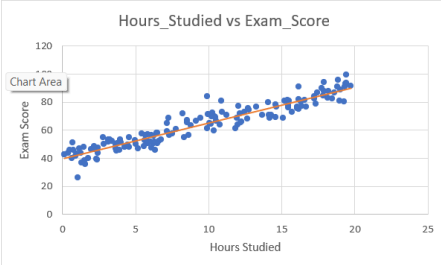
DS 203: Programming for Data Science - Exercise 1

Swayam Saroj Patel (22B1816)

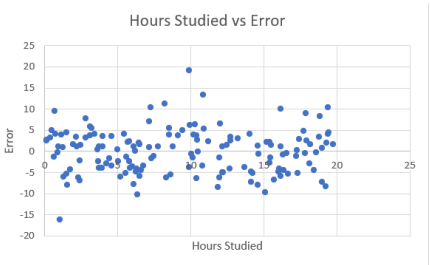
January 28, 2025

With Intercept:

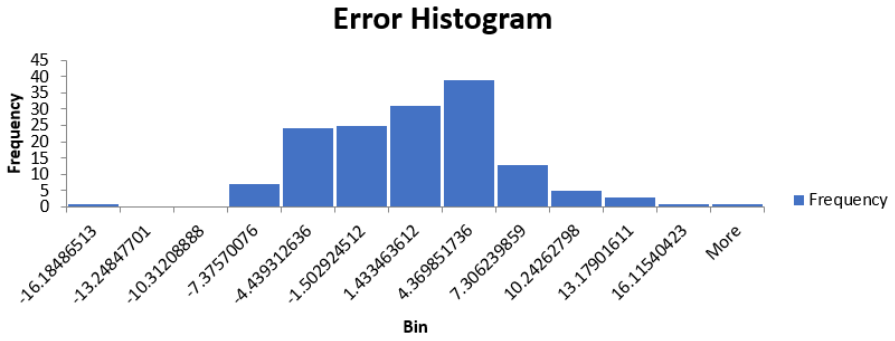
The plots:



(a) Scattered plot of hours vs score



(b) scattered plot of hours vs error



(c) Histogram if the error

Values:

Parameter	Value
Σx	1418.62
$\overline{x}$	9.45747
Σy	9581
$\overline{y}$	63.87333
Σxy	103942.5
Σx^2	18656.79
$a(slope)$	2.54388
$b(intercept)$	39.81467

Error Metrics: (Answer of Point 6)

Sum of Squared Errors (SSE)

Formula:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- It is used to evaluate the overall fit of the model.
- Lower SSE indicates a better fit, as it implies that the model's predictions are closer to the observed values.
- Total error measure, useful for comparing models on the same dataset.

Mean Squared Error (MSE)

Formula:

$$\text{MSE} = \frac{\text{SSE}}{n}$$

- It is widely used to assess the accuracy of regression models.
- Reflects the average squared error, penalizing large deviations.

Root Mean Squared Error (RMSE)

Formula:

$$\text{RMSE} = \sqrt{\text{MSE}}$$

- RMSE provides an interpretable error metric as it is in the same unit as the dependent variable y.
- It is used when it is important to understand the typical magnitude of errors in the same scale as the data.

- A smaller RMSE indicates better predictive performance.
- Offers a scale-dependent interpretation, helpful in understanding error magnitude.

Mean Absolute Error (MAE)
Formula:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Context:

- It is less sensitive to outliers compared to MSE and RMSE.
- MAE is preferred when the goal is to minimize the average absolute error in predictions.
- Represents average error magnitude, less sensitive to extreme values.

Values in this model:

Parameter	**Value**
SSE	3960.05
MSE	26.4003
RMSE	5.13813
MAE	10.3340
Skew	0.26146
Kurtosis	0.87142
SST	37871.34
R^2	0.895434

Skewness and Kurtosis Analysis

A normal distribution has Skewness 0 and Kurtosis near 3. **(Answer of point 9)**
For the error distribution ($e_i = y_i - \hat{y}_i$) of the regression model:

- **Skewness:** 0.261
  The skewness is close to zero, indicating that the error distribution is nearly symmetric.
- **Kurtosis:** 0.871
  The kurtosis value is less than 3, suggesting that the error distribution is slightly flatter than a normal distribution (platykurtic).

R-Squared and Goodness of Fit

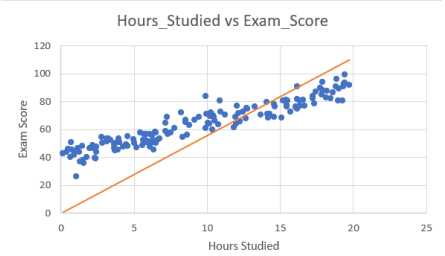
R^2 is a value that typically ranges between 0 to 1, and it tell us about how good your model is. Closer to 1, better the model. **(Answer to point 10)** Formula:

$$R^2 = 1 - \frac{SSE}{SST}$$

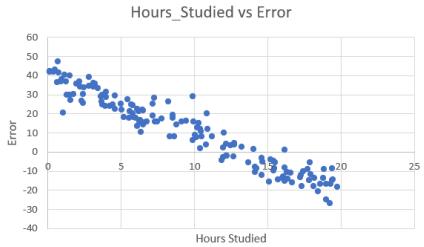
- **R-Squared:** 0.8954
  The high R^2 value demonstrates that the regression model effectively captures the relationship between the independent variable (x_i) and the dependent variable (y_i). This suggests that the model is highly reliable for prediction.

Without Intercept:

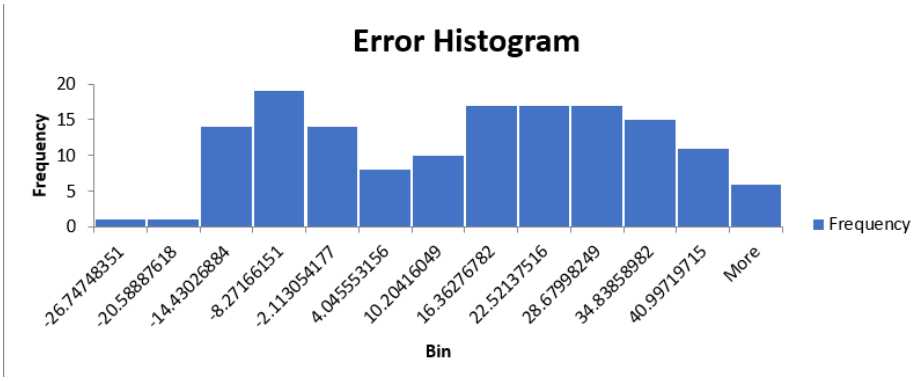
The plots:



(a) Scattered plot of hours vs score



(b) scattered plot of hours vs error



(c) Histogram if the error

Values:

Parameter	Value
Σx	1418.62
$\overline{x}$	9.45747
Σy	9581
$\overline{y}$	63.87333
Σxy	103942.5
Σx^2	18656.79
$a(slope)$	5.571298

Error Metrics:

Parameter	Value
SSE	70746.98
MSE	471.6466
RMSE	21.71743
MAE	8.583178
Skew	-0.10394
Kurtosis	-1.1626
SST	37871.34
R^2	-0.86809

Skewness and Kurtosis Analysis

A normal distribution has Skewness 0 and Kurtosis near 3.
For the error distribution ($e_i = y_i - \hat{y}_i$) of the regression model:

- **Skewness:** -0.10394
  The skewness is close to zero, indicating that the error distribution is approximately symmetric.
- **Kurtosis:** -1.1626
  The kurtosis value is significantly less than 3, suggesting that the error distribution is very flat (platykurtic) with light tails compared to a normal distribution.

R-Squared and Goodness of Fit

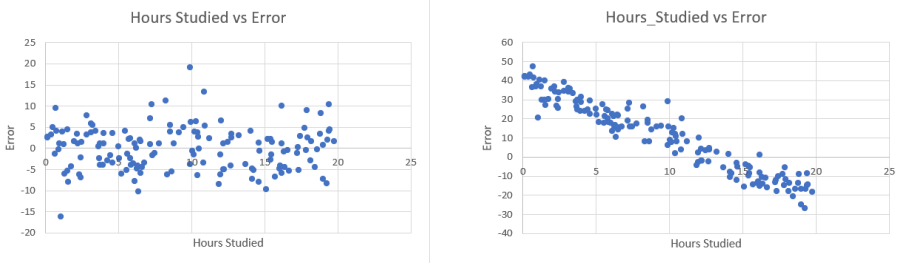
Formula:

$$R^2 = 1 - \frac{SSE}{SST}$$

- **R-Squared:** -0.86809
  The R^2 value is negative, which indicates that the model performs worse than a simple horizontal line at the mean of y_i. This suggests that the regression model fails to explain the variance in the data and is not a good fit.

Comparison:

The plots:



(a) With intercept (b) Without intercept

The error scatter plot with intercept is kind of **random and uniformly distributed** in the range of -16 to 20. While on the other hand the one with no intercept is **like a linear plot**.
I believe this is due to the fact that most of the data in the extremities are being treated like outliers. Resulting in higher error values. And the linear behavior due to the fact that the plot we are referring also seems linear so at one point it has to intersect as they are not parallel. So that is why this scatter plot also crosses 0.

Error Metrics:

- **Sum of Squared Errors (SSE):** The model with intercept has a lower SSE (3960.05) compared to the model without intercept (70746.98). This indicates that the model with intercept better minimizes the overall error.
- **Mean Squared Error (MSE):** MSE (26.40) is lower for the model with intercept, suggesting that it provides more accurate predictions. The no-intercept model has large prediction errors, as reflected by its MSE of 471.6466.

- **Root Mean Squared Error (RMSE):** The RMSE too give a similar deduction with values 5.13813(with intercept) and 21.71743 (without intercept).
- **Mean Absolute Error (MAE):** The MAE is slightly lower for the model without intercept (8.58 compared to 10.33).