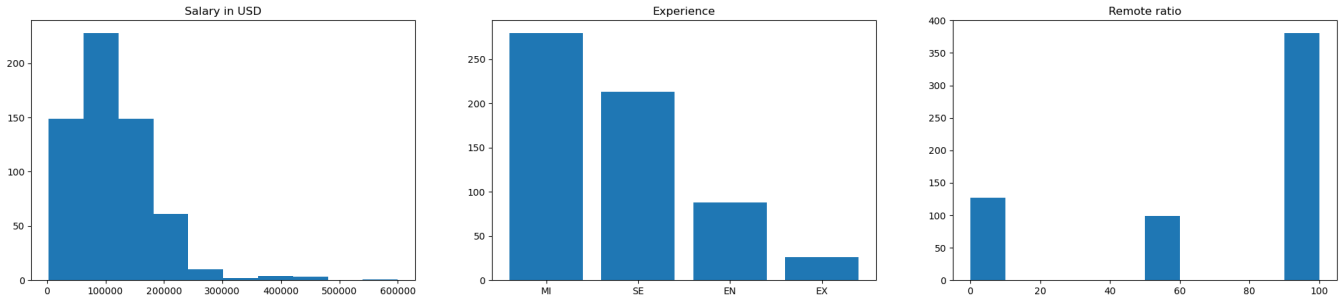
Statistics for Data Science - Assignment 2

Bouruc Petru-Liviu

24/06/2023

1 Exercise 1

For the first exercise, I chose as dataset the Data Science Job Salaries I chose to work with the numerical column salary in USD (because the others were in different currencies). As for the two columns taken into consideration for this experiment, I selected *experience_level* and *remote_ratio*. These are two categorical columns, the first one represents the experience of one individual and has 4 levels(EN, MI, SE and EX), and the second one represents the amount of work done remotely and has 3 levels: 0, 50 or 100. I first checked for 0 values, and there are none. We want to see if these attributes can have an impact on the salary a person receives.



One first important step was to convert my data into factors, because I could get different results if not. This happened because, as explained in lab 6, different levels of degrees of freedom can occur otherwise.

Next stept is to make groups from the data and extract samples, as presented in the laboratory. To reduce the dataset and keep it simple, I chose to concatenate the lowest levels of experience, respectively the highest two, and with these new two, we can make groups with the 3 levels of remote ratio. From each group, I will use a sample set of 32.

Before performing ANOVA analysis, we had to check for the homogenity of variance, using Barlett test (in the manual implementation I used the spicy documentation). The test gave me a p-value of 0.0033, so the data is not homogenous and we reject the the null hypothesis.

For the two-way ANOVA analysis, I used the formula: $salary\_in\_usd \sim Experience \cdot Remote$.

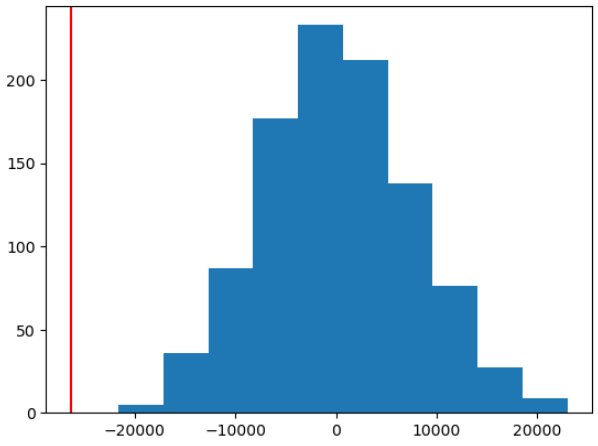
	sum_sq	df	F	PR(>F)
Experience	2.132797e+11	1.0	59.080065	8.167342e-13
Remote	2.263317e+10	1.0	6.269555	1.313418e-02
Experience:Remote	8.946609e+08	1.0	0.247828	6.191907e-01
Residual	6.786822e+11	188.0	NaN	NaN

In the ANOVA analysis, we can observe that that the null hypothesis is rejected for the factors *Remote* and *Experience* (having the p-values p-values 0.013134, respectively a very small one), compared to the p-value of *Experience* : *Remote* which is 0.61919, meaning that it is less significant over the salary than the other two.

In the following steps, I split the data into two groups: entry level jobs and middle level jobs.

Firstly, I used a Shapiro-Wilk test to see if that the two groups are normally distributed. Running the test on the two groups, it gave me for the both of them a very small p-value, meaning that the data is not normally distributed.

In order to see if there are differences between the means of the two groups, we used the t-test, giving a p-value of 0.0004 so we reject the null hypothesis. This suggests that there is a difference between the means of the two groups. To check this, I implemented permutation test, using the predefined function from spicy lib. For the implementation, I used the example from the spicy documentation. We can see that after the permutation, there is a significant statistical difference between the two groups. The p-value for this test is $0.0009 < 0.05$. Looking at the permutation test and at the t-test, we can observe that the values are quite similar, meaning that

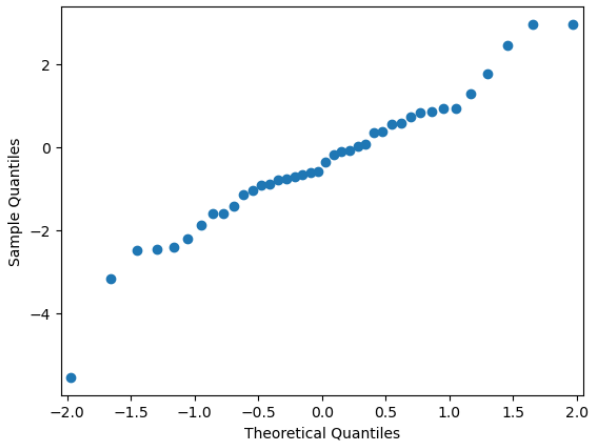


2 Exercise 2

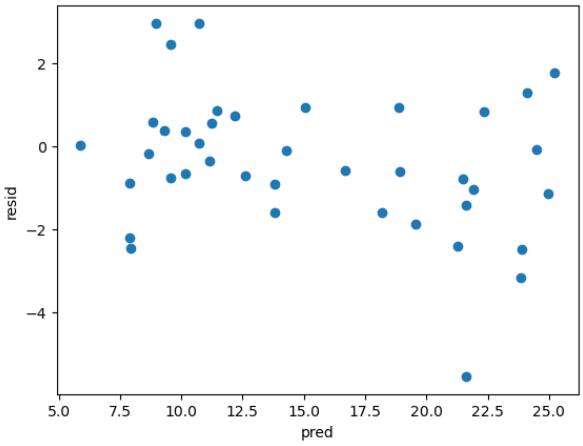
For the second exercise, I chose as dataset the Advertising dataset, which says how well a product is being sold based on the money spent for marketing on different platforms. My data consists in 3 attributes which can determine the sales of a product: money spent on TV, on Newspapers and on Radio. There are no missing values.

I first applied a simple Linear Regression from the sklean module, with a training and test data obtained from my dataset. To see if my model fitted well, I computed the R^2 score. A closer to 1 R^2 score means that the model fits very well the data (1 is perfect). We managed to obtain a score of 0.91 which means that our model learn the data very well.

Looking at the residuals, we can see that performing a Shapiro test we get a p-value of 0.360, which means that our data is normally distributed, rejecting the null hypothesis. The QQ-plot confirms what the the data is normally distributed.



We also have to look at the plot containing residuals and the predictions. The residuals are slightly spread on the the y=0 axis, which means homoscedasticity condition is met.

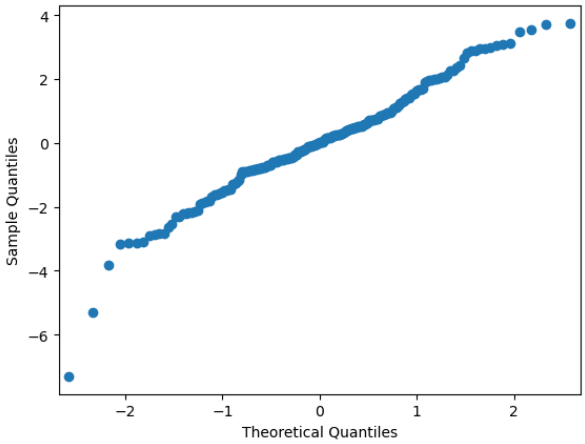


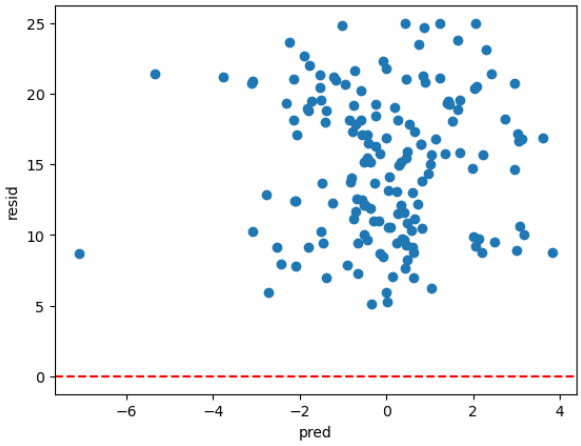
For the ANOVA analysis, I tried in this exercise to add the factors, using the formula: $Sales \sim TV + Radio + Newspaper$.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.903
Model:                            OLS   Adj. R-squared:                  0.901
Method:                 Least Squares   F-statistic:                     605.4
Date:                Sat, 24 Jun 2023   Prob (F-statistic):           8.13e-99
Time:                        16:22:05   Log-Likelihood:                -383.34
No. Observations:                 200   AIC:                             774.7
Df Residuals:                     196   BIC:                             787.9
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.6251      0.308     15.041      0.000       4.019       5.232
TV             0.0544      0.001     39.592      0.000       0.052       0.057
Radio          0.1070      0.008     12.604      0.000       0.090       0.124
Newspaper      0.0003      0.006      0.058      0.954      -0.011       0.012
==============================================================================
Omnibus:                       16.081   Durbin-Watson:                   2.251
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               27.655
Skew:                          -0.431   Prob(JB):                     9.88e-07
Kurtosis:                       4.605   Cond. No.                         454.
==============================================================================
...
TV           1.0  4512.435170  4512.435170  1634.211538  4.960902e-97
Radio        1.0   502.338264   502.338264   181.925492  9.267678e-30
Newspaper    1.0     0.009286     0.009286     0.003363  9.538145e-01
Residual   196.0   541.201230     2.761231          NaN           NaN
```

In the ANOVA analysis PR(>F) are the the p-value for the F-tests for each factor. These tests are used to see if the factors has significant influence on the sales or not. We can observe that only Newspapers have a p-value > 0.05, which means that it is not enough evident to say that it is significant. The other two have a value < 0.05, which means they are significant for the prediction of the sales.

Performing the Shapiro test on the residuals, it computes a p-value of 0.0015, meaning that the data is rejects the null hypothesis. Looking at the QQ-plot and at the way residuals are scattered in the predictions vs redisuals plot, we can confirm that.





3 Code

3.1 Exercise 1

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
from statsmodels.formula.api import ols
from scipy.stats import bartlett, distributions, shapiro, ttest_ind,
    permutation_test

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score


data = pd.read_csv('ds_salaries.csv')
used_data = data[['salary_in_usd', 'experience_level', 'remote_ratio']]
used_data.isna().sum()

fig, axs = plt.subplots(1, 3, figsize=(25, 5))

plt.subplot(1, 3, 1)
plt.title("Salary in USD")
plt.hist(used_data['salary_in_usd'])

plt.subplot(1, 3, 2)
plt.title("Experience")
plt.bar(used_data['experience_level'].unique(),
    used_data['experience_level'].value_counts())
```

```
plt.subplot(1, 3, 3)
plt.title("Remote ratio")
plt.hist(used_data['remote_ratio'])

used_data['Experience'] = used_data['experience_level'].apply(lambda x: 0 if x
    in ('EN', 'MI') else 1)
used_data['Remote'] = used_data['remote_ratio'].apply(lambda x: 0 if x == 0 else
    (1 if x == 50 else 2))

sample1 = used_data[(used_data['Experience'] == 0) & (used_data['Remote'] ==
    0)].sample(n=32, random_state=7)
sample2 = used_data[(used_data['Experience'] == 0) & (used_data['Remote'] ==
    1)].sample(n=32, random_state=7)
sample3 = used_data[(used_data['Experience'] == 0) & (used_data['Remote'] ==
    2)].sample(n=32, random_state=7)
sample4 = used_data[(used_data['Experience'] == 1) & (used_data['Remote'] ==
    0)].sample(n=32, random_state=7)
sample5 = used_data[(used_data['Experience'] == 1) & (used_data['Remote'] ==
    1)].sample(n=32, random_state=7)
sample6 = used_data[(used_data['Experience'] == 1) & (used_data['Remote'] ==
    2)].sample(n=32, random_state=7)

salaries = [sample1["salary_in_usd"], sample2["salary_in_usd"],
    sample3["salary_in_usd"], sample4["salary_in_usd"], sample5["salary_in_usd"],
    sample6["salary_in_usd"]]

stat, pvalue = bartlett(*salaries)
print(stat, pvalue)

r = len(salaries)
n = 0
N = np.empty(r)
for i in range(r):
    N[i] = len(salaries[i])
    n += N[i]

Si_2 = np.empty(r)
S_2 = 0
for i in range(r):
    xi_bar = np.mean(salaries[i])
    Si_2[i] = np.sum((salaries[i] - xi_bar)**2) / (N[i] - 1)
    S_2 += (N[i]-1) * Si_2[i]
S_2 = S_2 / (n - r)

numarator = (n*1 - r) * np.log(S_2) - np.sum(((N - 1)*np.log(Si_2)))
numitor = 1 + 1/(3*(r - 1)) * ((np.sum(1/(N - 1))) - 1/(n - r))
```

```
k_2 = numarator / numitor
pval = distributions.chi2.sf(k_2, r-1)
print(k_2, pval)

sampled_data = [sample1, sample2, sample3, sample4, sample5, sample6]
sampled_data = pd.concat(sampled_data, ignore_index=True)

model = ols('salary_in_usd ~ Experience * Remote', data=sampled_data).fit()
sm.stats.anova_lm(model, typ=2)

first_group = used_data[used_data['experience_level'] == 'EN']['salary_in_usd']
second_group = used_data[used_data['experience_level'] == 'MI']['salary_in_usd']

shapiro_stat_1, shapiro_p_value_1 = shapiro(first_group)
print(shapiro_stat_1, shapiro_p_value_1)
shapiro_stat_2, shapiro_p_value_2 = shapiro(second_group)
print(shapiro_stat_2, shapiro_p_value_2)

t_stat, t_p_value = ttest_ind(first_group, second_group)
print(t_stat, t_p_value)

def statistic(x, y, axis):
    return np.mean(x, axis=axis) - np.mean(y, axis=axis)

res = permutation_test((first_group, second_group), statistic, n_resamples=1000,
    vectorized=True, alternative='less')
print(res.pvalue)

plt.hist(res.null_distribution)
plt.axvline(statistic(first_group, second_group, 0), color='red')
plt.show()
```

3.2 Exercise 2

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
from statsmodels.formula.api import ols
from scipy.stats import bartlett, distributions, shapiro, ttest_ind,
    permutation_test

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
```

```
data=pd.read_csv('advertising.csv')
print(data.isnull().sum())

y = data['Sales']
X = data.drop('Sales',axis=1)

X_train, X_test, y_train, y_test = train_test_split(X,y, train_size = 0.8)

lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
predictions = lr_model.predict(X_test)

r2_score(y_test, predictions)

residuals = y_test - predictions
stat, p_value = shapiro(residuals)
print(stat, p_value)

fig = sm.qqplot(residuals)
plt.show()

plt.scatter(predictions, residuals)
plt.xlabel('pred')
plt.ylabel('resid')
plt.show()

model_sm = sm.formula.ols('y ~ TV + Radio + Newspaper', data=data)
results = model_sm.fit()
print(results.summary())
print(sm.stats.anova_lm(results))

stat, p_value = shapiro(results.resid)
print(stat, p_value)

fig = sm.qqplot(results.resid)
plt.show()

plt.scatter(results.resid, results.fittedvalues)
plt.xlabel('pred')
plt.ylabel('resid')
plt.show()
```