**Instituto Tecnológico y de Estudios Superiores de Monterrey**

# Regresión Lineal

**TC3004B.104 Inteligencia Artificial Avanzada para la Ciencia de Datos I**


Profesores:

*Ivan Mauricio Amaya Contreras*

*Blanca Rosa Ruiz Hernandez*

*Antonio Carlos Bento*

*Frumencio Olivas Alvarez*

*Hugo Terashima Marín*

*Julian Lawrence Gil Soares - A00832272*


6 de septiembre de 2023

Regresión lineal TC3004B.104 Inteligencia Artificial Avanzada para la Ciencia de Datos I

Profesores: Ivan Mauricio Amaya Contreras Blanca Rosa Ruiz Hernandez Antonio Carlos Bento Frumencio Olivas Alvarez Hugo Terashima Marín

Julian Lawrence Gil Soares - A00832272

6 de Septiembre de 2023

In [ ]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

In [ ]:
```
#Lectura de archivo y selecion de variables para analizar

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
from scipy import stats

data = pd.read_csv('/content/drive/MyDrive/Stats/Estatura-peso_HyM.csv')
```

```
     Estatura   Peso Sexo
0        1.61  72.21    H
1        1.61  65.71    H
2        1.70  75.08    H
3        1.65  68.55    H
4        1.72  70.77    H
..        ...    ...  ...
435      1.58  66.39    M
436      1.57  65.89    M
437      1.56  56.48    M
438      1.61  59.16    M
439      1.67  80.87    M

[440 rows x 3 columns]
```

In [ ]:
```
# Seperate by H and M
df_hombres = data[data['Sexo'] == 'H']
df_mujeres = data[data['Sexo'] == 'M']

# Correlation matrix
correlation_matrix = data.corr()
print(correlation_matrix)

# Mean and standard deviation
mean_values = data.mean()
std_deviation = data.std()

# Print the results
print("Mean values:")
print(mean_values)
```

```
print("\nStandard Deviation:")
print(std_deviation)

#Analize regression for men
X_hombres = sm.add_constant(df_hombres['Estatura'])
y_hombres = df_hombres['Peso']
model_hombres = sm.OLS(y_hombres, X_hombres).fit()

#Analize regression for women
X_mujeres = sm.add_constant(df_mujeres['Estatura'])
y_mujeres = df_mujeres['Peso']
model_mujeres = sm.OLS(y_mujeres, X_mujeres).fit()

#verify alpha for men
alpha = 0.03
if model_hombres.f_pvalue < alpha:
  print(model_hombres.f_pvalue)
  print("El modelo es significativo")
else:
  print(model_hombres.f_pvalue)
  print("El modelo no es significativo")

#verify coeficents men
for i, p_value in enumerate(model_hombres.pvalues):
    if p_value < alpha:
      print(p_value)
      print(f"El coeficiente β{i} es significativo")
    else:
      print(p_value)
      print(f"El coeficiente β{i} no es significativo")

#verify alpha for women
if model_mujeres.f_pvalue < alpha:
  print(model_mujeres.f_pvalue)
  print("El modelo es significativo")
else:
  print(model_mujeres.f_pvalue)
  print("El modelo no es significativo")

#verify coeficents women
for i, p_value in enumerate(model_mujeres.pvalues):
    if p_value < alpha:
      print(p_value)
      print(f"El coeficiente β{i} es significativo")
    else:
      print(p_value)
      print(f"El coeficiente β{i} no es significativo")



#results
print("\nModelo para hombres:")
print(model_hombres.summary())
print("\nModelo para mujeres:")
print(model_mujeres.summary())
```

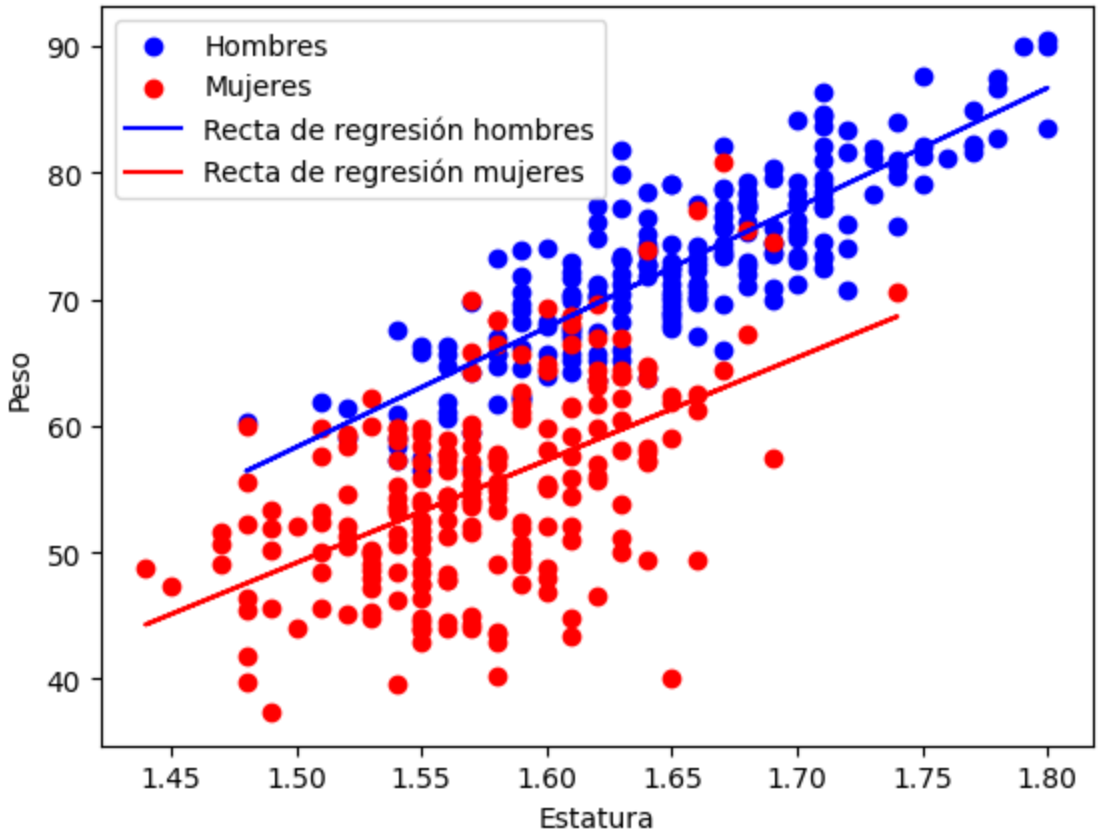
```
#Graph
plt.scatter(df_hombres['Estatura'], df_hombres['Peso'], label='Hombres', col
plt.scatter(df_mujeres['Estatura'], df_mujeres['Peso'], label='Mujeres', col
plt.plot(df_hombres['Estatura'], model_hombres.predict(X_hombres), color='bl
plt.plot(df_mujeres['Estatura'], model_mujeres.predict(X_mujeres), color='re
plt.xlabel('Estatura')
plt.ylabel('Peso')
plt.legend()
plt.show()

residuals_hombres = model_hombres.resid

# Normality
plt.figure(figsize=(12, 4))
plt.subplot(1, 2, 1)
stats.probplot(residuals_hombres, dist="norm", plot=plt)
plt.title("Normality")


# Mean of Zero/Homoscedasticity
plt.figure(figsize=(8, 4))
plt.scatter(model_hombres.fittedvalues, residuals_hombres)
plt.axhline(y=0, color="red")
plt.xlabel("Values")
plt.ylabel("Residuals")
plt.show()
```

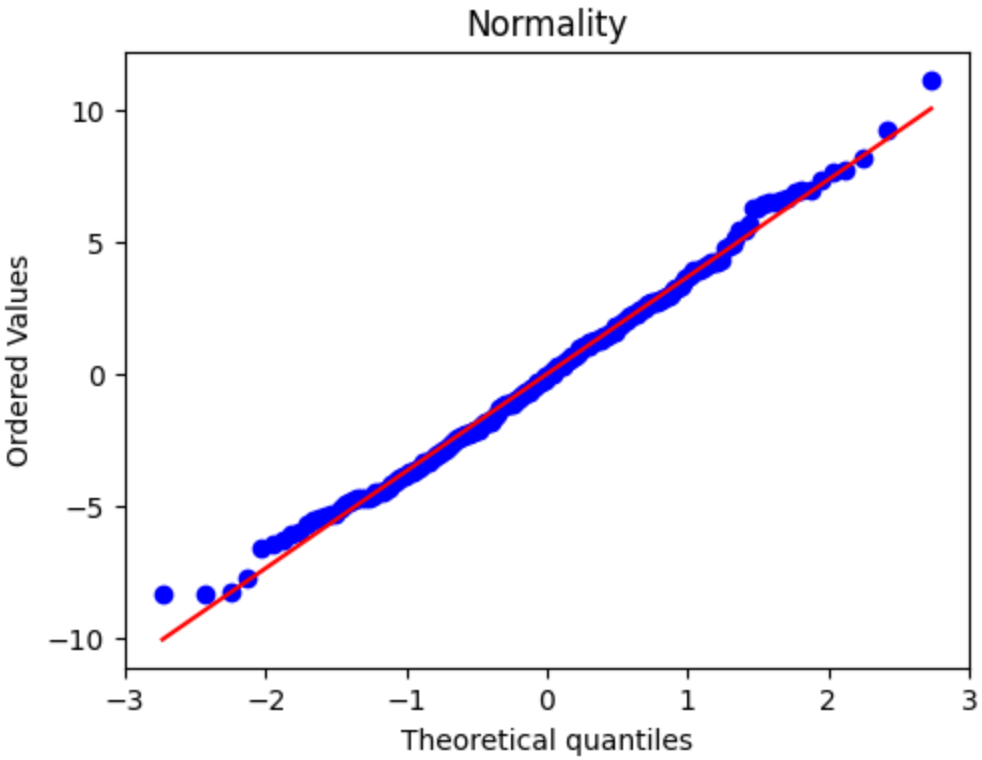
```
<ipython-input-14-3b51ed91bed4>:6: FutureWarning: The default value of numer
ic_only in DataFrame.corr is deprecated. In a future version, it will defaul
t to False. Select only valid columns or specify the value of numeric_only t
o silence this warning.
  correlation_matrix = data.corr()
<ipython-input-14-3b51ed91bed4>:10: FutureWarning: The default value of nume
ric_only in DataFrame.mean is deprecated. In a future version, it will defau
lt to False. In addition, specifying 'numeric_only=None' is deprecated. Sele
ct only valid columns or specify the value of numeric_only to silence this w
arning.
  mean_values = data.mean()
<ipython-input-14-3b51ed91bed4>:11: FutureWarning: The default value of nume
ric_only in DataFrame.std is deprecated. In a future version, it will defaul
t to False. In addition, specifying 'numeric_only=None' is deprecated. Selec
t only valid columns or specify the value of numeric_only to silence this wa
rning.
  std_deviation = data.std()
```

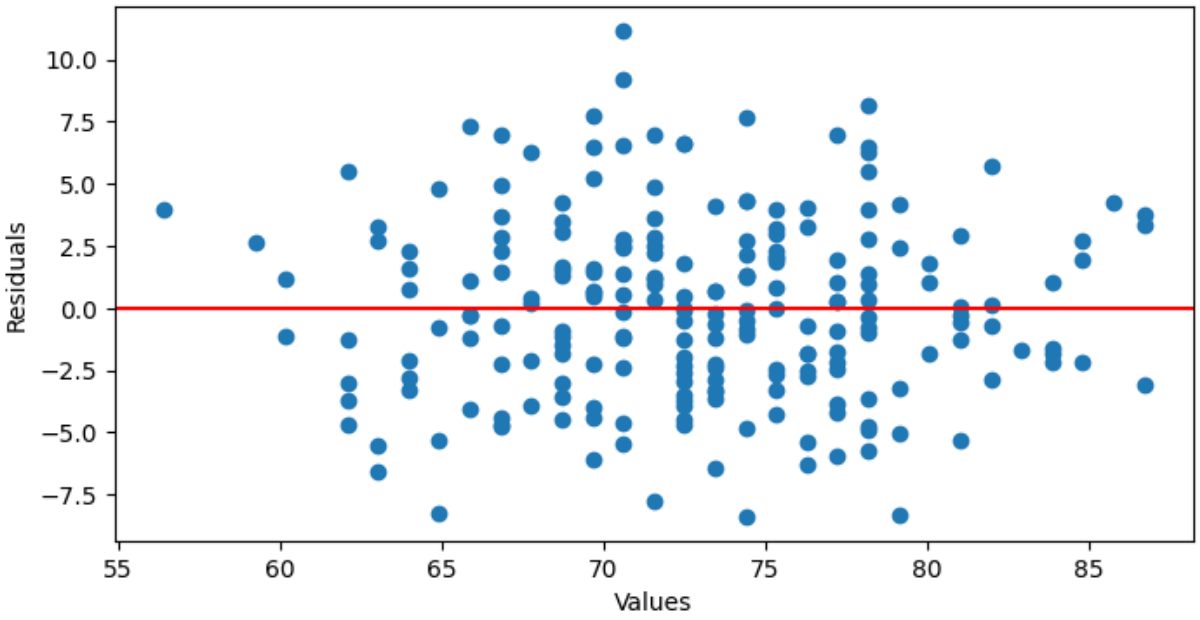
```
          Estatura      Peso
Estatura  1.000000  0.803245
Peso      0.803245  1.000000
Mean values:
Estatura     1.613341
Peso        63.970545
dtype: float64

Standard Deviation:
Estatura     0.069292
Peso        11.541615
dtype: float64
1.0635322278575254e-61
El modelo es significativo
1.4104534541878238e-27
El coeficiente β0 es significativo
1.0635322278576732e-61
El coeficiente β1 es significativo
5.99751674973596e-17
El modelo es significativo
5.342762065016259e-07
El coeficiente β0 es significativo
5.997516749736435e-17
El coeficiente β1 es significativo

Modelo para hombres:
                            OLS Regression Results
==============================================================================
==
Dep. Variable:                   Peso   R-squared:                       0.7
17
Model:                            OLS   Adj. R-squared:                  0.7
16
Method:                 Least Squares   F-statistic:                      55
2.7
Date:                Wed, 06 Sep 2023   Prob (F-statistic):           1.06e-
61
Time:                        22:32:03   Log-Likelihood:                -597.
71
No. Observations:                 220   AIC:                             119
9.
Df Residuals:                     218   BIC:                             120
6.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
--
const        -83.6845      6.663    -12.559      0.000     -96.818     -70.5
51
Estatura      94.6602      4.027     23.509      0.000      86.724     102.5
96
==============================================================================
```

```
==
Omnibus:                        2.048   Durbin-Watson:                   2.0
56
Prob(Omnibus):                  0.359   Jarque-Bera (JB):                2.1
27
Skew:                           0.215   Prob(JB):                        0.3
45
Kurtosis:                       2.782   Cond. No.                         6
0.7
================================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.

Modelo para mujeres:
                            OLS Regression Results
================================================================================
==
Dep. Variable:                   Peso   R-squared:                       0.2
75
Model:                            OLS   Adj. R-squared:                  0.2
72
Method:                 Least Squares   F-statistic:                     82.
73
Date:                Wed, 06 Sep 2023   Prob (F-statistic):           6.00e-
17
Time:                        22:32:03   Log-Likelihood:                -727.
98
No. Observations:                 220   AIC:                             146
0.
Df Residuals:                     218   BIC:                             146
7.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
--------------------------------------------------------------------------------
--
const        -72.5604     14.041     -5.168      0.000    -100.233     -44.8
87
Estatura      81.1491      8.922      9.096      0.000      63.565      98.7
33
================================================================================
==
Omnibus:                        0.523   Durbin-Watson:                   1.8
06
Prob(Omnibus):                  0.770   Jarque-Bera (JB):                0.5
19
Skew:                          -0.116   Prob(JB):                        0.7
72
Kurtosis:                       2.946   Cond. No.                         6
9.2
```

```
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
```






In [ ]:
```
'''
De mis valores de alfa se puede determinar si mi modelo es significativo, es
Correlación que se pueda utilizar para realizar una predicción. Los coeficie
Los coeficientes son los valores por los que voy a multiplicar mi variable i
si hay una relación entre mis datos y mis predicciones.

Para determinar si modelo es significativo hice una comparación entre mi alf
es significativo y mi alfa es el valor que determina para mi prueba de hipót
alfa, así que ambos son significativos. Para determinar si mis coeficientes
mis coeficientes son significativos.

Los residuos se refieren a la diferencia entre nuestros valores esperados y
El qqplot nos sirve para observar si nuestros residuos siguen una línea de t
que podemos decir que como nuestros datos obtenidos siguen una tendencia nor
La media cero quiere decir que nuestras predicciones por encima y por debajo
representa el valor 0 y podemos observar que nuestros puntos se ven distribu
puedo inferir que nuestras predicciones cumplen con la media de 0. Para la h
también a la varianza que se puede observar en nuestros residuos. Pero ahora
la varianza en el eje x (valores). Como podemos observar nuestros puntos est
que nuestros datos cumplen con la homoscedasticidad.
'''
```

Out[ ]:
```
'\nDe mis valores de alpha se puede determinar si mi modelo es significativ
o, esto quiere decir que puedo observar si mis datos de verdad tienen una
\ncorrelacion que se pueda utilizar para realizar una prediccion. Los coefi
cientes tambien sirven para determinar si mi modelo es significativo o n
o.\nLos coeficientes son los valores por los que voy a multiplicar mi varia
ble independiente asi que tambien son importantes para determinar \nsi hay
una relacion entremis datos y mis prediciones.\n\nPara determinar si miodel
o es significativo hice una comparacion entre mi alfa y mi valor p, si el v
alor de p es cercano a 0 entonces mi modelo \nes significativo y mi alfa es
el valor que determina para mi prueba de hipotesis, en ambos modelos el val
or de p fue menor que el valor de mi\nalfa asi que ambos son significativo
s. Para determinar si mis coeficientes son significativos hice la misma pru
eba y obtuve los mismos resultados,\nmis coeficientes son significativos.\n
\nLos residuos se refieren a la diferencia entre nuestros valores esperados
y nuestros valores obtenidos.\nEl qqplot nos sirve para observar si nuestro
s residuos siguen una linea de tendencia de distribucion normal. En nuestro
caso esto es cierto asi \nque podemos decir que como nuestros datos obtenid
os siguen una tendencia normal contra los datos esperados nuestra predicion
es valida.\nLa media zero quiere decir que nuestras prediciones por encima
y por debajo del valor real se cancelan. En la ultima grafica la linea roja
\nrepresenta el valor 0 y podemos observar que nuestros puntos se ven distr
ibuidos equitativamente por encima y por debajo de la linea roja, asi que\n
podeom inferir que nuestras prediciones cumplen con la media de 0. Para la
homocedasticidad utilizamos la misma grafica, este termino se refiere\ntamb
ien a la varianza que se puede observar en nuestros residuos. Pero ahora en
vez de observar la varianza en el eje y (residuos), observamos \nla varianz
a en el eje x (valores). Como podemos observar nuestros puntos estan bien d
istribuidos a lo largo del eje x asi que podemos inferir\nque nuestros dato
s cumplen con la homocedasticidad.\n'
```

In [ ]:
```
'''
Despues de todas las pruebas realizadas anteriormente podemos determinar que
modelo de regresion lineal es valido y puede generar prediciones valiosas.
'''
```