Practica2_CorrelacionRegresionLinealSimple

April 11, 2024

0.1 PRÁCTICA 2 - Estadística Descriptiva Bivariante. Correlación y regresión lineal simple.

Objetivos:

Manejo básico de estadística en Python:

- Análisis descriptivo de una y dos variables
- Búsqueda de relación entre variables
- Realización de análisis estadísticos y generación de informes

En esta práctica vamos a aprender a relacionar dos variables y analizar la regresión lineal simple entre dos variables. Para ello se va a utilizar el dataset `mpg`, utilizado en la Práctica 1 de estadística descriptiva básica. En esta práctica se va a aplicar la librería de Python `scipy` y sus módulos de correlación y regresión lineal.

[ ]:
```
#Importamos las librerías que vamos a utilizar

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

from scipy.stats import pearsonr
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
import statsmodels.api as sm
import statsmodels.formula.api as smf
from statsmodels.stats.anova import anova_lm
```

1. Regresión lineal simple y correlación de dos variables (presión y precipitación, peso y altura, potencia y aceleración).

[ ]:
```
#1a. Generar la matriz de dispersión

mpg = sns.load_dataset('mpg')
```

```
#Quitad los NaN que hay en el dataset


plt.scatter(mpg.horsepower, mpg.acceleration)
```

[ ]:
```
#1b. Regresión lineal simple.

# División de los datos en train y test

X = mpg.horsepower
y = mpg.acceleration

X_train, X_test, y_train, y_test = train_test_split(
                                        X.values.reshape(-1,1),
                                        y.values.reshape(-1,1),
                                        train_size   = 0.95,
                                        random_state = 1234,
                                        shuffle      = True
                                    )

# Creación del modelo

modelo = LinearRegression()
modelo.fit(X = X_train.reshape(-1, 1), y = y_train)

X_train = sm.add_constant(X_train, prepend=True)
modelo = sm.OLS(endog=y_train, exog=X_train,)
modelo = modelo.fit()
print(modelo.summary())
```

[ ]:
```
#1c. Correlación entre dos variables

# Correlación lineal entre las dos variables

corr_test = pearsonr(x = mpg.horsepower, y =  mpg.acceleration)
print("Coeficiente de correlación de Pearson: ", corr_test[0])
```

[ ]:
```
#1d. Calculamos el intervalo de confianza para los coeficientes del modelo

# Intervalos de confianza para los coeficientes del modelo
# ==============================================================================
modelo.conf_int(alpha=0.05)
```

[ ]:
```
#1e. Calculamos las predicciones con intervalo de confianza al 95% de nuestro␣
 ↪modelo previamente calculado

# Predicciones con intervalo de confianza del 95%
```

```
# ============================================================================
predicciones = modelo.get_prediction(exog = X_train).summary_frame(alpha=0.05)
predicciones.head(4)
```

[ ]:
```
#1f. Ploteamos nuestro modelo con su linea de regresión


# Predicciones con intervalo de confianza del 95%
# ============================================================================
predicciones = modelo.get_prediction(exog = X_train).summary_frame(alpha=0.05)
predicciones['x'] = X_train[:, 1]
predicciones['y'] = y_train
predicciones = predicciones.sort_values('x')

# Gráfico del modelo
# ============================================================================
fig, ax = plt.subplots(figsize=(6, 3.84))

ax.scatter(predicciones['x'], predicciones['y'], marker='o', color = "gray")
ax.plot(predicciones['x'], predicciones["mean"], linestyle='-', label="OLS")
ax.plot(predicciones['x'], predicciones["mean_ci_lower"], linestyle='--',␣
 ↪color='red', label="95% CI")
ax.plot(predicciones['x'], predicciones["mean_ci_upper"], linestyle='--',␣
 ↪color='red')
ax.fill_between(predicciones['x'], predicciones["mean_ci_lower"],␣
 ↪predicciones["mean_ci_upper"], alpha=0.1)
ax.legend();
```

2. El dataset mpg tiene varias variables sobre las características de un conjunto de coches:

1. Haz un gráfico de dispersión con potencia y aceleración. ¿Es razonable predecir la potencia de un coche de esta muestra en función de su aceleración, o viceversa? Justifica tu respuesta.
2. Utilizando la recta de regresión adecuada, ¿qué potencia se prevé para un coche que tenga una aceleración de 10 segundos?
3. Utilizando la recta de regresión adecuada, ¿qué aceleración se prevé para un coche que tenga una potencia de 10 Cv?
4. Un coche que tenga una aceleración de 17 segundos y una potencia de 130 Cv, ¿qué aceleración tendría un coche con una potencia de 300 Cv?
5. Distingamos ahora los coches con cuatro y ocho cilindros. ¿Qué potencia se prevé para un coche con cuatro cilindros? ¿Y para un coche de ocho cilindros? ¿Hay diferencia en la aceleración entre estos dos tipos de cilindros?
6. Analiza los residuos de los modelos anteriores
7. Identificar y eliminar los puntos extraños (outliers) en el modelo del apartado b, comprobando su efecto en el análisis estadístico. Guardar los datos en un fichero aparte.
8. Repetir los apartados 2), 3), 4), 5) y 6) con el nuevo modelo simplificado, comparando los resultados con el modelo completo.

9. Realizar un estudio similar, a partir del modelo completo, para hallar una posible relación entre aceleración y peso.
10. Repetir el apartado 9) con las variables desplazamiento y potencia.

[ ]:

[ ]:

[ ]:

Realizar un informe con los análisis realizados de cada uno de los apartados anteriores y exportarlo en un fichero .PDF

- El informe se realizará en un grupos de dos personas analizando los resultados obtenidos en cada uno de los apartados anteriores. Todas las gráficas deberán estar bien maquetadas: título, titulo en los ejes, ejes con un intervalo lógico y leyenda (si es el caso). El código empleado también se deberá incluir.
- En el Moodle se encuentra un trabajo tipo para que tengáis de referencia. El archivo incluirá los nombres de todos autores en el siguiente formato:
- Nombre del archivo: Practica2_ApellidosNombre1_ApellidosNombre2.pdf (extensión obligatoria en .pdf)
- Se avisará en Moodle la fecha límite de entrega de este informe.

[ ]: