ASSIGNMENT - 1

MATH_564-Regression

(1) (6 points) Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n, the number of data; and p, the number of predictors/features in each of the scenario.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

The type of problem here is Regression, Our goal in this scenario is to comprehend how different variables, or predictors, affect the continuous outcome, in this case the CEO compensation. Regression analysis is required since the CEO compensation is a numerical value rather than a category.

The interest here is Inference, Finding out how CEO pay relates to other variables like industry, number of employees, and profit is the aim of this research. An essential component of inference is examining the impact of changes in these variables on CEO remuneration.

Number of Data Points (n): 500, Data is collected from 500 firms, so each firm represents a single data point.

Number of Predictors/Features (p): 3, the predictors are profit, number of employees, and industry. Keep in mind that if it's categorical, industry might need encoding; yet, in this case, it counts as one characteristic.

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Type of Problem here is Classification, Based on past performance, the objective in this case is to predict if a new product would succeed or fail. Due to the fact that the product must be categorized into either success or failure, the result is categorical. As such, this is a problem of classification.

The Interest here is Prediction, Predicting the new product's outcome is the main goal. Based on the characteristics of comparable items that have been introduced in the past, you are

A20554334
AKHIL KUMAR MARNI

Due September 6th 2024, Friday, 11:59pm

interested in predicting whether the new product will succeed or fail. This entails developing a prediction model that, using the recorded features, can determine the probability of success or failure.

Number of Data Points (n): 20, the data set includes information on 20 previously launched products. Each product serves as one data point in the analysis.

Number of Predictors/Features (p): 13, the predictors includes, Whether the product was a success or failure. Price charged for the product, Marketing budget, Competition price and Ten other variables related to product attributes or market conditions

Therefore, the total number of predictors or features used to make predictions is 13.

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

The type of problem here is Regression, In this case, estimating the percentage change in the continuous variable USD/Euro exchange rate is the objective. Predictors that are continuous include the percentage change in different stock markets. Regression analysis applies here since we are predicting a numerical result.

The Interest here is Prediction, The primary goal is to use previous data to forecast future changes in the USD/EUR exchange rate. By simulating the correlation between the exchange rate and weekly fluctuations in various stock markets, you hope to predict the evolution of the exchange rate. This task involves prediction.

Number of Data Points (n): 52, the data set includes weekly observations for the entire year of 2012. Since there are 52 weeks in a year, there are 52 data points in total, with each week representing one data point.

Number of Predictors/Features (p): 3, the predictors includes, percentage change in the US market, percentage change in the British market, percentage change in the German market.

Therefore, there are 3 predictors or features used to predict the percentage change in the USD/Euro exchange rate.

(2) (6 points) What are the advantages and disadvantages of parametric and non-parametric methods?

A20554334
AKHIL KUMAR MARNI

Parametric Methods

- ✓ **Advantages:**

1. When using parametric models, particularly for small datasets, fewer parameters need to be estimated, which accelerates training convergence and reduces computing load.

2. Explaining the outcomes to non-technical stakeholders is made easier by parametric models' specified structure, which makes them simpler and easier to interpret.

3. For parametric models, a plethora of well-known statistical tools are available, such as confidence intervals, diagnostics, and hypothesis testing, all of which help improve the accuracy of conclusions drawn from the data.

4. In many situations, parametric models' limited structure makes them well-suited for generalizing to new data when assumptions are met, which lowers the risk of over-fitting.

5. In production situations with limited resources, parametric models can be more favorable because they require less memory and compute capacity to generate predictions once trained.

- ✓ **Disadvantages:**

1. A lot of the assumptions made by parametric models about the underlying data distribution, such as linearity and normality are crucial. The accuracy of the model may drastically decrease if these presumptions are broken.

2. Complicated, non-linear interactions are hard to express in parametric models due to their rigid structure, which limits their use in complicated, real-world scenarios.

3. Parametric models have a lower risk of over-fitting than more flexible techniques, but they can still over-fit, especially if the dataset is limited or the assumptions do not fit the data well.

4. Because parametric models rely on specified functional forms, they are vulnerable to bias. This might result in under-fitting if the underlying data structure is more complex than the model predicts.

5. A lot of parametric models, including linear regression, have a high threshold for outliers, which can cause predictions to be distorted and produce unreliable findings.

A20554334
AKHIL KUMAR MARNI

Due September 6th 2024, Friday, 11:59pm

Non-Parametric Methods

- ✓ **Advantages:**

  1. Non-parametric approaches can more successfully describe complicated, non-linear interactions since they do not make significant assumptions about the shape of the underlying data distribution.

  2. Many data types, notably those with outliers or irregular distributions, are better suited for these techniques since they are typically more resilient to assumptions being broken.

  3. When presented with adequate data, non-parametric models can closely conform to the underlying patterns, providing excellent predictive accuracy when parametric models would find it difficult to make predictions.

  4. Non-parametric techniques are more adaptable in practical applications since they do not depend on data to adhere to a certain probability distribution.

  5. Several non-parametric techniques, such random forests and decision trees, can efficiently handle high-dimensional data, spotting significant features and connections without the need for explicit feature engineering.

- ✓ **Disadvantages:**

  1. Since non-parametric models become more complex as more data is used, they frequently need greater processing power and memory, particularly when working with large datasets. It is possible for inference and training to become labor-intensive and slow.

  2. The outcomes of non-parametric models may be more difficult to interpret due to their greater complexity as a result of their flexibility. For example, decision boundaries in techniques such as k-nearest neighbors (KNN) are frequently hard to describe and depict.

  3. Although adaptable, non-parametric techniques have the potential to overfit the data, especially in cases of insufficient data or improper management of the model's complexity.

  4. One disadvantage of non-parametric approaches, particularly in applications needing statistical inference, is that they frequently lack a formal framework for confidence intervals and hypothesis testing, in contrast to parametric models.

A20554334
AKHIL KUMAR MARNI

Due September 6th 2024, Friday, 11:59pm

5. Non-parametric models frequently need a lot of data to function properly. Small datasets could make it difficult for them to identify trends and generate trustworthy forecasts.

6. While parametric methods use parameters to summarize the data, non-parametric approaches like KNN, store the complete dataset in order to generate predictions, using more memory.

5. (6 points) Consider a statistical learning model trained to predict a continuous outcome Y based on features X. The model is trained on a finite training dataset, and we are interested in understanding the tradeoff between variance and bias in the model's predictions.

(a) Explain what is meant by "bias" and "variance" in the bias-variance tradeoff in the context of statistical learning models.

Bias: The term "bias" describes the inaccuracy that results from using a simplistic model to approximate a complicated real-world problem. Prediction errors in a high-bias model are often caused by strong assumptions made about the data. For example, attempting to capture a complicated non-linear connection with a linear regression model will probably result in significant bias since it oversimplifies the underlying relationship between the input features (X) and the output (Y). A model that is too simplistic to fully capture the underlying patterns in the data due to high bias may be said to be under-fitted, producing predictions that are frequently off.

Variance: The term "variance" describes how sensitive the model is to minute modifications in the training set. Strongly adaptable and obsessive about the training set, a high variance model captures both the noise and the underlying patterns. Because of its inability to generalize, this might result in over-fitting, where the model performs well on training data but badly on untested data. Having a high variance means that the model is less stable and dependable in real-world scenarios since its predictions vary greatly across training datasets.

The bias-variance tradeoff: In statistical modeling and machine learning, the bias-variance tradeoff is a key idea. It explains the trade-off between the error resulting from variance (error from highly complicated models) and bias (error from overly simple models). Finding a balance where the total error, the sum of the bias, variance, and irreducible error, is minimized is the aim.

A20554334
AKHIL KUMAR MARNI

(b) Suppose you have the following scenarios:

Scenario 1: A model that is very simple, like a linear regression model with only one feature.

Scenario 2: A more complex model, like a polynomial regression model with many features.

(i) Which scenario is more likely to have high bias and low variance? Explain.
(ii) Which scenario is more likely to have low bias and high variance? Explain

High Bias and Low Variance:

Bias: Because it makes strong assumptions about the nature of the relationship between the input and output (such as that the relationship is linear), a simple model like linear regression with only one feature is likely to have high bias. Because the model is too simple to fully represent the underlying patterns in the data, under-fitting may result from this simplicity. For example, a linear model will consistently miss the target and produce skewed forecasts if the underlying relationship is non-linear.

Variance: The model's simplicity also makes it less susceptible to changes in the training set of data. If there is a small variation in the training data, the model is unlikely to alter dramatically, leading to low variance. As a result, the model is potentially erroneous but stable and consistent across many datasets.

Low Bias, High Variance:

Bias: A more complex model, such as polynomial regression with many features, is capable of capturing intricate patterns in the data, leading to low bias. The model can closely fit the training data, accommodating non-linear relationships and complex interactions between features. This reduces bias because the model does not oversimplify the problem.

Variance: However, the flexibility of a complex model comes with a downside: it is more sensitive to the specific training data. The model can over-fit by capturing noise and irrelevant details in the training data, leading to high variance. This means that while the model may perform well on the training data, it may struggle to generalize to new, unseen data, resulting in poor performance on out-of-sample predictions.

A20554334
AKHIL KUMAR MARNI

3. (6 points) In practice, we usually don't calculate least squares lines manually; instead, we use statistical software like R to handle the computations. However, in this exercise, we will manually compute the least squares line using the formula

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x},$$

to understand the process better. (Note: There are others equivalent way of writing the above two formulas.)
Given the following data points:

x	y
1	10
2	14
4	12
6	13
7	15
8	12
10	13

Calculate the least squares line y = β0 + β1x, without using software such as R, Excel, Python and so on, a typical hand-held calculator is fine. Write your answer in fraction, do not simplified to decimals.

A20554334
AKHIL KUMAR MARNI

Due September 6th 2024, Friday, 11:59pm

Solutions

Given →

x →	1	2	4	6	7	8	10
y →	10	14	12	13	15	12	13

By using the given formula compute following.

Calculate → x $\sum_{i=1}^{n} x_i \Rightarrow 1+2+4+6+7+8+10$

$\Rightarrow 38$

Calculate y $\sum_{i=1}^{n} y_i \Rightarrow 10+14+12+13+15+12+13$

$\Rightarrow 89$

Calculate x thsoughy $\sum_{i=1}^{n} x_i y_i$

$\Rightarrow 1(10)+2(14)+4(12)+6(13)+7(15)+8(12)+10(13)$

$\Rightarrow 10+28+48+78+95+105+96+130$

$\Rightarrow 495$

A20554334
AKHIL KUMAR MARNI

Calculate x^2

$$\sum_{i=1}^{n} x_i^2 \Rightarrow (1)^2 + (2)^2 + (4)^2 + (6)^2 + (7)^2 + (8)^2 + (10)^2$$

$$\Rightarrow 1 + 4 + 16 + 36 + 49 + 64 + 100$$

$$\Rightarrow 270$$

Substitute all the values to compute the formula $\hat{\beta}_1$ & $\hat{\beta}_0$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$\Rightarrow \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

A20554334
AKHIL KUMAR MARNI

Calculate $\hat{\beta}_1$

$$\hat{\beta}_1 \Rightarrow \frac{(7)(495) - (89)(38)}{(7)(270) - (38)^2}$$

$$\Rightarrow \frac{(3465) - (3382)}{(1890) - (1440)}$$

$$\Rightarrow \frac{83}{450} \Rightarrow \frac{83}{446}$$

$$\hat{\beta}_1 \Rightarrow \frac{83}{446}$$

Calculate $\hat{\beta}_0$ $\quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\bar{y} = \sum_{i=1}^{n} y_i / n$$

$$\Rightarrow \frac{89}{7}$$

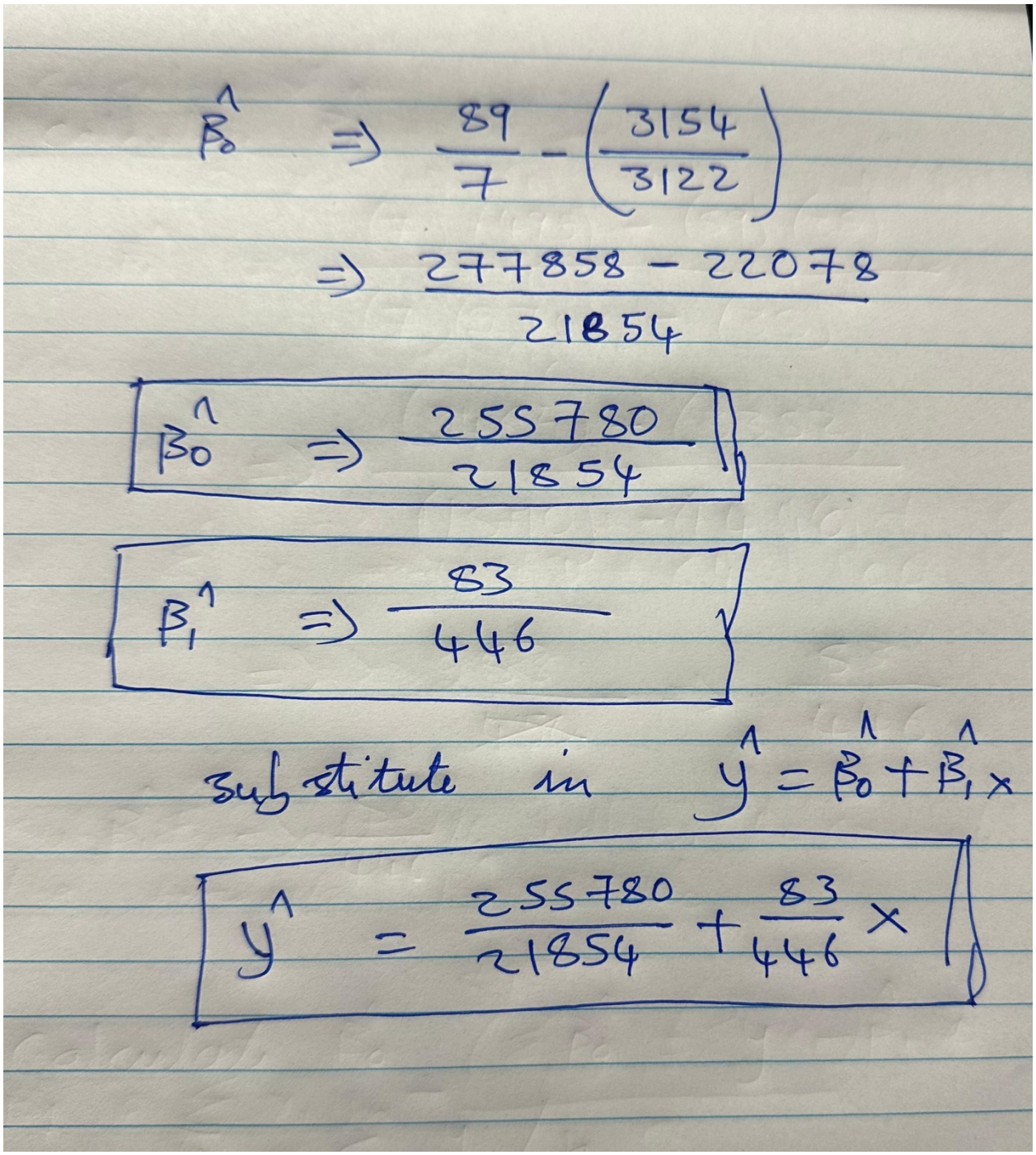
$$\bar{x} \Rightarrow \sum_{i=1}^{n} x_i / n$$

$$\Rightarrow \frac{38}{7}$$

$$\hat{\beta}_0 \Rightarrow \frac{89}{7} - \left[\left(\frac{83}{446}\right) \cdot \left(\frac{38}{7}\right)\right]$$

A20554334
AKHIL KUMAR MARNI

Due September 6th 2024, Friday, 11:59pm



A20554334
AKHIL KUMAR MARNI

Due September 6th 2024, Friday, 11:59pm

4. (6 points) In Problem 3, notice that with by = bβ0 + bβ1x we have the following:

x	y	$\widehat{y}$	$y-\widehat{y}$
1	10	5303/446	-843/446
2	14	2693/223	429/223
4	12	2776/223	-100/223
6	13	2859/223	40/223
7	15	5801/446	889/446
8	12	2942/223	-266/223
10	13	3025/223	-126/223

Take note that the values in the 4th column sum to zero!

This is actually not a coincidence, in fact, show that for the least squares line by = bβ0 + bβ1x obtained from the training data (x1, y1), ・ ・ ・ , (xn, yn), we have

$$\sum_{i=1}^{n}(y_i-\widehat{y}_i)=0.$$

SOL

A20554334
AKHIL KUMAR MARNI

Due September 6th 2024, Friday, 11:59pm

Least square line: $\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$

we know that residuals are defined as

$$\hat{\varepsilon_i} = y_i - \hat{y_i}$$

$$\hat{y_i} \Rightarrow \hat{\beta_0} + \hat{\beta_1} x_i$$

$$\hat{\varepsilon_i} = y_i - \hat{\beta_0} + \hat{\beta_1} x_i$$

we have to prove that
Sum of residuals is zero.

i.e. $\sum_{i=1}^{n} \hat{\varepsilon_i} = 0$

substitute above equations

$$\Rightarrow \sum_{i=1}^{n} \left(\hat{\varepsilon_i}\right) = \sum_{i=1}^{n} \left[y_i - \hat{\beta_0} - \hat{\beta_1} x_i\right]$$

$$\Rightarrow \sum_{i=1}^{n} \left(\hat{\varepsilon_i}\right) = \sum_{i=1}^{n} (y_i) - \sum_{i=1}^{n} \left(\hat{\beta_0}\right) - \sum_{i=1}^{n} \left(\hat{\beta_1} x_i\right)$$

β_0 & β_1 are constants

A20554334
AKHIL KUMAR MARNI

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

By the following

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = S_{xy} / S_x^2$$

$\bar{y}$ is the Mean of the y_i

$\bar{x}$ is the Mean of the x_i

$$\sum_{i=1}^{n} x_i = n\bar{x}$$

$$\sum_{i=1}^{n} y_i = n\bar{y}$$

By substituting

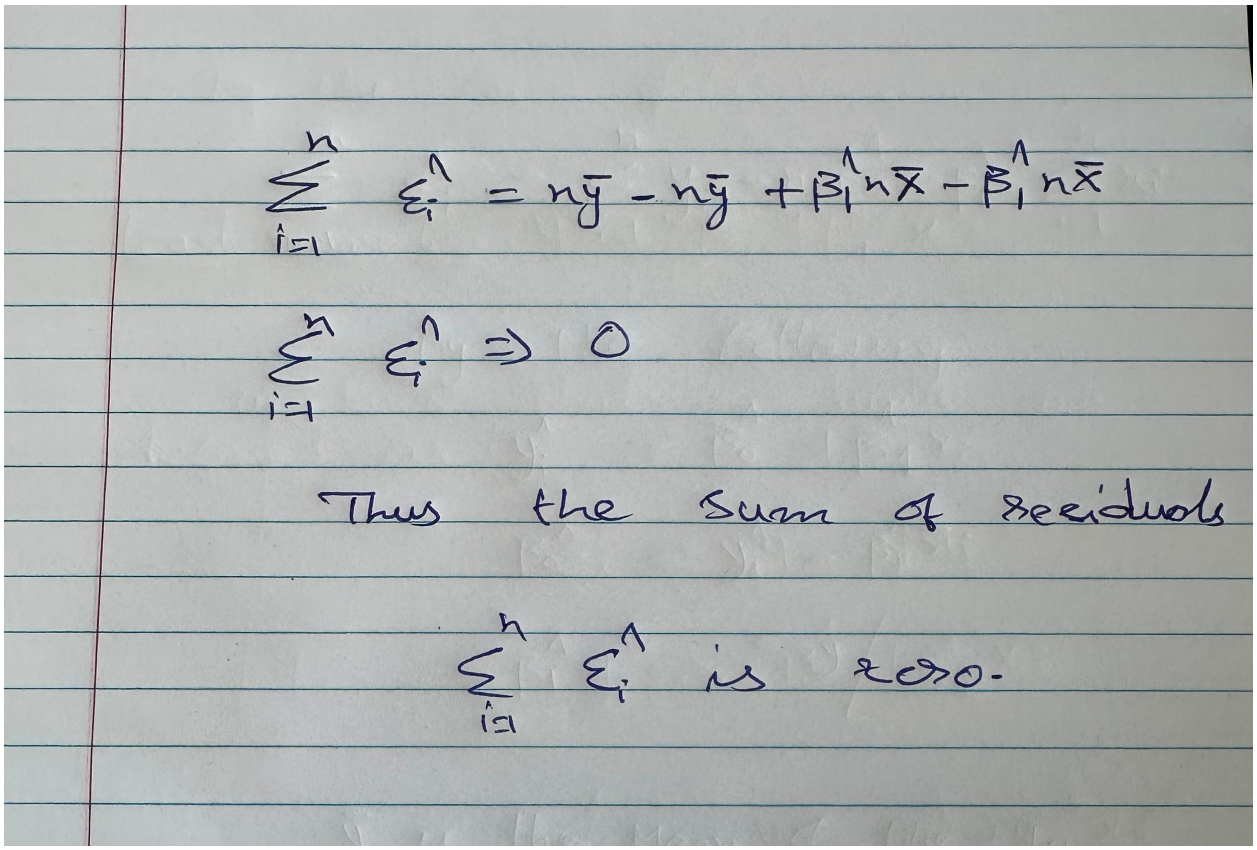
$$\sum_{i=1}^{n} \hat{\varepsilon}_i = n\bar{y} - n\hat{\beta}_0 - \hat{\beta}_1 n\bar{x}$$

we know $\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\sum_{i=1}^{n} \hat{\varepsilon}_i \Rightarrow n\bar{y} - n\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) - \hat{\beta}_1 n\bar{x}$$

A20554334
AKHIL KUMAR MARNI

Due September 6th 2024, Friday, 11:59pm



A20554334
AKHIL KUMAR MARNI

Due September 6th 2024, Friday, 11:59pm