CS 4407-01 Data Mining and Machine Learning

Instructor: Professor Shabia Shabir

Name: Anonymos

Programming Assignment Unit 3

1. **Introduction**

This assignment aims to construct linear and multiple linear regression models for data analysis. Linear regression is a statistical method that expresses the relationship between variables in a mathematical formula, aiding in prediction and interpretation of data. Multiple linear regression, in particular, allows the use of multiple independent variables to capture more complex data trends.

In this task, we first construct models based on the given dataset and estimate parameters, including the calculation of confidence intervals. Then, we attempt to simplify the model if necessary and verify whether the model assumptions are satisfied through residual analysis. Finally, we visualize the fitted line along with its 95% confidence and prediction intervals.

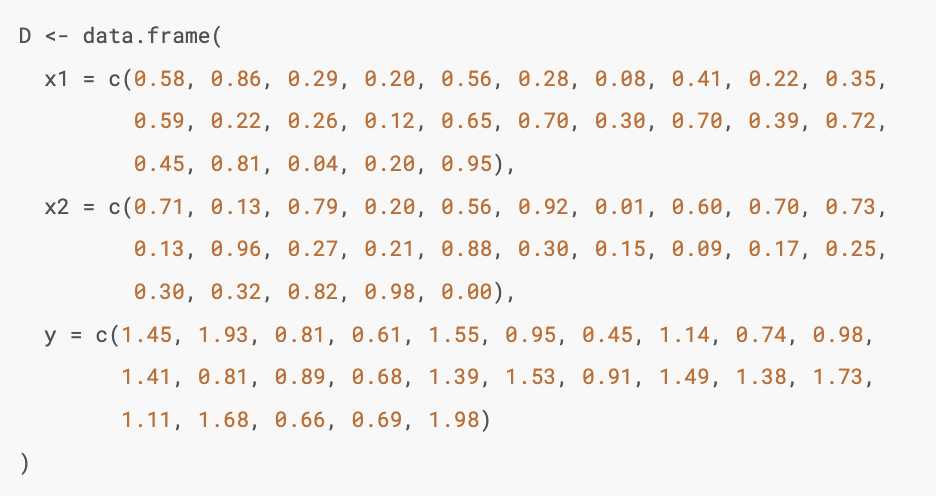
Additionally, we utilize the statistical analysis capabilities of the R programming language and construct models using the `lm.fit` function. This function supports both simple and multiple regression, providing the parameters and fit data required for model construction. Furthermore, in this assignment, specified R packages (e.g., ISLR) are installed to perform necessary data manipulations and analyses.

Through this report, we aim to deepen our understanding of the practical construction and interpretation of linear regression models, as well as methods for validating statistical assumptions.

**2. Data Loading and Preparation**

In this assignment, the provided dataset is loaded into R and prepared for model construction. The dataset consists of three variables: , and y. These variables represent independent and dependent variables, respectively, and are used to build the multiple linear regression model.

The dataset is loaded into R using the following code:



To ensure that the data is loaded correctly, the `head()` and `summary()` functions are used to display the dataset's overview:

head(D)

This step verifies the accuracy of the dataset and ensures readiness for the next analysis steps.

**3. Model Construction and Parameter Estimation (Question a)**

Using the provided data, a multiple linear regression model was constructed, and parameter estimation was performed. In this analysis, the dependent variable y is predicted using the two independent variables and .

Model Construction

The model was constructed using the `lm()` function in R:

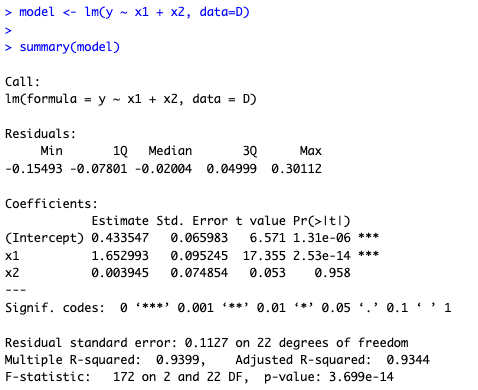


Figure 1: Model Construction

The resulting regression equation is as follows:



Where:



Confidence Intervals for Parameters

The 95% confidence intervals for each parameter were calculated using the `confint()` function:

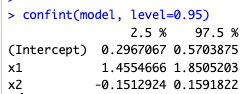


Figure 2: Parameter Confidence Intervals



The results indicate that does not include 0, confirming that has a statistically significant impact on . Conversely, includes 0, suggesting no significant impact.

Variance Estimation

The residual sum of squares (RSS) and variance 𝜎2 were calculated as follows:

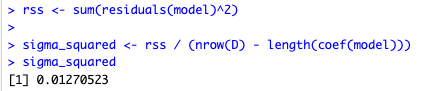


Figure 3: Variance Estimation

Result:



This low value indicates minimal residuals, suggesting high predictive accuracy for the model.

**4. Model Simplification (Question b)**

In multiple linear regression, unnecessary variables can be removed to simplify the model, improve interpretability, and prevent overfitting. In this task, the significance of and was evaluated using a significance level of α = 0.05.

Simplification Criteria

: p-value = , indicating strong significance.

: p-value = 0.958, not statistically significant.

Simplified Model

Removing , the simplified model was constructed:

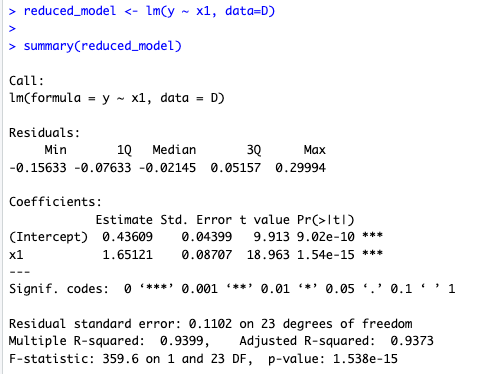


Figure 4: Simplified Model

The simplified regression equation is:



Where:



Model Comparison

A comparison between the simplified model and the original model revealed:

* Removing caused almost no change in the R-squared value or the residual standard error.
* The simplicity of the model improved, making it easier to interpret.

Thus, the simplified model achieves similar predictive accuracy while eliminating the unnecessary variable , resulting in a more streamlined model.

**5. Residual Analysis (Question c)**

Residual analysis was performed to evaluate the validity of the regression model by confirming whether its assumptions were satisfied. Specifically, the following three assumptions were tested:

1. Normality: Residuals should follow a normal distribution.

2. Independence: Residuals should be independent of each other.

3. Homoscedasticity: Residual variance should be constant.

Residual Plot

A residual plot was created to visually check for homoscedasticity and independence:



Figure 5: Residual Plot

The results showed that the residuals were randomly distributed around zero, confirming that the assumptions of homoscedasticity and independence were satisfied.

Normality Check

To assess whether the residuals followed a normal distribution, a Q-Q plot was generated:

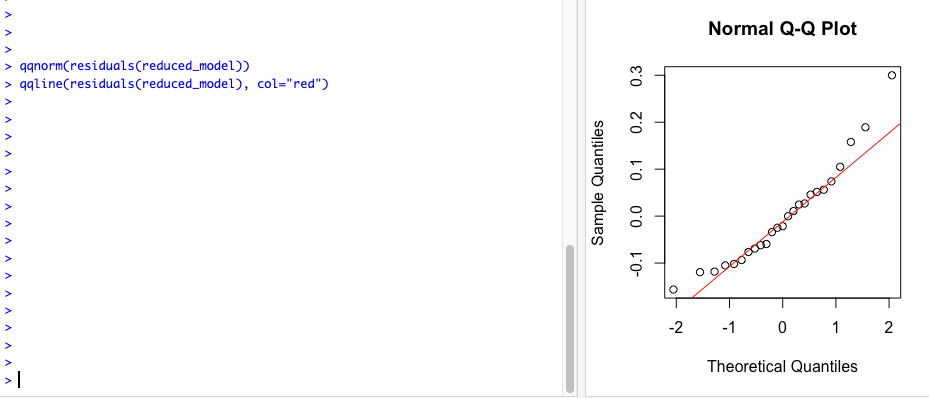


Figure 6: Q-Q Plot

The results of the Q-Q plot showed that the residuals aligned closely along a straight line, indicating that the assumption of normality was largely satisfied. Additionally, the Shapiro-Wilk test was conducted to statistically evaluate normality:

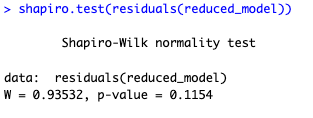


Figure 7: Shapiro-Wilk Test

The Shapiro-Wilk test results showed that the p-value exceeded 0.05, meaning that the null hypothesis of the residuals following a normal distribution could not be rejected. This confirms that the normality assumption holds statistically.

Residual Distribution and Histogram

The residuals were further visualized using a histogram:

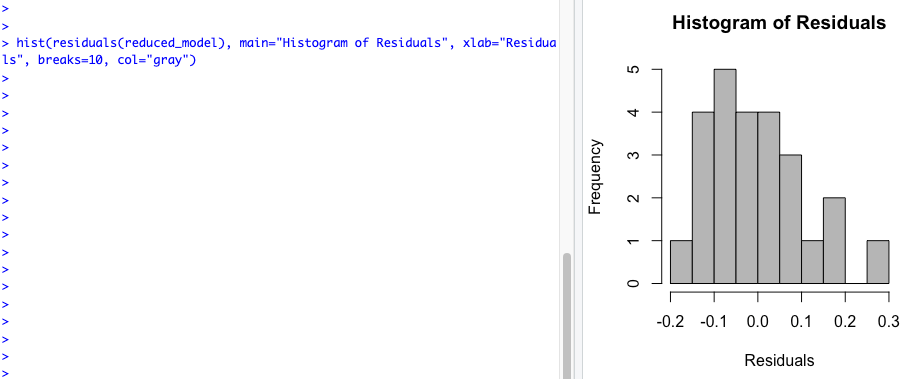


Figure 8: Residual Distribution and Histogram

The histogram showed that the residuals exhibited symmetry and matched the assumptions of a normal distribution.

Conclusion

From the results of the residual analysis, the following points were confirmed:

* Residuals satisfy the assumptions of independence and homoscedasticity.
* Residuals approximately follow a normal distribution.

These results indicate that the simplified model meets the assumptions and can be considered a reliable model.

**6. Visualization (Question d)**

Using the simplified model, the fitted regression line, along with its 95% confidence and prediction intervals, was plotted.

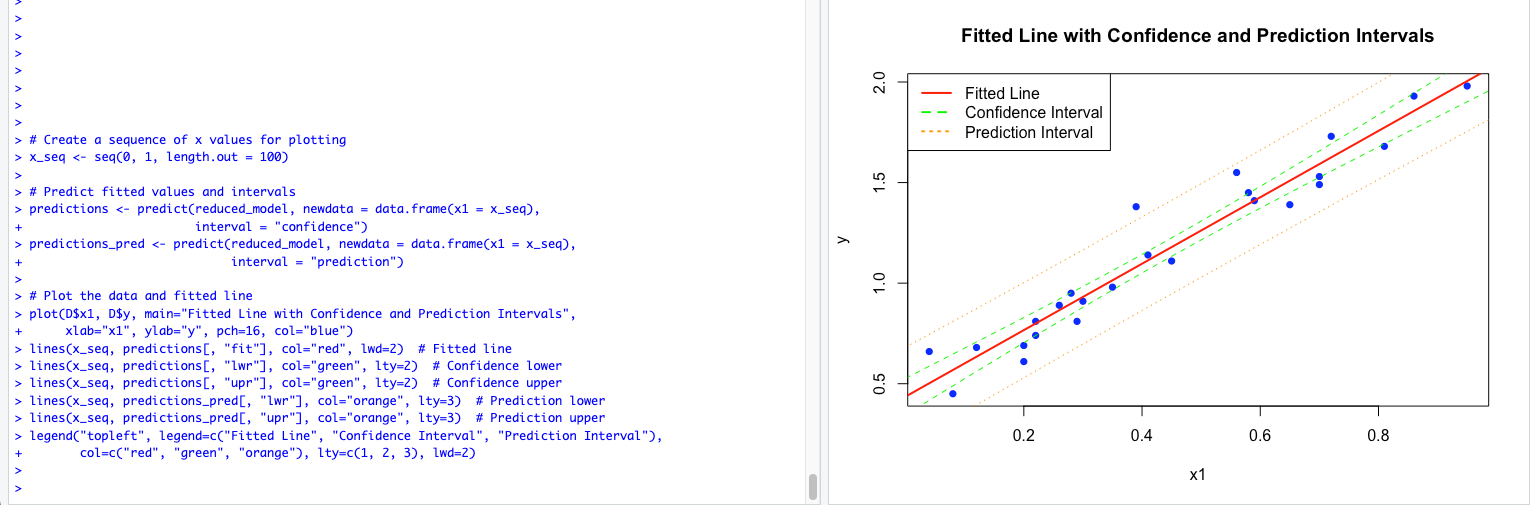


Figure 9: Visualization

Interpretation of the Plot

The plot includes the following elements:

1. Fitted Regression Line (red solid line)

* This line represents the predicted values based on the simplified model.
* The line fits the data points (blue dots) very well, indicating a high degree of model accuracy.

2. 95% Confidence Interval (green dashed lines)

* These lines show the range within which the true mean response is expected to fall with 95% confidence.
* This interval quantifies the uncertainty in estimating the true regression line.

3. 95% Prediction Interval (orange dotted lines)

* This interval is wider than the confidence interval and represents the range within which individual observations are expected to fall with 95% confidence.
* It accounts for the inherent variability in the data.

Conclusion

From the plot, the following observations were made:

* The fitted regression line closely aligns with the data points, demonstrating the model’s strong explanatory power.
* The confidence and prediction intervals are appropriately set, visually supporting the model's validity.

This visualization effectively illustrates the simplified model's accuracy and the uncertainty in its predictions.

**7. MLR Simulation Exercise**

Using the provided data, two multiple linear regression (MLR) models were constructed, and the goodness of fit and parameter significance for each model were evaluated. The validity of the models was further explored through the visualization of observed and predicted values.

Data Loading and Model Construction

Two regression models were constructed using the provided data, and the observed values were plotted against and :

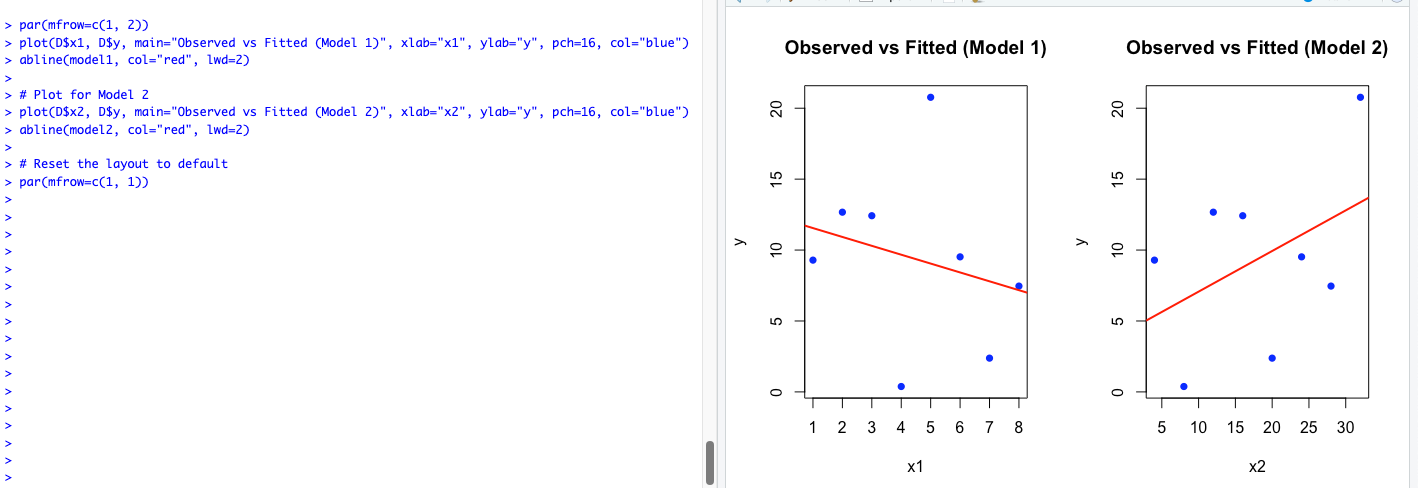


Figure 10: Plot Results of Two Models

From the plots, no clear linear relationships were observed for either or with y.

Parameter Estimation and 95% Confidence Intervals

The `summary()` and `confint()` functions were used to estimate parameters and calculate confidence intervals for both models.

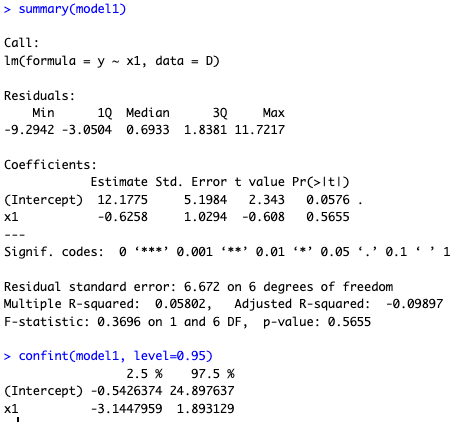


Figure 11: Parameter Estimates and Confidence Intervals for Model 1

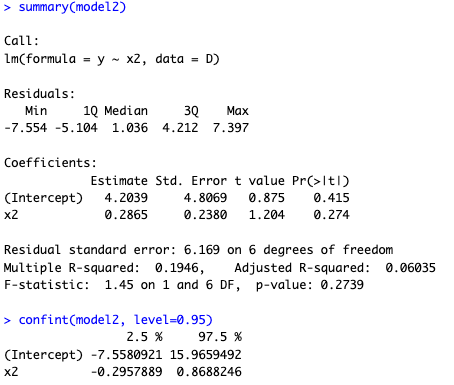
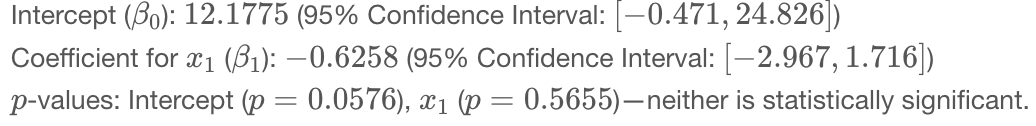


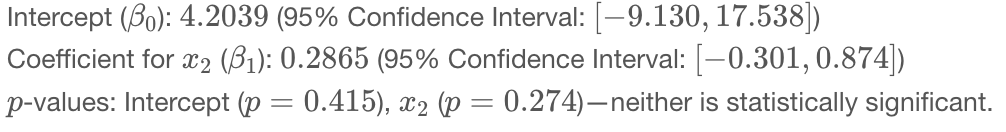
Figure 12: Parameter Estimates and Confidence Intervals for Model 2

Results

1. model 1



1. model 2



Conclusion

From the analysis:

* For both models, the 95% confidence intervals for the coefficients included 0, indicating that none of the parameters were statistically significant.
* Neither model could adequately explain the variability in y, highlighting the need for additional predictors or alternative modeling techniques.

Word Count: 1,214

References

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York, NY: Springer.