto Luna Denilson
Zúñiga Galván Diego Antonio

28 de Noviembre del 2023

Regresión Lineal Múltiple

Un modelo de regresión lineal múltiple es un modelo estadístico versátil para evaluar las relaciones entre un destino continuo y los predictores. Los predictores pueden ser campos continuos, categóricos o derivados, de modo que las relaciones no lineales también estén soportadas. El modelo es lineal porque consiste en términos de aditivos en los que cada término es un predictor que se multiplica por un coeficiente estimado. La regresión lineal se utiliza para generar conocimientos para los gráficos que contienen al menos dos campos continuos con uno identificado como el destino y el otro como un predictor.

Supuestos del Modelos de regresión lineal múltiple

Recordemos que partimos de los siguientes supuestos de modelo de regresión múltiple

- $\epsilon \sim N(0, \sigma^2)$
- σ^2 es constante
- Rango de X es completos
- Colinealidad
- Autocorrelación

Aplicaciones del Modelo de Regresión Lineal

- **Predicción** La regresión lineal múltiple se puede utilizar para predecir el valor de una variable en el futuro.
- **Explicación** La regresión lineal múltiple se puede utilizar para explicar la relación entre dos o más variables.
- **Control** La regresión lineal múltiple se puede utilizar para controlar el efecto de una variable sobre otra

Serie de Datos

Los datos para el proyecto de egresión lineal múltiple fueron de la liga Santander (Española) desde la temporada 2014-2015 hasta la temporada 2022-2023, es decir se tomaron en cuenta, las siguientes temporadas

Los datos son valores reales donde x_1, x_2, y son variables continuas en $\mathbb{R}$ y ademas existe independencia entre x_1 y x_2

Temporadas Seleccionadas

1. Temporada 2014-2015
2. Temporada 2015-2016
3. Temporada 2016-2017
4. Temporada 2017-2018
5. Temporada 2018-2019
6. Temporada 2019-2020
7. Temporada 2020-2021
8. Temporada 2021-2022
9. Temporada 2022-2023

Propiedades de los datos

Se tomaron los puntos hechos por equipo en cada temporada además se su respectivo diferencia de goles X_1 y su % de posesión X_2
Los datos fueron extraídos de las siguiente paginas:Whoscored, FBref.De esta manera los datos quedarían:

Ejemplo de datos trabajados

Y	X_1	X_2
41	-16	45.1
51	4	51.6
77	37	50.6
88	50	64.3
60	5	50.6
42	-23	41.9
43	-10	50.2
25	-37	45.5
37	-17	43.1
42	-11	39.9
49	3	51.3
50	-6	41.1

Regresión Lineal Múltiple

Nuestro Modelo de regresión lineal múltiple seria el siguiente

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Valores de las Variables

- $\hat{Y}$ = Valor estimado de Y
- Y = Puntos esperados por temporada
- $\beta_0, \beta_1, \beta_2$ = Son los estimadores muéstrales
- X_1 = Diferencia de Goles
- X_2 = % Posesión por Temporada
- ϵ = Error

Interpretación de Variables

Es decir la variable dependiente Y que intentamos predecir son los puntos que hará un equipo durante una temporada completa mientras que las variables independientes que nos ayudaran a realizar esta predicción son su diferencia de goles (X_1) y % posesión por temporada (X_2)

Elección de Datos

Se eligieron estas variables independientes ya que notamos que la diferencia de goles es un valor mas fiable que otras variables como eran los goles a favor, los goles en contra y los goles esperados por partido, mientras que se eligió la posesión ya que no están tan correlación con la diferencia de goles y es un dato muy polémico respecto a el fútbol muchos piensan que es insignificante mientras que otros piensan que es lo mas importante en el fútbol

Otros posibles modelos

Nota: No agregamos mas variables pues el coeficiente de correlación era alto y en síntesis las v.a. explicaban los mismo puesto que no eran independientes entre si. Ya que de alguna forma si agregamos una variable $X_3 =$ Valor del equipo este esta directamente correlacionado con $X_1 =$ Diferencia de goles

Análisis de Resultados

En este caso se comprobó por el teorema de Gauss-Markov en clase que el método de estimación correcto era por mínimos cuadrados ordinarios, además de hacer el código en clase por que lo para calcular nuestros estimadores simplemente ajustamos nuestros modelo

Código

Importación de librerías

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from sklearn.metrics import r2_score
import matplotlib.pyplot as pltm.
```

Código

Subida de Datos y Asignacion de Variables

```
df = pd.read_excel('datosproyecto.xlsx')
df.head()
X = df[['X1', 'X2']]
y = df['Y']
```

Código

Modelacion

```
model = LinearRegression()
model.fit(X, y)
y_pred = model.predict(X)

df["Y_hat"] = y_pred
df["ERROR"] = df["Y"]-df["Y_hat"]
```

Código

Valor de los estimadores y varianza

```
print("Estimadores para los Parámetros\n")
sigma_2 = (n/(n-2))*ECM
print(f"sigma^2 = {sigma_2}")
betas = model.coef_
B0 = model.intercept_
print('B0 =', B0)
i = 1
for coef in betas:
  print(f"B{i} = {coef}")
  i += 1
```

Código

Evaluacion de la prediccion

```
# Diferencia de gol
x1 = 50
# Posesión
x2 = 50
# Nuevos valores de x1 y x2 para hacer una predicción
valores_dados = [[x1, x2]]
# Realiza la predicción
y_hat = model.predict(valores_dados)
# Muestra la predicción
print('Estimación de puntos:', y_hat[0]) {coef}")
```

Resultados obtenidos

Por tanto el valor de los estimadores en nuestro modelo de regresión lineal múltiple fueron

$$\sigma^2 = 25{,}2212$$

$$\beta_0 = 48{,}7148$$

$$\beta_1 = 0{,}5996$$

$$\beta_2 = 0{,}0673$$

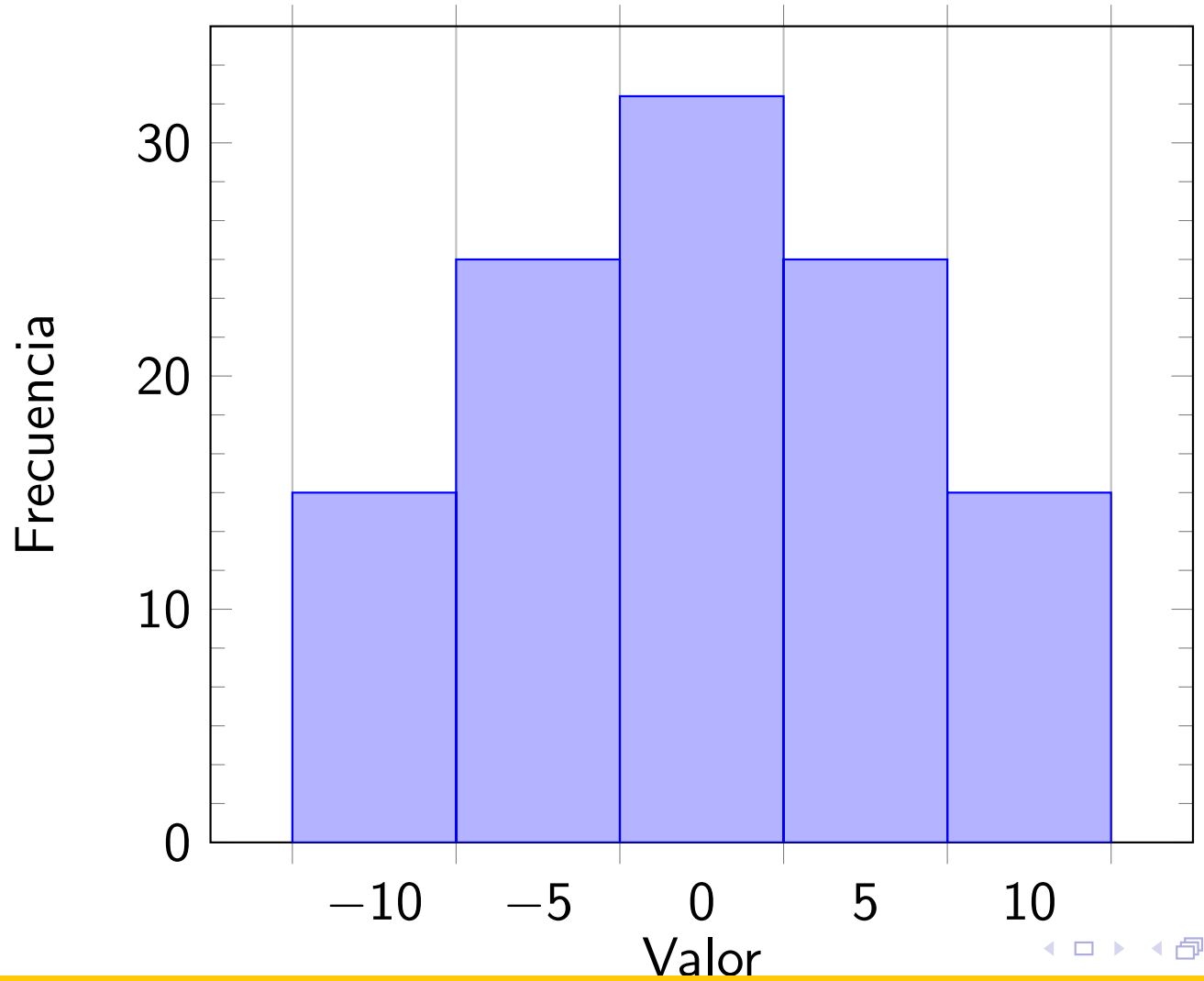
Mientras que el rendmiento del modelo lo podemos observar con

$$EMA = 4{,}0360$$

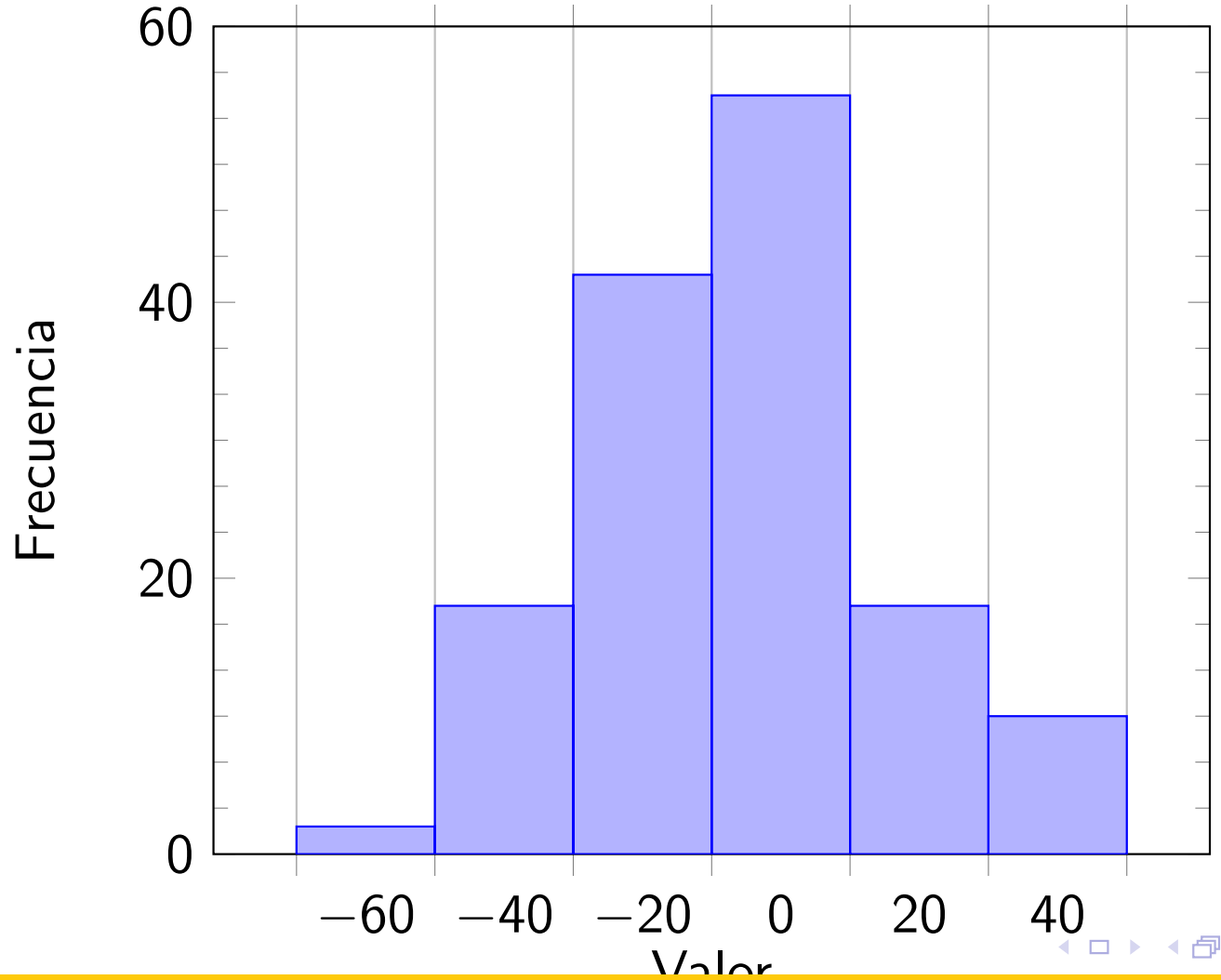
$$ECM = 24{,}9409$$

$$R^2 = 0{,}9151$$

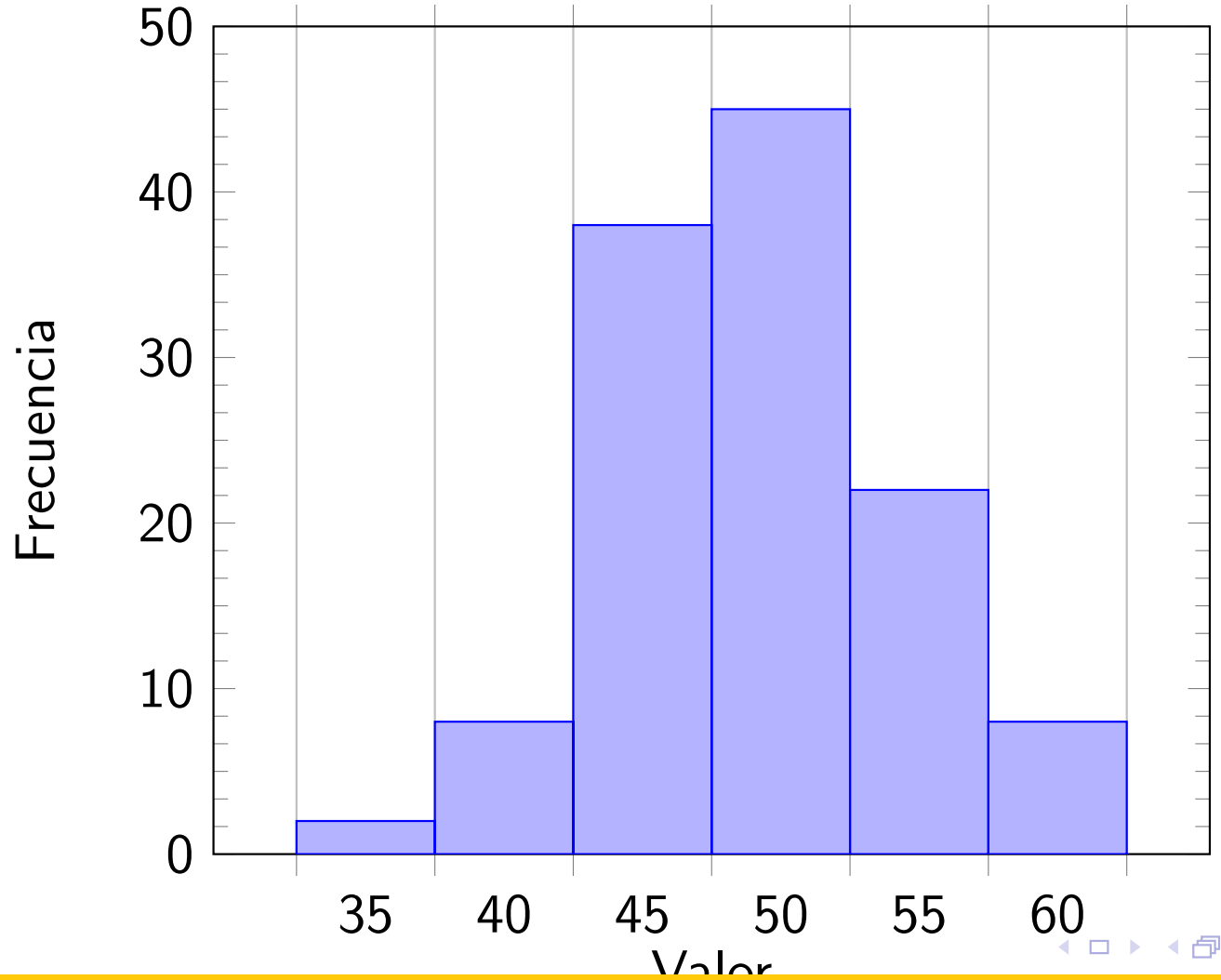
Histograma de Errores



Histograma de Diferencia de Goles



Histograma de Posesión



Conclusión

Como conclusión podemos decir que nuestro modelo de regresión lineal múltiple sirve para hacer una estimación sobre los puntos que hará un equipo en la Liga Española (Santander) en un temporada completa, partiendo de los datos recolectados y aplicando nuestro modelo hecho, agregando mas datos la modelo tendríamos una aproximación mas exacta

Referencias

- IBM documentation. (s. f.). https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=tests-multiple-linear-regression
- WhoScored https://es.whoscored.com/
- FBref Estadísticas e historia del fútbol https://fbref.com/es/