Applicability of Linear Regression Model

Linear Regression

Linear Regression is a common technique for analyzing statistical data (quantitative data on any subject such as stock price variations etc.). This model can be considered as a supervised learning method because it should be used known data to train the model. Then this trained model will use to predict unknown or future data (which is called 'label'). For example, the data analyzer can refer past stock data to predict next month stock prices.

Basically, there are two types of linear regression methods such that simple linear regression and multiple linear regression. In simple linear regression method, a single independent variable which is technically referred to as 'feature' is used to predict the dependent variable. Usually, this model fits with the general equation of a line which is y = mx + b. Therefore, for any particular data set, it should be found the suitable values for m (gradient) and b (bias) which has a minimum error for the prediction. Since this model can represent using a line, this can be illustrated using the two-dimensional plot. On the other hand, the multiple linear regression method uses more than one independent variables as features to predict a dependent variable. Then this cannot be simply represented using a two-dimensional work area. Therefore, it may be needed a three-dimensional view to illustrate this model.

Before creating the model using currently available data, it must be followed data preprocessing steps to reduce the error between predicted and real value. Therefore, it should be handled noisy and missing data to build a clean model. After that, it can be selected suitable independent variable/s or feature/s which have a linear relationship (correlation) with the dependent variable. Otherwise, it may predict the future values for label

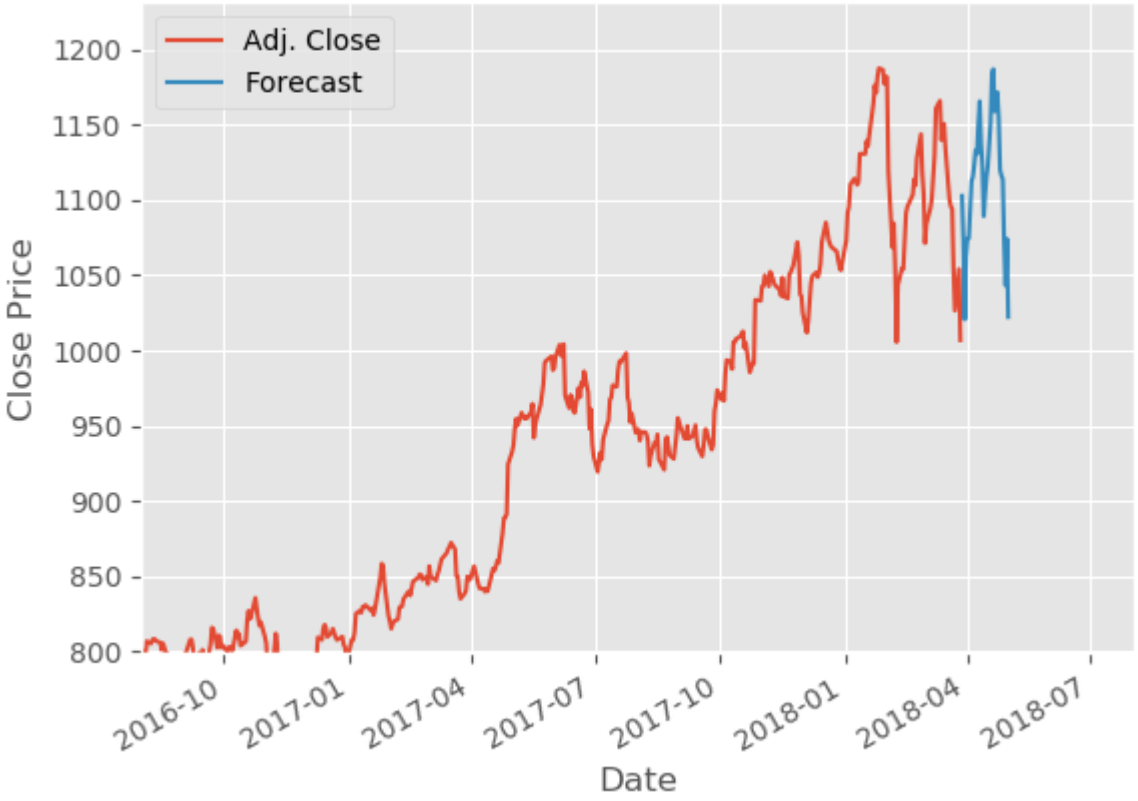


Figure 1

incorrectly. Basically, if using the correct feature/s to model the current data, the predicted data will also follow the similar behavior/pattern as previous data. If not, it may predict unexpected values which will deviate from the current data behavior. For example, in Figure 1 illustrates the future prediction (next month) for 'Close prices' in stock data. Here the multiple linear regression model was used to build the model with several important features. But if the model is trained after dropping important feature/s, it has predicted unexpected values for 'Close price' data as illustrates in Figure 2.

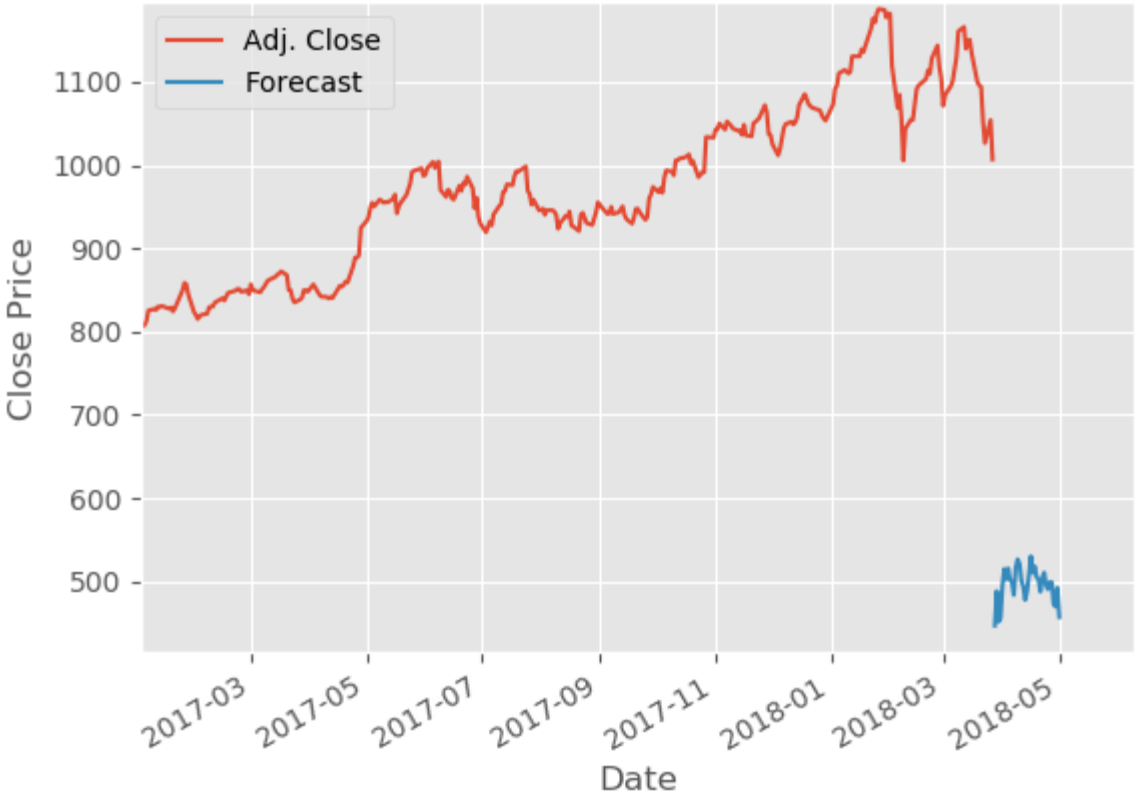


Figure 2

This model is considered as a fundamental supervised learning technique in Machine learning and Artificial Intelligence fields. Practically, this model is used for finding the relationships between variables. Therefore, this model applies in biology, weather, finance and economy fields to forecast the future. Since Python programming language has implemented the most of these models, the user can directly use the linear regression model to predict data. In Python, '**Pandas**' and '**Numpy**' libraries can be used to preprocess the current data. As well as the '**Sklearn**' library can be applied to train the current data and for the linear regression model.

R – squared value for linear regression

R – squared value (R^2) can be used to evaluate 'goodness of fit' for the linear model. Therefore, this value represents a statistical measure that gives how close the data are fitted with the regression line. Statistically, this is also known as the '**coefficient of determination**' for simple linear regression and '**coefficient of multiple determination**' for multiple linear regression. This R^2 value is always between the 0 and 1. If the value of the $R^2 = 1$ indicates, all data points lie on the regression line. This means that there is no difference between current values for the dependent variable and predicted values for the dependent variable using the linear model. If $R^2 = 0$ indicates, data points are not fitted with the regression line. This basically happened if, there is no correlation between the independent variable and dependent variable. When considering about

random data values without any correlation, the value of the R^2 is almost zero. R – squared value can be calculated using the following equation.

R-Squared Error = 1 – (Squared Error for $y_{predicted}$ values / Squared Error for y_{mean} values)

= 1 – [SE(y_p) / SE(y_m)]

According to the above equation, for the best case (R^2 = 1) **SE(y_p)** going to be zero and for the worst case, **SE(y_p)** is equal to the **SE(y_m)**. Note that here **y_{mean} values** denote a horizontal line (**mean line: $y = y_{mean}$**). As well as **$y_{predicted}$ values** represent the regression line. Therefore, theoretically, it should be **SE(y_p) ≤ SE(y_m)**. But there are some rare situations for **SE(y_p) > SE(y_m)** practically. Then the R-Squared Error will take a negative value for this situation. This means that prediction using the mean line is better than making the prediction using the regression line. Actually, obtain a negative value for the R^2, is not a mathematical violation. But it simply says that the chosen model fits the current data really bad way. Figure 3 illustrates such a situation which gives R^2 value of **-0.0108**. Therefore, the mean line is the best fit line than the regression line.

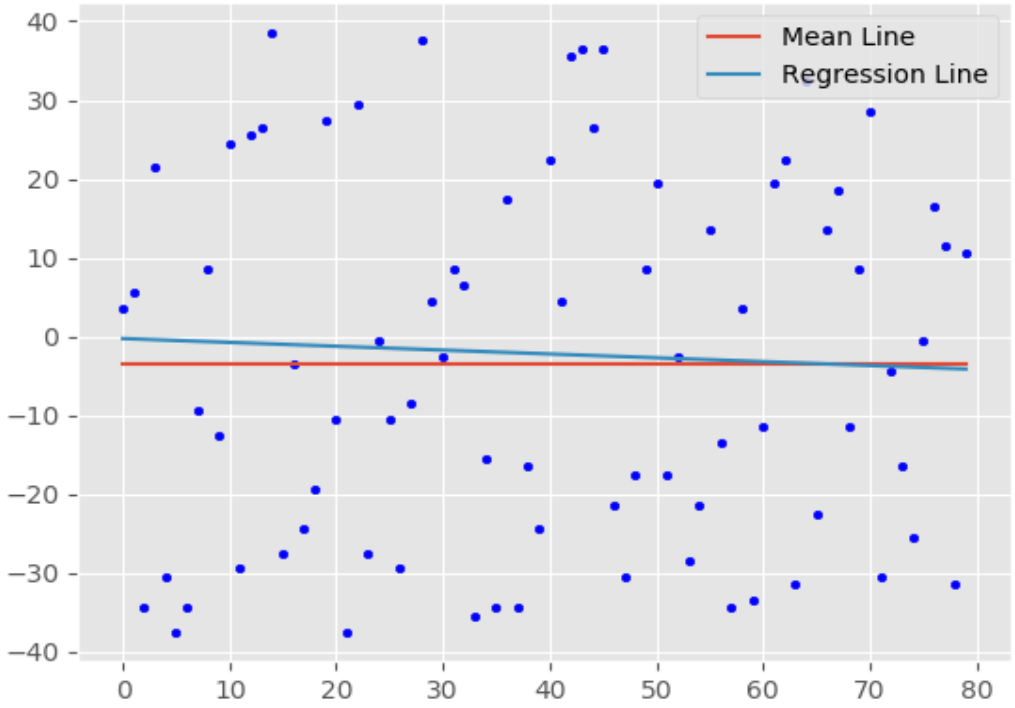


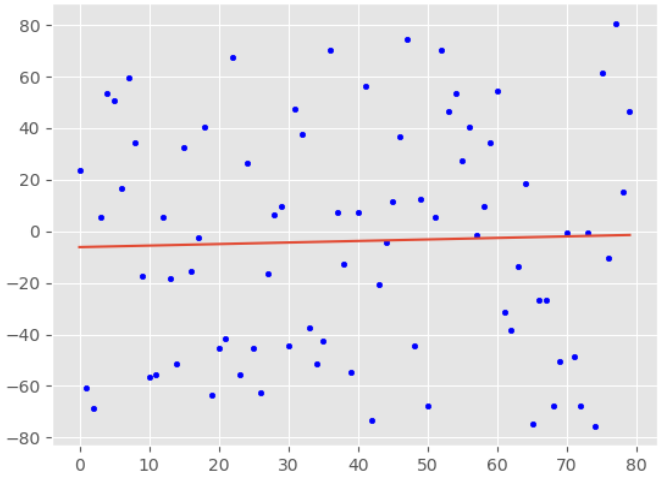
Figure 3

Figure 3Following Table 1 illustrates the R^2 value variations according to the correlation of the independent and dependent variable. Then it can get an idea about the applicability of linear regression model by using the R^2 value. According to the following Figure 4, the simple linear regression model is applicable for the data set that refer in 4.c and 4.d

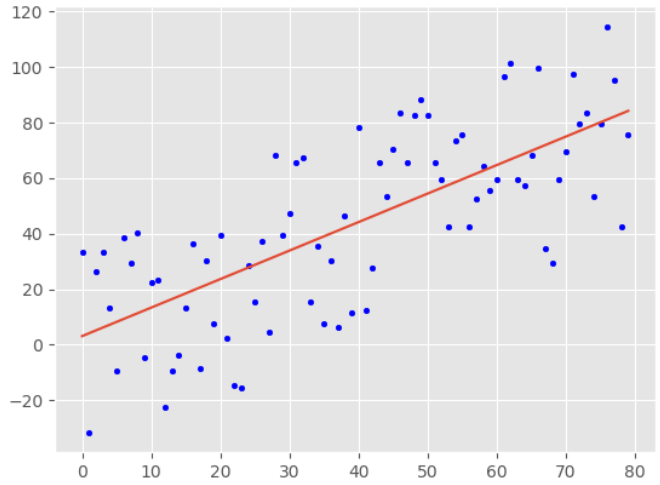
Table 1

Figure number	Figure 4.a	Figure 4.b	Figure 4.c	Figure 4.d
R^2 value	0.0010	0.5136	0.8757	0.9855

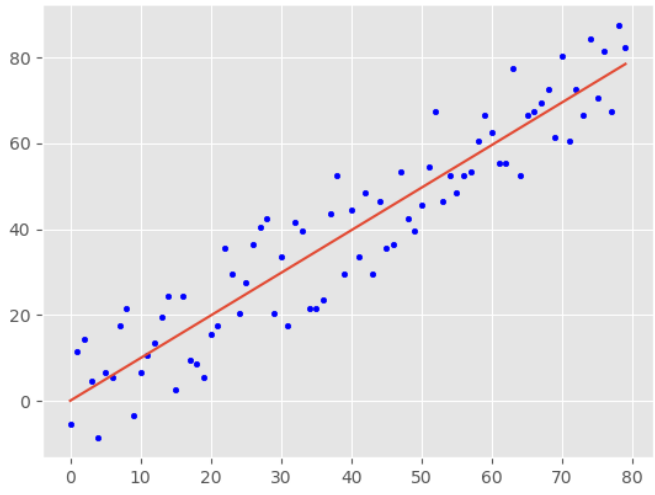
.



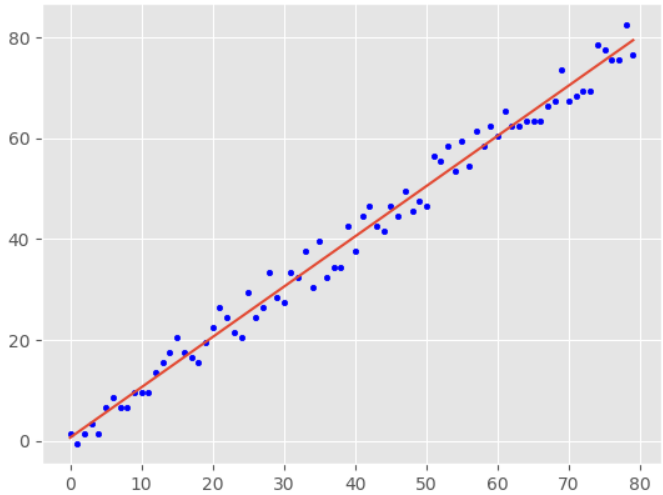
4.a



4.b



4.c



4.d

Figure 4