DS203 : Exercise 1

Nirav Bhattad

Last updated August 12, 2024

§1. Part-A (Simple Linear Regression Derivation)

1.1. Introduction

We will derive the formula for simple linear regression. We will use the following notation:

- X is the dependent variable.
- Y is the independent variable.
- n is the number of data points.
- (x_i, y_i) is the i^{th} data point.
- $\overline{x}$ is the mean of X.
- $\overline{y}$ is the mean of Y.
- $\overline{xy}$ is the mean of XY.
- $\overline{x^2}$ is the mean of X^2.
- $\overline{y^2}$ is the mean of Y^2.

1.2. Derivation

The formula for the line of best fit is given by:

$$y = ax + b \tag{1.1}$$

where a is the slope and b is the intercept. We can derive the formula for a and b by minimizing the sum of squared errors. The sum of squared errors is given by:

$$\sum_{i=1}^{n}(y_i - (ax_i + b))^2 \tag{1.2}$$

We can minimize this expression by taking the partial derivatives with respect to a and b and setting them to zero. The partial derivative with respect to a is given by:

$$\frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - (ax_i + b))^2 \tag{1.3}$$

The partial derivative with respect to b is given by:

$$\frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - (ax_i + b))^2 \tag{1.4}$$

Setting these partial derivatives to zero gives us the following two equations:

$$\sum_{i=1}^{n}(y_i - (ax_i + b))x_i = 0 \tag{1.5}$$

$$\sum_{i=1}^{n}(y_i - (ax_i + b)) = 0 \tag{1.6}$$

Solving these equations gives us the following formulas for a and b:

$$a = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} \tag{1.7}$$

$$b = \frac{\overline{y}\,\overline{x^2} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2} \tag{1.8}$$

1.3. Conclusion

We have derived the formula for simple linear regression. The equations for slope and intercept are given by:

$$a = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} \tag{1.9}$$

$$b = \frac{\overline{y}\,\overline{x^2} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2} \tag{1.10}$$

§2. Part-B

Step 1

Using a spreadsheet create a dataset comprising 100 pairs (x_i, y_i) as per the following guidelines:

- Create 100 random values of x_i lying between 0 and 1 (both inclusive)
- Corresponding to each x_i, create y_i such that the scatter plot of y_i *v/s* x_i looks like the following:

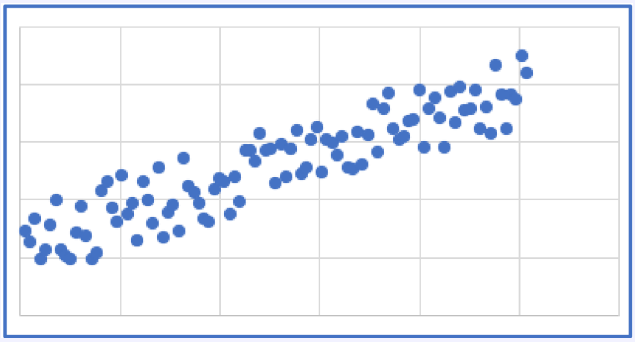

Figure 1: Randomized Dataset

Here, I choose to create a dataset with 100 random values of x_i between 0 and 1 (both inclusive). The corresponding y_i values were calculated using the following formula:

$$y_i = 2 \times x_i + 1 + (\texttt{RAND()} - 0.5) \times 0.4 \tag{2.1}$$

Step 2

Create the scatter plot resulting from your dataset (x_i, y_i)

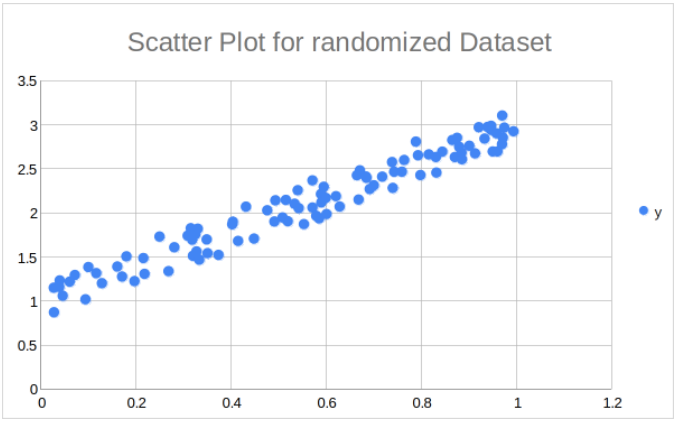

Figure 2: Scatter Plot for our Dataset

Step 3

Using the (x_i, y_i) data, calculate the regression coefficients a and b (all calculations should be entirely done using the spreadsheet). The equation of the resulting regression model (line) will be as shown below.

$$\widehat{y_i} = a \cdot x_i + b \tag{2.2}$$

Parameters obtained from Linear Regression	
a	1.98148553913676
b	1.01359291754755

Figure 3: Regression Coefficients

Step 4 and Step 5

Using this regression line predict $\widehat{y_i}$ corresponding to every x_i. Superimpose the regression line over the scatter plot created in Step 2.3, as shown below.

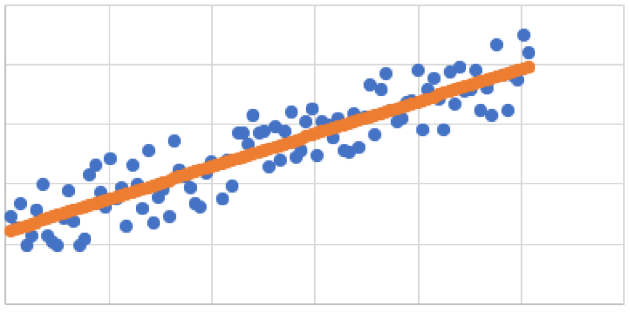

Figure 4: Regression Line Superimposed on the Scatter Plot

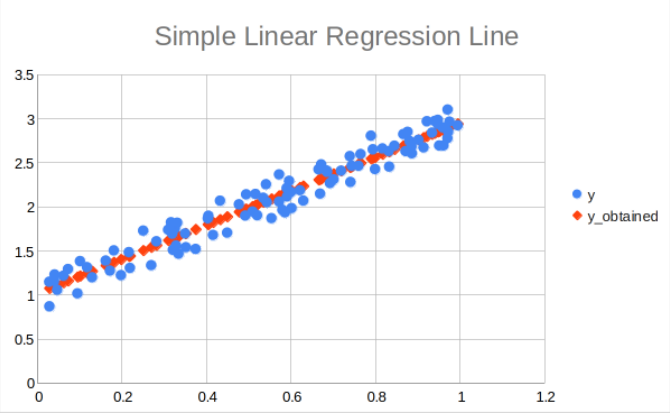

Figure 5: Regression Line Superimposed on the Scatter Plot for my dataset

Step 6

Calculate the prediction error e_i corresponding to every y_i, and calculate the error metrics Sum of Squared Errors (SSE) and Mean Absolute Error (MAE).

SSE (Sum of squared errors)	1.8863572051536
MAE (Mean Absolute Error)	11.8525873814639

Figure 6: Error Metrics

Step 7

Create a scatter plot of e_i v/s x_i.

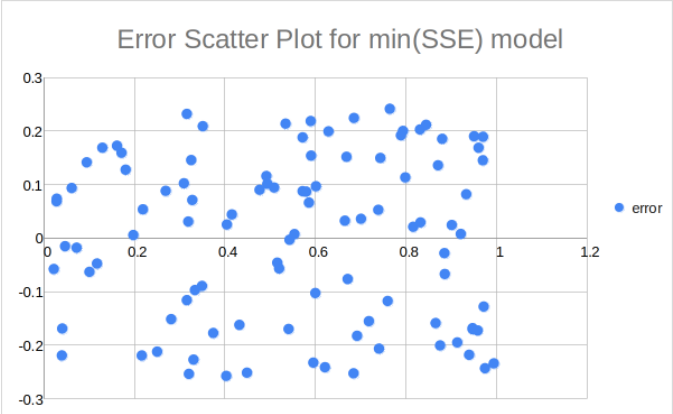

Figure 7: Scatter Plot of e_i v/s x_i

Step 8

As discussed in class, the simplest (and naïve) model is one that predicts the mean, as shown below. Using this model calculate e_i, SSE and MAE and create the scatter plot of e_i v/s x_i.

$$\widehat{y_i} = \overline{y} \tag{2.3}$$

SSE (Sum of squared errors)	36.9657797718942
MAE (Mean Absolute Error)	51.7021117917332

Figure 8: Scatter Plot of e_i v/s x_i for Naïve Model

Figure 9: Error Metrics for Naïve Model

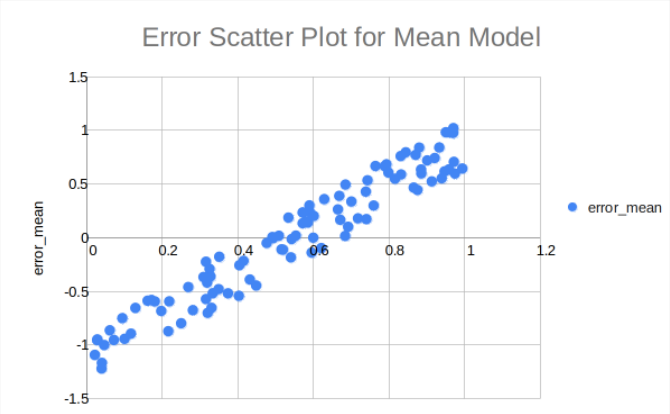

Figure 10: Scatter Plot of e_i v/s x_i for Naïve Model

Step 9

Compare the error metrics and error scatter plots resulting from the above two distinct models and record your analysis and explain the differences between the two error scatter plots. (Note: Stating obvious facts is NOT analysis!)

2.9.1. Error Metrics

1. **Minimum SSE Model:** This model is optimized to minimize the Sum of Squared Errors (SSE). Hence in general we observe a lower SSE and Mean Absolute Error (MAE) value as compared to the Naïve Model, because the model's parameters are specifically adjusted to fit the training data as closely as possible.

2. **Mean Model:** This model is a simple model that predicts the mean value of the target variable for all inputs. This model is not optimized for any specific metric, and hence we observe a higher SSE and MAE value as compared to the Minimum SSE Model.

2.9.2. Error Scatter Plots

1. **Minimum SSE Model:** The error scatter plot for the Minimum SSE Model shows a random distribution of errors around the $y = 0$ line. This indicates that the model is able to predict the target variable with a certain degree of accuracy, and the errors are distributed randomly around the regression line with almost equal probability to be positive or negative.

2. **Mean Model:** The error scatter plot for the Mean Model shows a systematic distribution of errors around the $x = 0.5$ line. This indicates that the model is not able to predict the target variable accurately, and the errors are systematically biased towards one side of the regression line till x is around 0.5 and then flips signs. In this case, the errors are consistently negative for $x < 0.5$ and consistently positive for $x > 0.5$. The error scatter plot looks almost like a straight line with a slope equal to the slope of the regression line.

2.9.3. Analysis

The analysis of the error metrics and error scatter plots for the two models is as follows:

1. The Minimum SSE Model is able to predict the target variable more accurately as compared to the Mean Model, as indicated by the lower SSE and MAE values.
2. The error scatter plot for the Minimum SSE Model shows a random distribution of errors around the regression line, indicating that the model is able to predict the target variable with a certain degree of accuracy.
3. The error scatter plot for the Mean Model shows a systematic distribution of errors around the $x = 0.5$ line, indicating that the model is not able to predict the target variable accurately and the errors are systematically biased towards one side of the regression line.
4. The Mean Model is a simple model that predicts the mean value of the target variable for all inputs. This model is not optimized for any specific metric, and hence the errors are systematically biased towards one side of the regression line.
5. The Minimum SSE Model is optimized to minimize the Sum of Squared Errors (SSE) and hence is able to predict the target variable more accurately as compared to the Mean Model.
6. The error scatter plot for the Minimum SSE Model shows a random distribution of errors around the regression line, indicating that the model is able to predict the target variable with a certain degree of accuracy.

The main distinction between the error scatter plots of the two models is that the minimum sum of squared errors (SSE) model typically shows a better fit, with residuals spread out randomly around zero, indicating a well-fitted model. In contrast, the residuals of the mean model are more likely to exhibit identifiable patterns, highlighting its failure to accurately capture the nuances of the data. The minimum SSE model's ability to more closely match the data generally leads to significantly improved error metrics, underscoring the effectiveness of model-specific optimization compared to simple average predictions. Finally, we can conclude that the minimum SSE model is more effective in predicting the target variable, as evidenced by its lower error metrics and more evenly distributed residuals.