LAB 5 ASSIGNMENT

LINEAR REGRESSION AND CORRELATION

In this lab assignment, you will use simple linear regression to examine the relationship between two variables. In particular, you will use scatterplots to display the pattern of the data and correlation to measure the strength and direction of the relationship. Different regression models will be compared in terms of their ability to produce reliable predictions. Moreover, the regression assumptions will be tested with residual plots. Before you start working on the assignment problem, you should get familiar with the course material about regression and correlation and with the *Lab 5 Instructions*.

Rollercoasters: A Variable Design

Roller coasters first appeared in the 17th century, the first roller coaster appearing in North America in 1912 and then rising to greater popularity in the 1920s. Several variables could relate to the top speed of a roller coaster (the maximum speed it can achieve), such as the height of the initial climb, the length of the entire track, and the number of inversions (the number of times the roller coaster is fully upside down). This lab will identify the relationships between top speed and height as well as length for 21 selected roller coasters. Data for this list were gathered from a combination of online sources.

The necessary Excel file is available in the Data link located in the Lab 5 tab display in the Labs section on eClass. The data are not to be printed in your submission. **Unless otherwise stated, numerical answers should be to four significant decimal places.**

The following is a description of the variables in the data file:

Column	Name of Variable	Description of Variable
1	Name	name of the roller coaster,
2	Location	amusement park and nation of each roller coaster,
3	Height	height of the roller coaster's initial climb (in m),
4	Length	full length of the roller coaster (in km),
5	Inversions	number of times the roller coaster is fully upside down,
6	TopSpeed	the top speed of the car (in km/h).

Use the data to answer the following questions:

1. Before carrying out statistical analysis for the data, examine the study design. Is this an observational study or an experiment? Are population inferences applicable? Causal inferences? Explain briefly.

2. First, analyze the relationship between top speed and each explanatory variable mentioned below with a scatterplot.

   (a) Make a scatterplot of *TopSpeed* vs. *Height*. You may follow the instructions in the *Lab 5 Instructions* to obtain the plot. Paste the scatterplot with the title and names of the axes into your report. Make sure that the format of your scatterplot is consistent with the format in the *Lab 5 Instructions* (title, no lines or grids, axes rescaled to display only the observed values).

   (b) Describe the overall pattern of the relationship. Is it linear? Is the relationship strong, moderate, or weak? Is there a positive or negative association, or neither?

   (c) Repeat parts (a) and (b) with a scatterplot of *TopSpeed* vs. *Length*.

Now you will use the *Correlation* feature in the *Data Analysis* menu to assess the strength and direction of the relationship for each pair of variables with correlation.

(d) What is the correlation coefficient for each pair of variables in part (a) and (c)?

(e) Do the sign and magnitude of the coefficient confirm the conclusions you reached above? Explain briefly.

(f) Regress *TopSpeed* on *Height* for the dataset. Report the percentage of variation in *TopSpeed* that is explained by its linear regression on *Height*. Does the regression model fit the data well?

(g) Repeat part (f) while regressing *TopSpeed* on *Length* for the dataset.

3. Now you will perform the regression of *TopSpeed* on *Height*. In the *Regression* dialog box, check the *Residual Plots* boxes.

   (a) Paste the regression output into the report. Find the equation of the least-squares regression line to predict *TopSpeed* from *Height.*

   (b) Carry out an appropriate test to check for a positive slope. (See schema for all required components.)

   (c) Predict the top speed for the Kondaa roller coaster. Compare to the observed value of *y*.

   (d) Paste the plot of residuals against height. (Consider rescaling the axes first to display only the observed values.) Describe the pattern of residuals. Is there any evidence that the linear regression assumptions may be violated? Explain briefly.

4. Repeat parts (a) through (d) of Question 3 while regressing *TopSpeed* on *Length* for the dataset.

5. Compare briefly the results of regression of the two models. Which regression model is better? If an engineer wanted to design a new roller coaster to maximize top speed, which explanatory variable is more reliable?

LAB 5 ASSIGNMENT MARKING SCHEMA

Question 1 (6)

Type of study: 2 points
Population inferences: 2 points
Causal inferences: 2 points

Question 2 (32)

(a) Scatterplot of x_1 vs. y: 6 points
(b) Description of the pattern: 3 points
(c) Scatterplot of x_2 vs. y: 6 points
Description of the pattern: 3 points
(d) Correlation: 2 points each (4 points total)
(e) Comparison with scatterplots: 2 points each (4 points total)
(f) Percentage of variation: 2 points
Regression model fit: 1 point
(g) Percentage of variation: 2 points
Regression model fit: 1 point

Question 3 (30)

(a) Regression output: 6 points
Equation of least-squares regression: 3 points
(b) Test for positive slope: 8 points
(assumptions: 2, hypotheses: 1, distribution: 1, test statistic: 2, p-value: 1, conclusion: 1)
(c) Prediction: 2 points
Comparison: 2 points
(d) Residual plot: 6 points
Description of the pattern and relation to assumptions: 3 points

Question 4 (30)

(a) Regression output: 6 points
Equation of least-squares regression: 3 points
(b) Test for positive slope: 8 points
(assumptions: 2, hypotheses: 1, distribution: 1, test statistic: 2, p-value: 1, conclusion: 1)
(c) Prediction: 2 points
Comparison: 2 points
(d) Residual plot: 6 points
Description of the pattern and relation to assumptions: 3 points

Question 5 (6)

Summary: 2 points
Best model: 2 points
Explanatory variable of focus: 2 points

TOTAL = 104