SLIIT Discover Your Future

Lab Sheet 05

Regression Analysis

Activity 1: Simple Linear Regression

Open Jupyter Notebook and create a folder named lab 06. Download the data sets for lab 06 from the courseweb.

Explore the content of the salary_data.csv file. The file contains years of experience of employees and salary in a particular domain.

Now create a new Notebook under Jupyter Notebook. Similar to the previous lab sheets you need to import the pandas, numpy and matplotlib.pyplot library first. Additionally add the following statement to add the statmodels library which we want to use for regression.

```
import statsmodels.api as sm
from sklearn.linear_model import LinearRegression
```

Import the data from the salary_data.csv file and store it in a data frame. View the content of the data using methods known to you.

Create a new variables named x and y and store the content of independent variable in x and dependent variable in y. Use the scatter function in the matplotlib library to plot the data.

Fitting the regression model

Next we have to fit our data to the model and see how well the data fits the model. This could be done using the following statements.

```
model = LinearRegression()

model.fit(x, y)
```

We can see the resultant regression coefficients using the following commands

```
print(model.intercept_)

print(model.coef_)
```

We can also use our model for prediction. For example following code segment will show you the salary for an employee with 5 years of experience.

```
regressor.predict(np.array([[5]]))
```

Using the above output write the regression equation for the dataset.

Obtaining model summary

Although we can write a regression equation referring to the outputs we obtained earlier, for checking the significance of the regression coefficients, checking adequacy of the model and comparing models the above output is not sufficient. We will use a different approach now to obtain a summary of the results.

Simple linear regression equation takes the following format

$$y = \beta 0 + \beta 1x + \epsilon$$

We already have our independent and dependent variables. Now, we need to indicate that we are going to have a constant. Add the following statement to the program.

```
x1 = sm.add_constant(x)
```

The above statement will add a column of one's to the data series x. Now use the following statements in the program. They will estimate the regression coefficients for the data using the Ordinary Least Square (OLS) method.

```
model = sm.OLS(y,x1).fit()

model.summary()
```

Explore the resultant model summary.

How would you interpret the results obtained? What is the regression equation for the model? Is it same as what you obtained earlier?

Activity 2: Multiple linear regression

Create a new notebook in python. Import the libraries you have used in the previous activity. Import the Startup.csv file which is available in the courseweb to the program.

Explore the content of the file using methods known to you. Note that the 'state' column is in text rather than numerical. Convert the content to integers using the method you learned in the previous week.

Follow the instructions below to identify the optimal regression equation.

1. Use simple linear regression between each independent variables and the dependent variable. Obtain the model summary.
2. Identify which independent variables have considerable explanatory power over the model using R^2
3. Next, select the variables which have a considerable explanatory power and add a combination of two of those variables and run multiple linear regression.
4. Does adding variables to the model improves the Adjusted R^2?
5. Determine the most suitable regression equation for the model.

Make a prediction based on the model developed.