STA 3180 Fall 2024

LAB 1: Simple Linear and Multiple Regression

Objective: Find the best model for predicting a quantitative response variable from a set of explanatory variables.

Key Words: scatterplots, simple linear regression, residual plots, correlation, r^2, multiple regression, r^2-adjusted.

Set-up: The data set, "Hikes", provides various measurements of the forty-six mountains in the Adirondacks of upstate New York. They are known as the High Peaks with elevations near or above 4000 feet. Researchers would like to find a regression model for the expected *Time* (in hours) to hike these peaks using the variables: *Elevation* (in feet), *Ascent* (in feet), *Difficulty* (on a 1-7 scale with 7 being the most difficult) and *Length* (of roundtrip in miles). **Use the provided data set to answer the following questions and provide supporting output with code.**

1. The researchers would like to fit a simple linear regression model, so they must decide which of the four explanatory variables to select. First, create a correlation matrix using the four explanatory variables and the response variable. Next, create four separate scatterplots with one for each of the four explanatory variables with the response variable. Use the information in the correlation matrix and the scatterplots to recommend to the researchers which variable would be their best choice to predict *Time*. Justify your selection.

Explanatory Variable selected: _Length (of roundtrip in miles)_____

Justification of the Variable Selected: __*Length* would be the best explanatory variable to select because it has the highest correlation value (0.8585079) with *Time* when compared to the other explanatory variables and is close to 1. _____

2. Use the variable you selected in question 1 to fit a simple linear model. Use this information to write the least squares regression equation for your model. Make sure to use variable names in your equation.

Predicted time = 2.04817 + 0.68427* Length (of roundtrip in miles)

3. Interpret the meaning of the slope value in your regression equation in the context of the question.

For each 1 additional roundtrip mile in length, the predicted time will increase by 0.68427%, on average.

4. Find and interpret the values of r and R^2 for your model.

R: 0.8585 → There is a strong positive linear correlation between length (of roundtrip in miles) and time. As length increases the expected time to hike the mountain peaks would also increase.

R^2: 0.7370 → 73.70% of the variation in Time can be explained by the linear regression of Time onto length in roundtrip miles.

5. Obtain a Normal Quantile plot of the residuals (show plot). Which assumption of linear regression can be verified from this plot? Is this assumption met? Explain.

The Normal Quantile Plot of the residuals suggests that the condition of **normal distribution** of the residuals may not have been met as the points do not follow a linear pattern. The points are seen to have a U-pattern. Additionally, the points in the center (to the diagonal line), suggest that the middle part of the data distribution is approximately normal, with the points at the ends implying the presence of outliers.

6. Obtain a plot of the residuals vs fitted values (predicted values) – show plot. Which assumption(s) of linear regression can be verified using this plot? Is this assumption met? Explain.

The Residuals vs. Predicted Plot suggests that the condition of **equal variance** may not have been met since the spread increases as the predicted values increase. The points are seen to have a funnel-shaped pattern.

7. Now, use the correlation matrix from question 1 to discuss the relationships between **ONLY** the four explanatory variables.

1. 3 of the 4 explanatory variables (Ascent, Difficulty, and Length) are positively correlated and are moderately to highly correlated.

2. As said earlier, Time and Length have the highest correlation at 0.8585079.

3. The high correlation between the explanatory variables could cause problems with multicollinearity, which suggests some variables may need to be eliminated from the model.

8. An ecologist for the New York Department of Natural Resources contacts you. You explain to the ecologist that predicting *Time* for the low-difficulty mountains may be completely different from the *Time* for high-difficulty mountains. Split the data into two parts based on *Difficulty,* with low-difficulty having ratings from 1-4, and above 4 would be high-difficulty peaks. **Create two separate multiple regression models (one for low difficulty and one for high difficulty) using the three other additional explanatory variables of *Elevation, Ascent*, and *Length* to predict *Time*.** Provide the output for both models.

9. Consider the model for the low-difficulty peaks. Which, if any, of the predictors are significant in predicting *Time*? Justify your answer using the computer output.

The predictor that is significant in predicting *Time* is *Length*. *Length* has the lowest p-value (0.00298) and is the only predictor that has a significant value at 99% and 95% confidence intervals.

10. Consider the model for the high-difficulty peaks. Which, if any, of the predictors are significant in predicting *Time*? Justify your answer using the computer output.

The predictors that are significant in predicting *Time* are *Elevation* and *Length*. Both predictors have the lowest p-values (0.00234 and 1.9e-08 respectively) and are the only predictors that have a significant value at 99% and 95% confidence intervals.

11. You now decide to use a stepwise procedure to determine the "best" model separately for both low-difficulty and high-difficulty peaks. Show your labeled outputs and write the best model for each level.

Length is the most significant factor affecting hike time for both low (p-value = 0.000256) and high-difficulty (p-value = 1.9e-08) hikes, with Elevation also affecting high-difficulty hikes negatively (p-value = 0.00234). Additionally, the low-difficulty model is the best because it has a larger R^2 value (0.7896) than the high-difficulty model (0.7134). This also means that the simpler model (with only Length) is a better fit compared to the more complex models involving Elevation and Ascent.

12. We would like to predict the hike time for Mt Haystack. Use the appropriate model from question 11 to predict the time to hike Mt Haystack. Show work.

Mt Haystack length = 17.8 roundtrip miles

Predicted time =β 0 + β1*Length (of roundtrip in miles)

Predicted time = 1.9551 + 0.6683(17.8)

Predicted time = 13.8508 (of roundtrip in miles)

13. Now, use the original model from question #2 (which did not differentiate between difficulty levels) to predict the time to hike Mt Haystack. Show work.

Predicted time = 2.04817 + 0.68427* Length (of roundtrip in miles)

Predicted time = 2.04817 + 0.68427(17.8)

Predicted time = 14.2282 (of roundtrip in miles)

14. Compare your two predictions in #12 and #13 to the actual time to hike Mt Haystack. Which model worked best at making this prediction? Compute residuals to justify your answer.

Mt Haystack actual time = 12

The model that worked best at making a time prediction was the low-difficulty model from question #12. This was because the residuals gathered from the low-difficulty model (-1.8508) were lower and closer to the actual time than the original model from question #2 (-2.2282).