

## LAB ASSIGNMENT 2

### REGRESSION AND CORRELATION

This lab assignment is based on a subset of data from the Framingham Heart Study. The study followed a cohort of 5209 men and women for over 25 years to identify and examine risk factors associated with cardiovascular disease. In this lab assignment you will use simple linear regression to explore the relationship between three variables: gender, systolic blood pressure, and age in a random sample of 50 men and women. In particular, you will use scatter plots to explore the relationships, 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. Before you start working on the assignment, you should get familiar with the course material about regression and correlation and with *Lab 2 Instructions*.

### The Framingham Heart Study

In this lab assignment you will apply both graphical and regression analysis tools in StatCrunch to examine the heart study data. The data are available in the StatCrunch file *lab2.txt* located on the *STAT 151* Laboratories web site at <http://www.stat.ualberta.ca/statslabs/stat151/index.htm> (click *Stat 151* link, and *Data* for *Lab 2*). The data are not to be printed in your submission.

The following is a description of the variables we have selected from the study for the purpose of this assignment:

<u>Column</u>	<u>Variable</u>	<u>Description of Variable</u>
1	Id	Subject Number,
2	Gender	Gender (1-Male, 2-Female),
3	Age	Age (30-64 years),
4	Systolic	Systolic blood pressure (82-300 mm).

Answer the following questions using the data:

1. First you will compare systolic blood pressure of the 50 males and females with a side-by-side boxplot.
  - (a) Obtain a side-by-side boxplot of systolic blood pressure for females and males. Paste the boxplot into your report.
  - (b) Given the side-by-side boxplot obtained in part (a), what are the appropriate measures of center and spread to compare the two distributions? Compare the centers, spreads, and shapes (symmetric, skewed) of the two distributions.
2. Now you will examine the relationship between systolic blood pressure and age for the two gender groups.
  - (a) Obtain a scatterplot of systolic blood pressure vs. age with different marking symbols for each gender. Paste the scatterplot into your report. The format of your scatterplot should be consistent with the format used in *Lab 2 Instructions* (title, names of the axes, and the legend for the two gender groups).
  - (b) Use the scatterplot obtained in part (a) to describe the relationship between systolic blood pressure and age for each gender. In particular, comment on the overall form (line, curve), direction (positive or negative) and strength (size of the scatter) of the relationship for males and females.

- (c) How does the relationship between systolic blood pressure and age for the males differ from the one for the females?
3. Now you will use the *Correlation* feature to assess the strength of the linear relationship between systolic blood pressure and age for men and women.
- (a) Obtain the correlation coefficients between systolic blood pressure and age for males and females. Paste the related output into your report.
- (b) Do the signs and magnitudes of the coefficients confirm your conclusions you have reached in Question 2? Explain briefly.
4. Now you will use the *Regression* feature in the *Stat* menu to obtain the least-squares regression lines separately for men and women.
- (a) Find the equations of two least-squares regression lines to predict systolic blood pressure from age for each gender group. Compare the slopes of the least-squares regression lines. Which systolic blood pressure increases faster with age, the one for men or the one for women?
- (b) What percent of the variation in systolic blood pressure for males and females is explained by their age?
- (c) Predict the systolic blood pressure of a man and a woman of 40 years old. Would you be able to predict the systolic blood pressure of a man and a woman of 70 years old? Explain briefly.
5. Obtain the residuals for each regression line and answer the following questions:
- (a) Use the *Summary Statistics (Column)* feature in the *Stat* menu to calculate the mean, standard deviation, median, quartiles, minimum and maximum of the residuals for each gender. Paste the outputs into your report.
- (b) What is the mean, and standard deviation of the residuals for each gender group? Identify the female and male subjects with largest residual.
6. Now you will investigate the effect of removing a single observation from the data on the regression results and the predictions.
- (a) Find the equation of the least-squares regression line to predict systolic blood pressure from age for males when the male subject with Id=14 (age 49, systolic blood pressure 100) is excluded from the regression analysis. Moreover, obtain the new value of R-square (fraction of the variation in systolic blood pressure that is explained by age for males) in this case.
- (b) Compare the slope of the new regression line with the slope of the regression line for all male observations in Question 4. Also report the value of R-square (fraction of the variation in systolic blood pressure that is explained by age for males) in this case. Explain the change in the value of R-square the slope after removing the single observation. You may refer to the scatterplot in Question 2 in your explanations.
- (c) Obtain the predicted systolic blood pressure of a male of 40 years old. Compare the new value with the value obtained in Question 4 part (c) and comment.

## LAB 2 ASSIGNMENT MARKING SCHEMA

Proper header and appearance: 10 points

### Question 1

- (a) Side-by-side boxplot of systolic blood pressure: 6 points
- (b) Appropriate measures of center and spread for comparison: 2 points  
Comparison of centers and spreads: 3 points

### Question 2

- (a) Properly formatted scatterplot with observations for both males and females: 6 points
- (b) Description of the relationship for each gender: 3 points each (6 points total)
- (c) Comparing the relationship for males and females: 3 points

### Question 3

- (a) Correlation coefficient for men (output): 2 points  
Correlation coefficient for women (output): 2 points
- (b) Comments about the correlation coefficients: 2 points

### Question 4

- (a) Equation of the least-squares regression line for men: 2 points  
Equation of the least-squares regression line for women: 2 points  
Comparison of the slopes and the question about the rate: 3 points
- (b) Percent of variation explained by age for men and women: 2 points each (4 points total)
- (c) Prediction for a 40 years old male and female: 2 points each (4 points total)  
Explanation why the prediction for a 70 years old subject would be unreliable: 2 points

### Question 5

- (a) Output for residuals for each gender: 2 points each (4 points total)
- (b) Mean and standard deviations of the residuals for each gender: 2 points each (4 points total)  
Cases with the largest absolute residuals: 2 points

### Question 6

- (a) Equation of the least-squares regression line for men when Id =14 removed: 2 points  
New R-square value: 2 points
- (b) Comparison of the slope and the R-square: 2 points  
Explanation: 3 points
- (c) New predicted systolic blood pressure for a male of 40 years old: 2 points  
Comments: 2 points

**TOTAL = 82**