**ASSIGNMENT 1**

**REGRESSION MODELS**

**Sahil Khan - 202482066**

Question1:

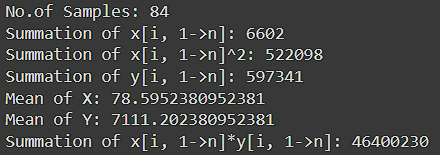
1. Obtain the estimated regression function.

****

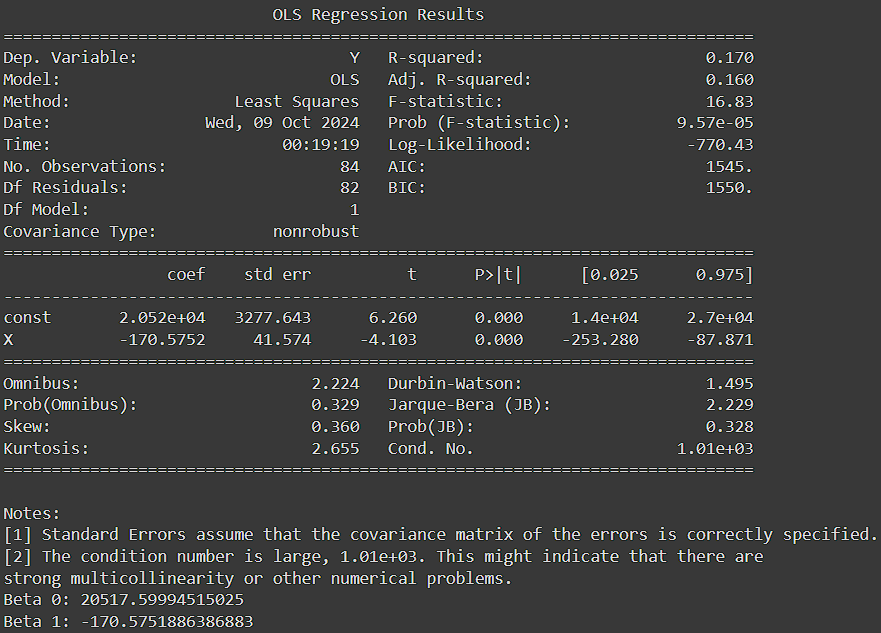
Where:

*β*0 is the intercept and *β*1 is the slope.

* Variables used to solve for β0 and *β*1.

****

Summary of the Model:



**Regression Equation**

The regression equation obtained from the data is:

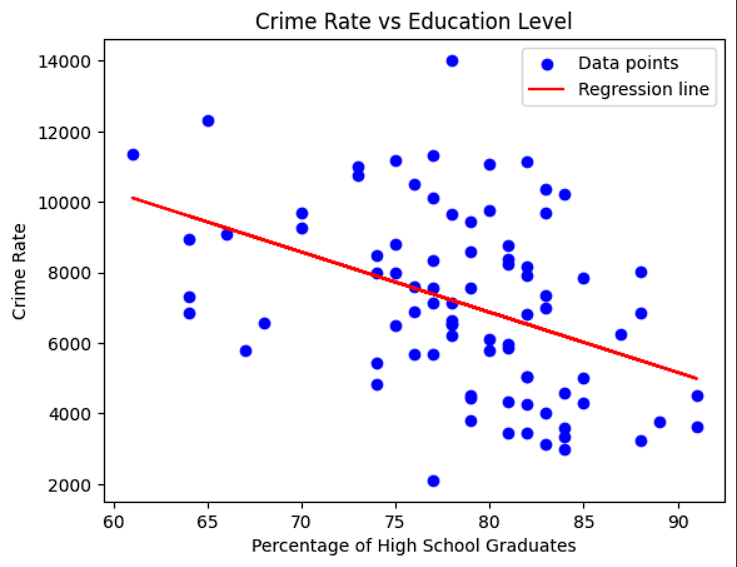
ŷ= **20517*.*60** − **170*.*58***x*

Where:

• ŷis the predicted crime rate.

• *x* is the percentage of individuals with at least a high school diploma.

1. Scatterplot of Crime Rate v/s Education Level



**Scatterplot Interpretation:**

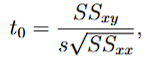
The scatter plot shows a negative correlation between the percentage of high school graduates and the crime rate. The regression line, which represents the best linear fit to the data, slopes downwards. This suggests that as the percentage of high school graduates increases, the crime rate tends to decrease. However, the points are quite dispersed around the regression line, indicating that the relationship is not perfectly linear and there are other factors influencing the crime rate beyond just the high school graduation rate. The model doesn't perfectly predict the crime rate based solely on graduation rates.

(i) Null and Alternative hypotheses.

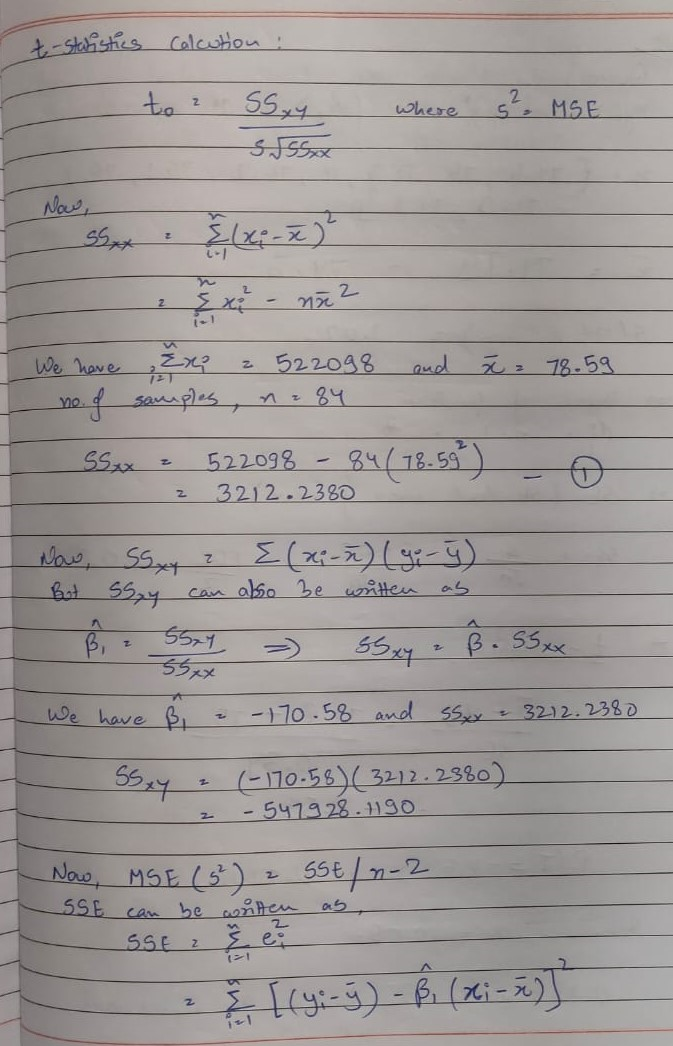
 Null hypotheses, H0: β1​=0 (No difference in crime rate based on education)

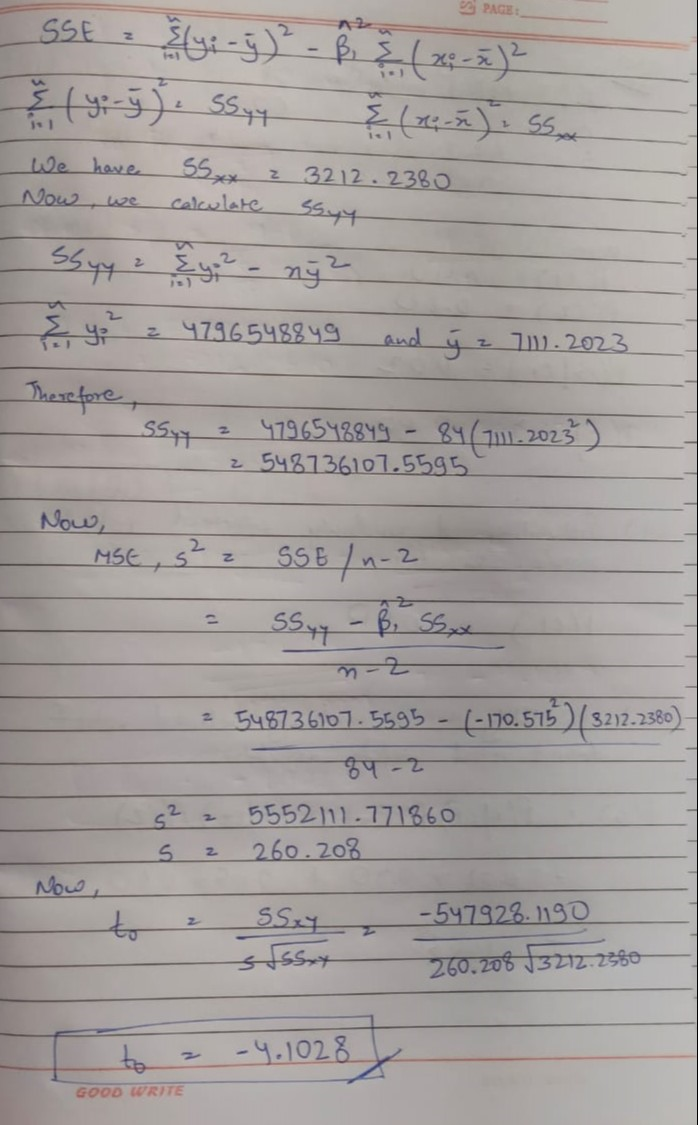
 Alternative hypotheses, H1: β1≠0 (There is a difference)

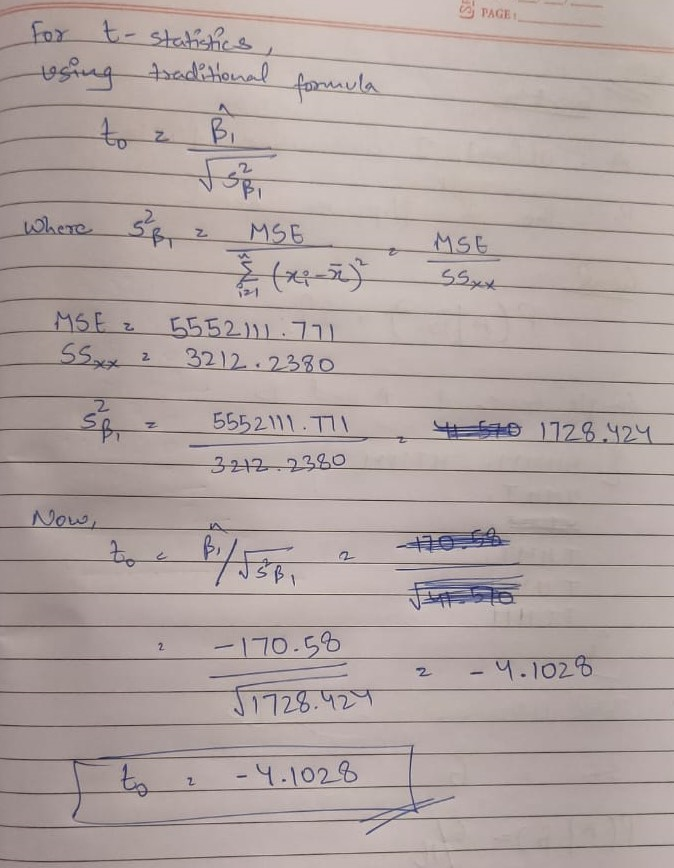
(ii) t-Statistics Calculation:

 where s2 = MSE

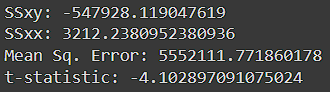
SSxy = Σ(xi - x̄)(yi - ȳ), SSxx = Σ(xi - x̄)2 and MSE = Mean Sq. Error







Output:



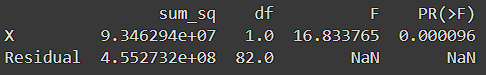
(iii) p – value from the model comes out to be:

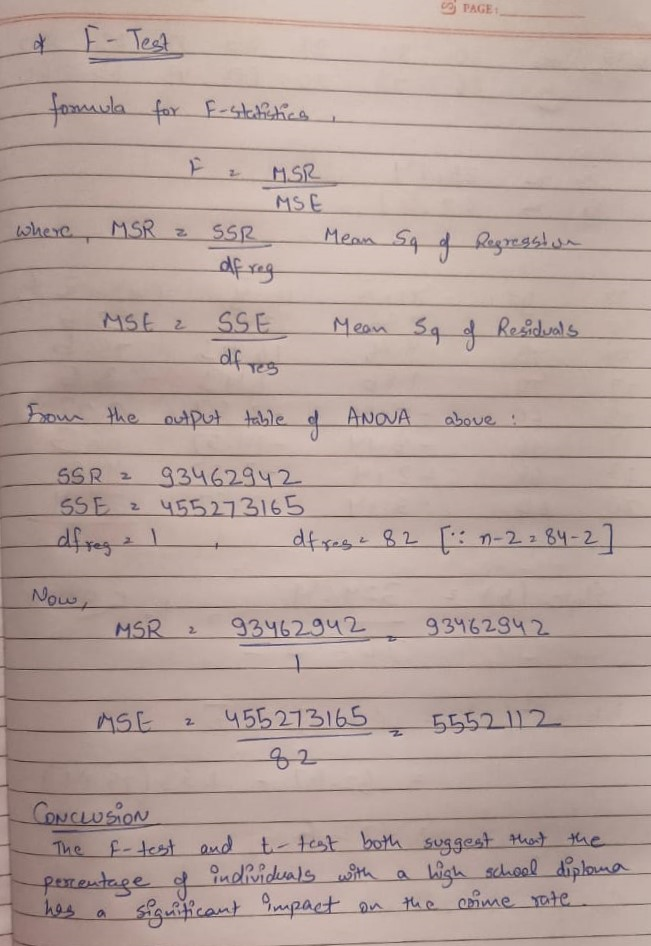


The p-value associated with this t-statistic is **0.0000957** (or 9.57 × 10−5).

The low p-value (less than 0.05) leads us to **reject the null hypothesis**. This provides strong evidence of a statistically significant relationship between high school graduation rates and crime rates.

(iv) Perform ANOVA Test:



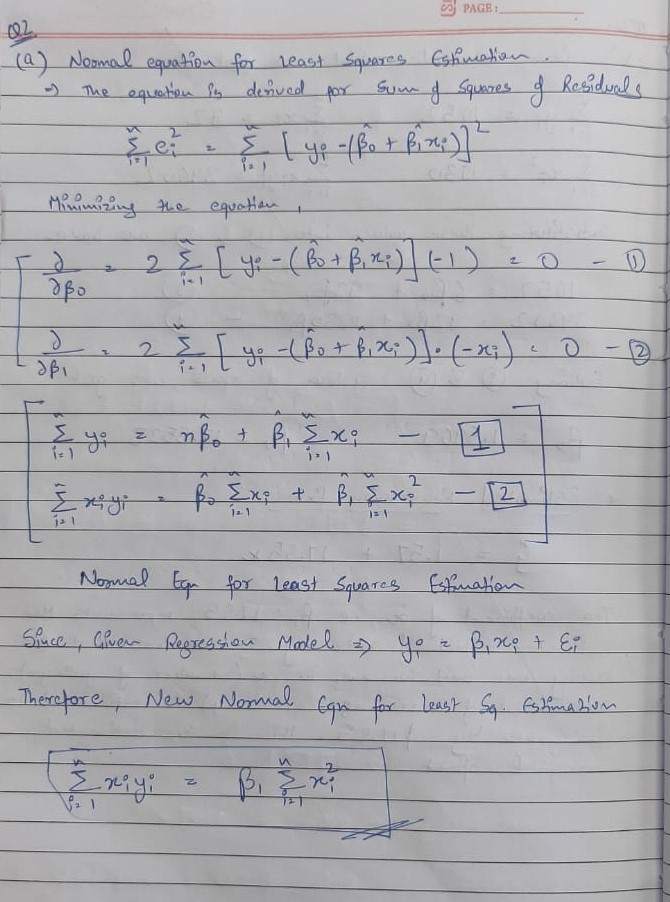


**Interpretation and Conclusions:**

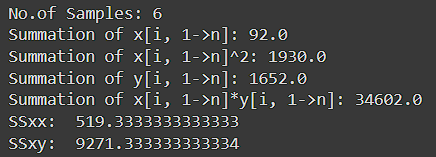
1. **What does MSR estimate?**
   * MSR estimates the portion of the total variation in the crime rate that is explained by the regression model. In this case, the value of **93,462,942** indicates the variation in the crime rate that can be attributed to changes in the percentage of high school graduates. A higher MSR suggests that the model is effectively explaining a significant amount of the variation in the dependent variable (crime rate).
2. **What does MSE estimate?**
   * MSE estimates the variation in the crime rate that is not explained by the model; it is the average squared residuals (the differences between the observed and predicted values). The value of **5,552,112** indicates the degree of unexplained variance. Lower values of MSE suggest that the model fits the data well, as there is less residual error.
3. **When do MSR and MSE estimate the same quantity?**
   * MSR and MSE estimate the same quantity when the model perfectly fits the data perfectly.

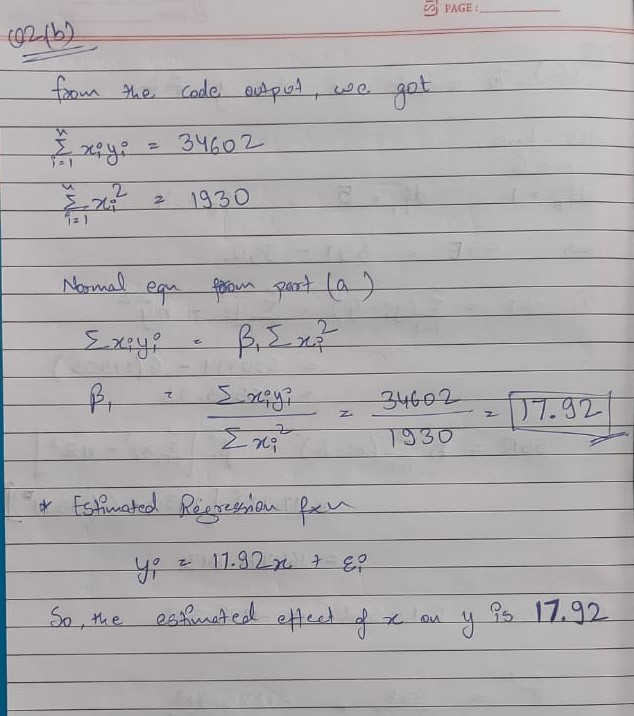
Question 2:

(a)

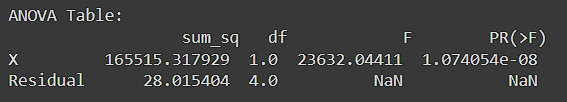


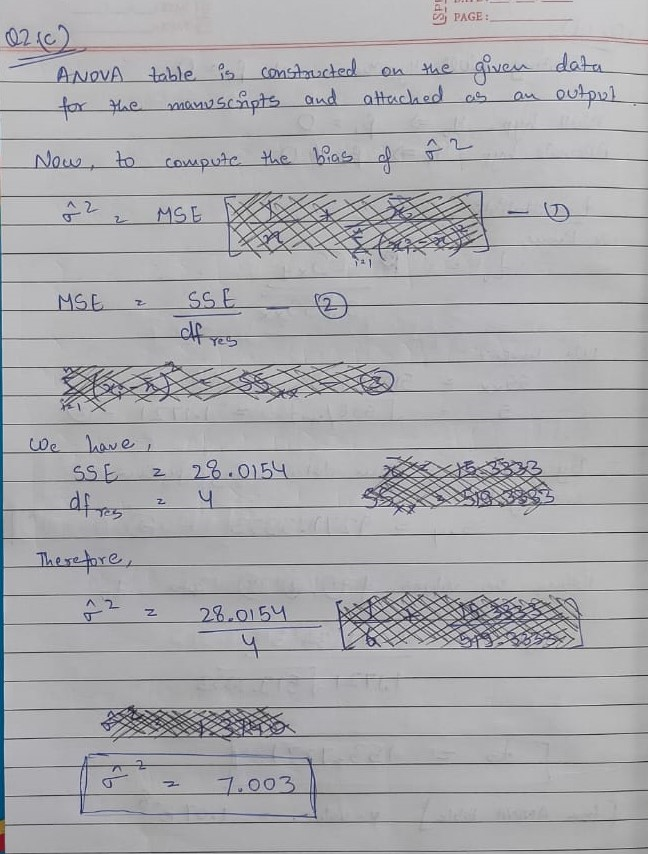
(b)



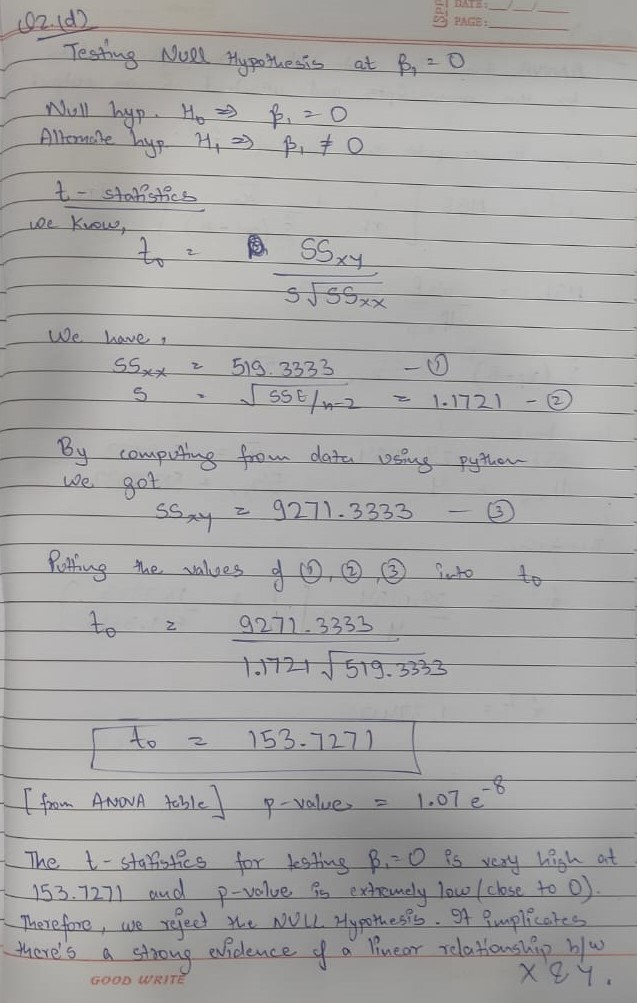


(c)





(d)



Question 3:

(a)



After generating five normal random numbers to serve as error terms, we calculated the y values for each corresponding x value using the regression model:

**E(y) = 20 + 4x + ϵ**

Using Least Squares Estimation:

Estimated intercept (b₀) = **24.180**

Estimated slope (b₁) = **4.256**

(b)



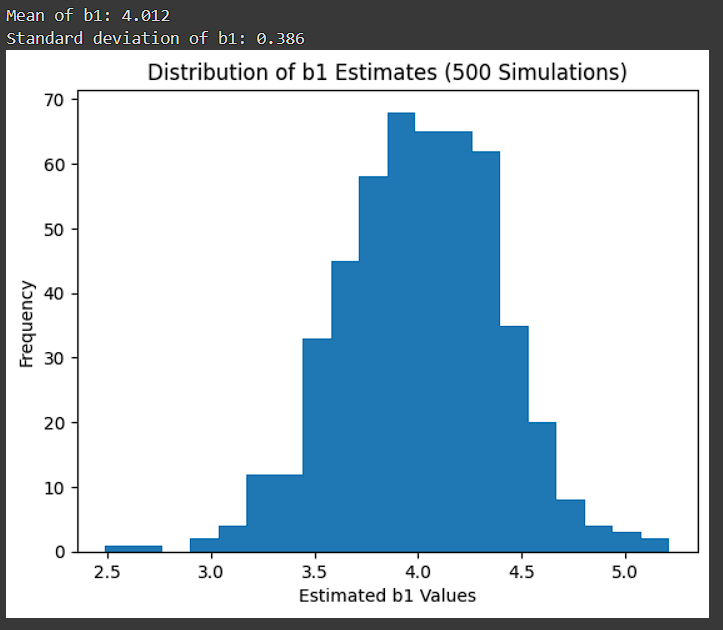
Using the regression model from part (a):

Predicted value of y (ŷ) at xℎ = 10: ŷℎ = **66.739**

95% Confidence Interval for E(yℎ) at xℎ = 10:

([Lower bound: **60.887**], [Upper bound: **72.592**])

(c)



After repeating parts (a) and (b) 500 times, we computed the following statistics for the estimates of the slope (β^1​):

* Mean of β^1 estimates: **4.021**
* Standard deviation of β^1​ estimates: **0.386**

These results align well with theoretical expectations. The mean is very close to the true slope of 4 from the regression model E(y) = 20 + 4x.

The standard deviation indicates that most of the slope estimates are clustered around this true value, with some variation. The frequency distribution of the slope estimates, displayed in the histogram, shows that the estimates are approximately normally distributed around the true slope of 4.

(d)

Confidence Interval Coverage:



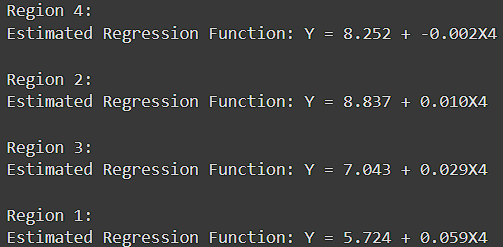
Proportion of the 500 confidence intervals that contain the true E(yℎ) at xℎ = 10: **0.978**

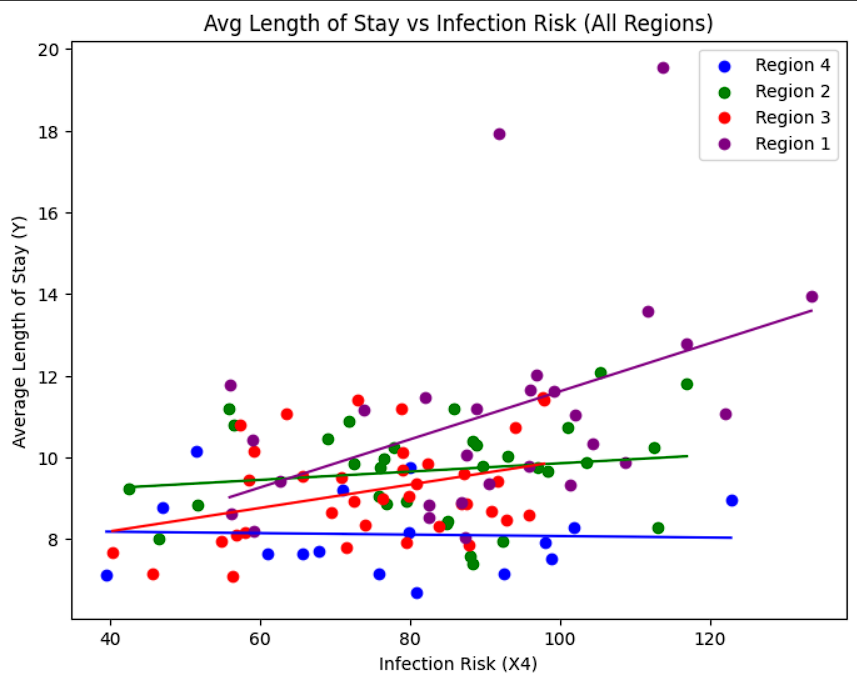
The observed proportion of confidence intervals containing the true value of E(yℎ) is close to the expected 95%. This demonstrates that the confidence intervals are performing as they should, providing a reliable way to estimate the true mean response at xℎ = 10.

Question 4:

(a)

**Regression Functions and Scatterplots**





I performed a linear regression analysis on each geographic region to explore the relationship between average length of stay and infection risk. I used various colors in the scatter plot to differentiate data points from each region, and plotted distinct lines to show the regression functions.

• **Region 1 (Purple)**: The slope is significantly steep, showing a strong positive correlation; as infection risk rises, the average length of stay increases substantially.

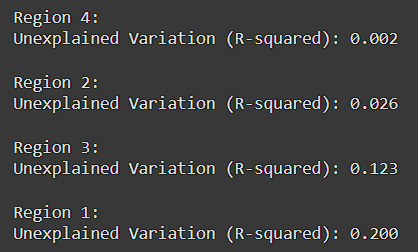
• **Region 2 (Green)** and **Region 3 (Red)**: Both exhibits moderately positive slopes, suggesting a moderate impact of infection risk on length of stay.

• **Region 4 (Blue)**: The line is almost flat, indicating a minimal relationship between infection risk and length of stay.

Overall, the plot highlights that the impact of infection risk on length of stay varies notably across regions, with **Region 1** experiencing the most pronounced effect.

(b)

**Unexplained Variation**



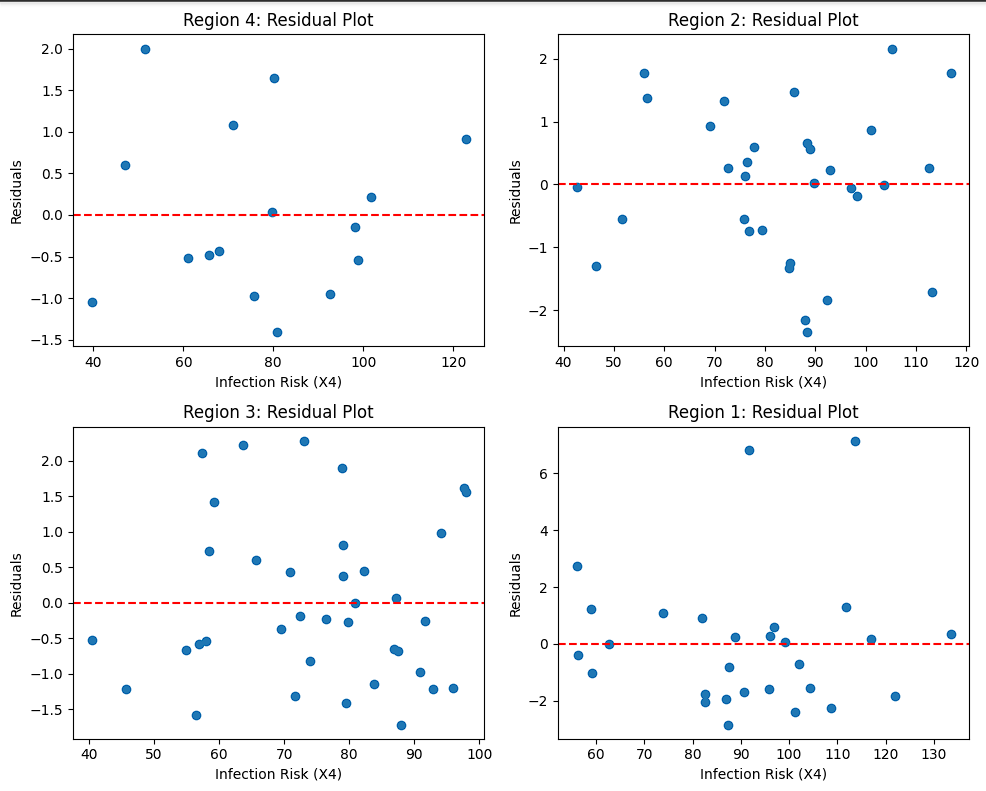
**Comparison:**

The unexplained variation reflects the amount of total variability in average length of stay (Y) that is not accounted for by the linear relationship with infection risk (X4).

Region 4 exhibits the lowest unexplained variation **(0.002)**, indicating that the linear model with infection risk explains almost all the variability in the average length of stay in this region. Conversely, Region 1 has the highest unexplained variation **(0.200)**, implying that infection risk is a weaker predictor of average length of stay in this region, or other unmeasured variables are playing a more substantial role here. Regions 2 and 3 fall in between, with Region 2 showing a slightly better fit than Region 3.

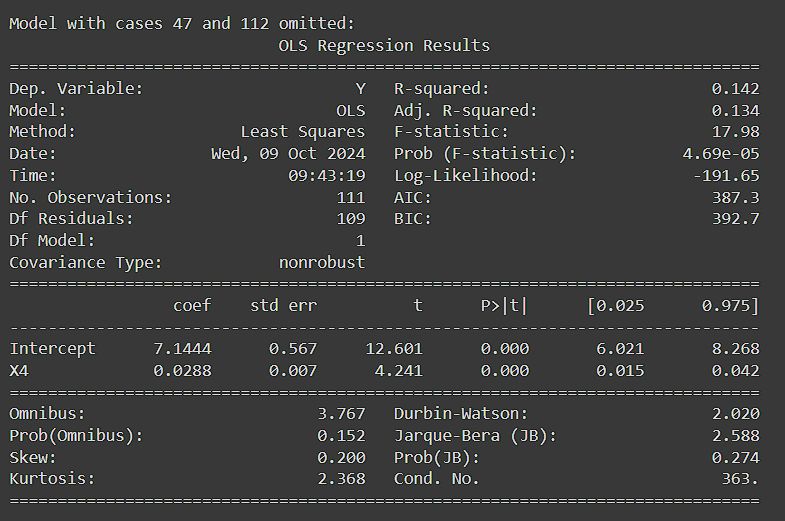
The variation in unexplained variation across regions suggests that the relationship between average length of stay and infection risk may vary geographically.

(c)



The residual plots for each region help evaluate how well the models fit the data. In all regions, the residuals are dispersed around zero, suggesting that the model fits the data fairly well without any noticeable patterns. There is no clear sign of heteroscedasticity (changing variance), indicating that the model appears suitable for the data.

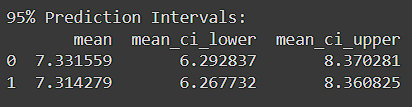
(d)



In comparison to the original model, there was a slight improvement in the fit, and the positive relationship between infection risk and length of stay persisted, albeit with some impact on the model.

(e)

95% prediction intervals for new y observations at x = 6.5 and x = 5.9



Observations y47 and y112:



I used the model from part (d) to compute 95% prediction intervals for two new observations with infection risks of 6.5 and 5.9:

• For x = 6.5, the predicted value of y is **7.33** with a confidence interval from **6.29** to **8.37**.

• For x = 5.9, the predicted value of y is **7.31** with a confidence interval from **6.26** to **8.36**.

It's evident that both case 47 (with y = 19.56) and case 112 (with y = 17.94) fall **outside the prediction intervals**, indicating that these two observations are clear outliers. This clarifies why removing them had a significant impact on the model in part (d).