Innovative Assignment
Regression Analysis

Course Code - 2HSOE52

Course Name - Introduction to Econometrics

Tutorial Batch - T2

Submitted By,

18BCE085 – Parth Jasani

Group Members

18BCE081 - Ishika Shah

18BCE083 - Jaineet Shah

18BCE084 - Kaushal Jani

18BCE085 - Parth Jasani

18BCE089 - Pranav Kansagara

Index

Contribution:

- Selection of Sample, Code, General Analysis and Conclusion: All
- Elaborative Analysis on Sign of Beta Coefficient: 18CE081
- Elaborative Analysis on Value of Beta Coefficient: 18BCE083
- Elaborative Analysis on P-value: 18BCE084
- Elaborative Analysis on R Squared Value: 18BCE085
- Elaborative Analysis on F Statistic Probability Value: 18BCE089
- Elaborative Analysis on Skewness and Kurtosis: All

Language used for Analysis:

- Python

Parameters for Analysis:

1. Model Estimation: Beta Coefficient (Value and Sign)
2. Model Evaluation: P-value, R squared value, F statistic value
3. Model Distribution: Skewness, Kurtosis

Population:

X Y	80	100	120	140	160	180	200	220	240	260
Weekly Family Consumption Expenditure Y, $	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	106	118	135	145	157	175	180
	_	88	_	113	125	140	_	160	189	185
	_	_	_	115	_	_	_	162	_	191

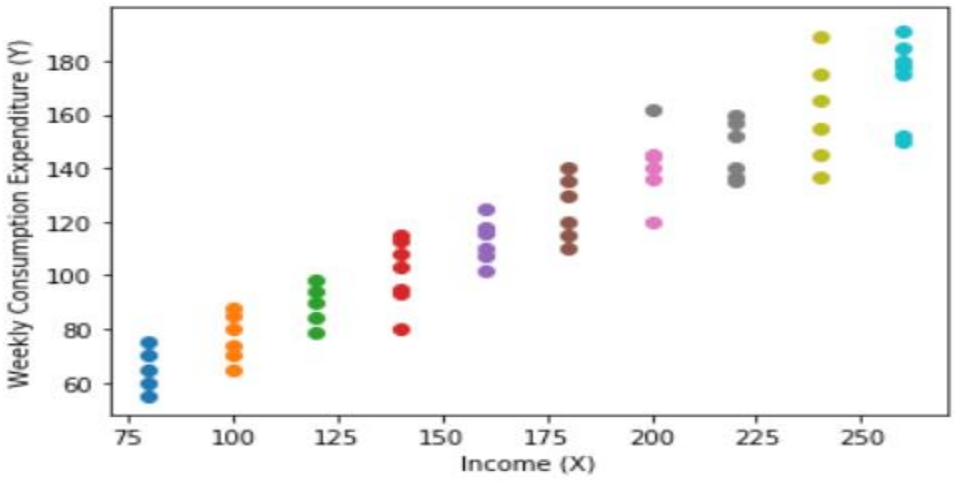
Scatter Plot of Population:

- **Code for Scatter Plot**

```
In [3]: # Population Scatter Plot
from matplotlib import pyplot as plt

x = [[80, 80, 80, 80, 80],
     [100, 100, 100, 100, 100, 100],
     [120, 120, 120, 120, 120],
     [140, 140, 140, 140, 140, 140, 140],
     [160, 160, 160, 160, 160, 160],
     [180, 180, 180, 180, 180, 180],
     [200, 200, 200, 200, 200],
     [220, 220, 220, 220, 220, 220, 200],
     [240, 240, 240, 240, 240, 240],
     [260, 260, 260, 260, 260, 260, 260]]
y = [[55, 60, 65, 70, 75],
     [65, 70, 74, 80, 85, 88],
     [79, 84, 90, 94, 98],
     [80, 93, 95, 103, 108, 113, 115],
     [102, 107, 110, 116, 118, 125],
     [110, 115, 120, 130, 135, 140],
     [120, 136, 140, 144, 145],
     [135, 137, 140, 152, 157, 160, 162],
     [137, 145, 155, 165, 175, 189],
     [150, 152, 175, 178, 180, 185, 191]]
plt.xlabel("Income (X)")
plt.ylabel("Weekly Consumption Expenditure (Y)")
for i in range(10):
    plt.scatter(x[i], y[i])
```

- **Scatter Plot**



Here, from the scatter plot, it is clear that given data is linearly distributed. Hence, our model must be linear in nature. So, Weekly Consumption, Y = (B0) + (B1) * Income + Error Term (e) is our model.

Sample - 1:

Income (X)	Weekly Family Consumption Expenditure (Y)
80	65
100	80
120	90
140	103
160	118
180	135
200	140
220	152
240	155
260	178

Here, Sample Regression Function is:

Weekly Consumption, $y_i = \hat{\beta_0} + \hat{\beta_1} * x_i + \hat{e_i}$

- **Regression Code of Sample-1**

In [2]:
```
# Sample - 1
import pandas as pd
import numpy as np
from sklearn import datasets, linear_model
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
from scipy import stats

income=np.array([80,100,120,140,160,180,200,220,240,260])
weekly_consumption=np.array([65,80,90,103,118,135,140,152,155,178])

temp = sm.add_constant(income)
est = sm.OLS(weekly_consumption, temp)
est2 = est.fit()
print(est2.summary())
```

- **Regression Result of Sample-1**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.987
Model:                            OLS   Adj. R-squared:                  0.986
Method:                 Least Squares   F-statistic:                     631.9
Date:                Sat, 07 Nov 2020   Prob (F-statistic):           6.71e-09
Time:                        12:25:01   Log-Likelihood:                -27.743
No. Observations:                  10   AIC:                             59.49
Df Residuals:                       8   BIC:                             60.09
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         19.6000      4.283      4.576      0.002       9.723      29.477
x1             0.6000      0.024     25.138      0.000       0.545       0.655
==============================================================================
Omnibus:                        3.018   Durbin-Watson:                   1.988
Prob(Omnibus):                  0.221   Jarque-Bera (JB):                0.512
Skew:                          -0.386   Prob(JB):                        0.774
Kurtosis:                       3.796   Cond. No.                         561.
==============================================================================
```

Regression Analysis of Sample - 1:

1. **Sign of Beta Coefficient:**
    - If the sign of the beta coefficient is positive, it means for every unit increase in the explanatory variable, the explained variable increases by the value of the beta coefficient. If the sign of the beta coefficient is negative, it means for every unit increase in the explanatory variable, the explained variable decreases by the value of the beta coefficient.
    - As per Keynes empirical consumption law as income increases consumption also increases but less than proportionately.
    - Here, the value of the beta coefficient is **+0.60** and the value of the constant is **+19.6**.
    - Hence, in our sample, as the income increases, consumption increases by **60%** (considering only beta1).

2. **Value of Beta Coefficient:**
    - In general, the higher the value of beta coefficient the better the explanatory variable is. Also, t-value of a beta coefficient determines if the beta coefficient is significant or not. The more the t value is closer to 0, the lesser significant the corresponding beta coefficient is.
    - In our sample, beta coefficient, beta1 has a value of **0.6** with t-value **25.138**, and beta0 has a value of **19.6**, with a t-value of **4.576**. Hence, by the t values we can determine beta1 is highly significant, while beta0 is less significant.

3. **P-value:**
    - P value stands for "probability value", it is used to check whether an explanatory variable is significant or not.
    - It is calculated from t statistic.
    - Ideally, p value of parameter must be less than type 1 error rate (alpha) so that we can consider it is significant variable, in general it must be less than **0.05**.
    - In our sample, p value of B0 is **0.002** and p value of B1 is equal to **0.0** which is less than 0.05, hence, B1 is a very significant variable and so it significantly explains the change in the variable y, which in our case is the weekly consumption.

4. R Squared value:

- R squared value is used to check how an independent variable explains change in dependent variable.

$$R^2 = Explained\ Sum^2 + Total\ Sum^2$$

- It is also used to check how an independent variable is related to the dependent variable. The general range of R squared value is from 0 to 1.
- R squared value of 0 suggests that the independent variable doesn't explain change in Y, further suggesting insignificant coupling with Y.
- R squared value of 1 suggests complete dependency of independent variable on dependent variable, showing complete coupling between the two.
- The thumb rule is, a R squared value greater than **0.60** suggests our model is well fitted on sample.
- In our sample, the R squared value was found to be **0.987**, suggesting high coupling between the dependent and independent variables.

5. F-Statistic Probability value:

- F-statistic probability value suggests whether regression results (regression model) can be generalised for the entire population.
- Generally, F-statistic probability value is expected to be less than the type 1 error rate, alpha, which we have considered as **0.05**.
- In our sample, F-statistic probability value was found to be **6.71e-09** which is less than 0.05 suggesting that our model is not only valid for this sample it is valid for the entire population.

6. Skewness:

- Skewness is defined as the lack of symmetry in distribution. It reflects the presence of outliers in one versus the other tail. If the mean, median and mode coincide then the distribution is called symmetric.
- If the skewness is **between -0.5 and 0.5**, the data is said to be **fairly symmetrical**.
- If the skewness is **between -1 and – 0.5 or between 0.5 and 1**, the data is said to be **moderately skewed**.
- If the skewness is **less than -1 or greater than 1**, the data is said to be **highly skewed**.
- In our sample, the value of skewness is **-0.386**, which lies in the range [-0.5, 0.5], hence, suggesting that our sample is **fairly symmetrical**.

7. Kurtosis:

- Kurtosis gives a measure of flatness / peakedness of distribution or tailedness of curve. The degree of kurtosis of a distribution is measured relative to that of a normal curve.
- For kurtosis, the general guideline is that if the number is greater than **+1**, the distribution is too peaked, and is called **leptokurtic**. Likewise, a kurtosis of less than **–1** indicates a distribution that is too flat, which is called **platykurtic**. A kurtosis value of **0** indicates a normal distribution, which is called **mesokurtic**.
- In our sample, we found the value of kurtosis to be **+3.796**, which suggests that the distribution is peaked, showing that the curve is **leptokurtic**.

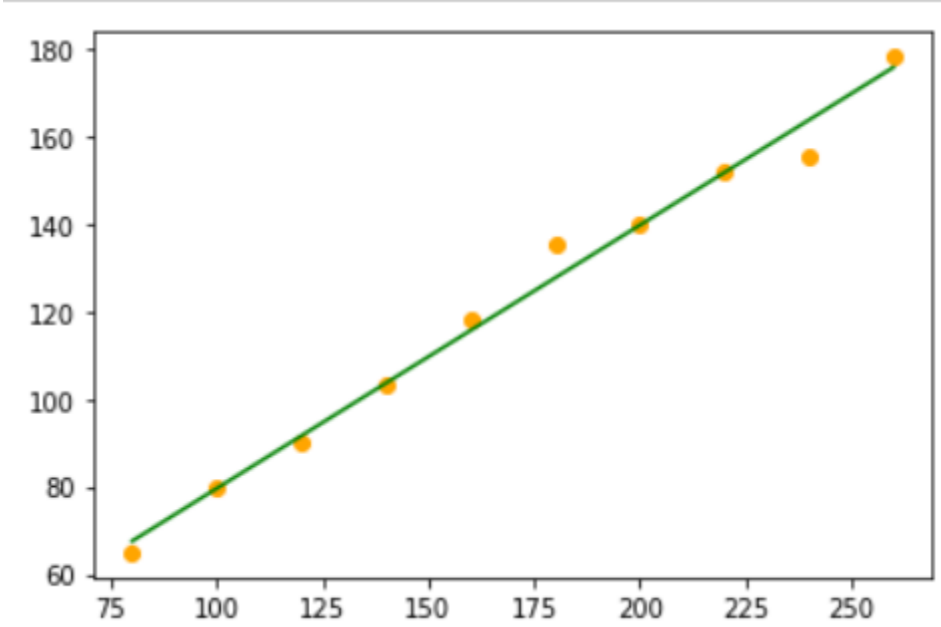
- **Code for Regression Line of Sample - 1:**

```
In [11]: # Sample - 1
from matplotlib import pyplot as plt
from sklearn.linear_model import LinearRegression

x = [[80], [100], [120], [140], [160],
     [180], [200], [220], [240], [260]]
y1 = [65, 80, 90, 103, 118,
      135, 140, 152, 155, 178]


re = LinearRegression().fit(x, y1)
y_pred = re.predict(x)
plt.scatter(x, y1, color = "orange")
plt.plot(x, y_pred, color = "green")
plt.show()
```

- **Regression Line of Sample - 1:**



Sample - 2:

Income (X)	Weekly Family Consumption Expenditure (Y)
80	70
100	74
120	94
140	115
160	125
180	110
200	136
220	160
240	175
260	191

- **Regression Code for Sample - 2**

In [3]:
```
# Sample - 2
import pandas as pd
import numpy as np
from sklearn import datasets, linear_model
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
from scipy import stats


income=np.array([80,100,120,140,160,180,200,220,240,260])
weekly_consumption=np.array([70,74,94,115,125,110,136,160,175,191])
temp = sm.add_constant(income)
est = sm.OLS(weekly_consumption, temp)
est2 = est.fit()
print(est2.summary())
```

- **Regression Result for Sample - 2**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.948
Model:                            OLS   Adj. R-squared:                  0.941
Method:                 Least Squares   F-statistic:                     144.7
Date:                Sat, 07 Nov 2020   Prob (F-statistic):           2.11e-06
Time:                        12:25:57   Log-Likelihood:                -36.049
No. Observations:                  10   AIC:                             76.10
Df Residuals:                       8   BIC:                             76.70
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         13.0061      9.828      1.323      0.222      -9.658      35.670
x1             0.6588      0.055     12.028      0.000       0.532       0.785
==============================================================================
Omnibus:                        7.358   Durbin-Watson:                   1.627
Prob(Omnibus):                  0.025   Jarque-Bera (JB):                3.082
Skew:                          -1.302   Prob(JB):                        0.214
Kurtosis:                       3.783   Cond. No.                         561.
==============================================================================
```

Regression Analysis of Sample - 2:

1. **Sign of Beta Coefficient:**
    - Here, the value of the beta coefficient is **+0.6588** and the value of the constant is **+13.0061**.
    - Hence, in our sample, as the income increases, consumption increases by **65.88%** (considering only beta1).

2. **Value of Beta Coefficient:**
    - In our sample, beta coefficient, beta1 has a value of **0.6588** with t-value **12.028**, and beta0 has a value of **13.0061**, with a t-value of **1.323**. Hence, by the t values we can determine beta1 is highly significant, while beta0 is less significant.

3. **P-value:**
    - In our sample, p value of B1 is equal to **0.0** which is less than 0.05, hence, it is a highly significant variable and so it significantly explains the change in the variable y, which in our case is the weekly consumption.

4. **R Squared value:**
    - The thumb rule is, a R squared value greater than **0.60** suggests our model is well fitted on sample.
    - In our sample, the R squared value was found to be **0.948**, suggesting high coupling between the dependent and independent variables.

5. **F-Statistic Probability value:**
    - In our sample, F-statistic Probability value was found to be **2.11e-06** which is less than 0.05 suggesting that our model is not only valid for this sample it is valid for the entire population.

6. **Skewness:**
    - In our sample, the value of skewness is **-1.302**, which is less than **-1**, hence, suggesting that our sample is **highly skewed**.

7. **Kurtosis:**
    - In our sample, we found the value of kurtosis to be **+3.783**, which suggests that the distribution is peaked, showing that the curve is **leptokurtic**.

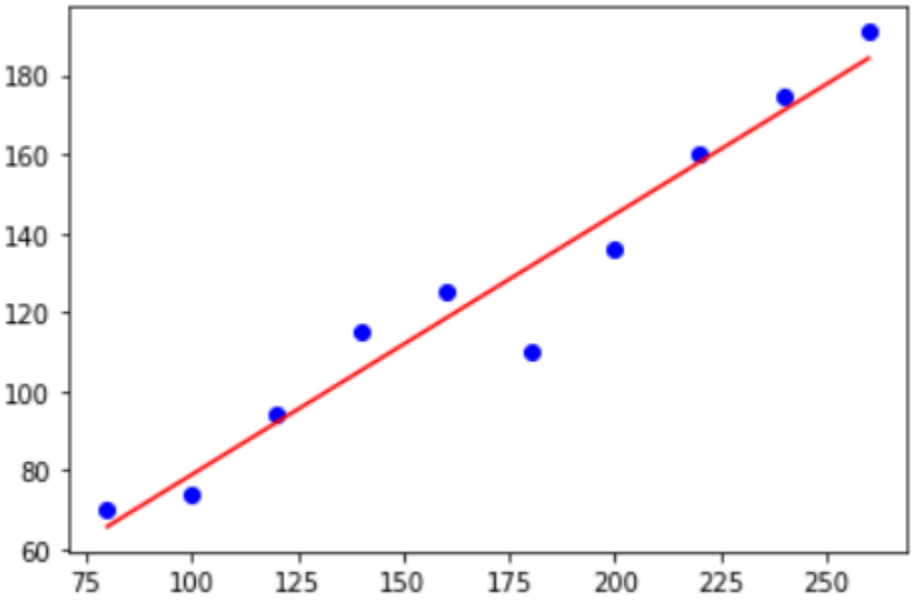
- **Code for Regression Line of Sample - 2:**

[12]:
```
# Sample - 2
from matplotlib import pyplot as plt
from sklearn.linear_model import LinearRegression

x = [[80], [100], [120], [140], [160],
     [180], [200], [220], [240], [260]]
y2 = [70, 74, 94, 115, 125,
      110, 136, 160, 175, 191]

re = LinearRegression().fit(x, y2)
y_pred = re.predict(x)
plt.scatter(x, y2, color = "blue")
plt.plot(x, y_pred, color = "red")
plt.show()
```

- **Regression Line of Sample - 2:**



- **Conclusion:**

  Here, R squared value is **0.987** greater than 0.6 and F-statistic probability value is **6.71e-09** less than 0.5. Based on this analysis, it is concluded that our model weekly consumption, Y = 19.6 + 0.6 * Income is appropriate for Sample Regression as well as Population Regression.