ASSIGNMENT 3

SIMPLE LINEAR REGRESSION

In simple linear regression model the mean of a response variable is a linear function of a single explanatory variable. The model and the associated inferences are valid as long as the assumptions of independence, normality and constant variance are not violated.

In this assignment you will use a linear regression model to examine the relationship between the volume of air exhaled during a forced breath (an important measure of pulmonary function) and one of the following explanatory variables: age, height, sex, and parental smoking. In particular, the effect of parental smoking on the volume will be discussed. You will also compare the inferences based on the t-tools with those produced by simple linear regression model.

Second-hand Smoking and Childhood Respiratory Disease

Almost half of all Canadian children under the age of 15, some 2.8 million children, are exposed to second-hand smoke on a regular basis. The effect of second-hand smoking on the pulmonary function of children was studied by many researchers. In particular, studies have shown that children (especially infants) of parents who smoke have more lung illnesses, such as bronchitis and pneumonia, and can develop asthma. Still 85% of adults who smoke and who live with a child do not ensure that the child is not exposed to the smoke from their cigarettes.

One of the important measurements of lung function and effective tool to diagnose pulmonary disease is the volume of air expelled in the first second from the level of total lung capacity (FEV). It is expected that the volume may be lower for the children exposed to second-hand smoke at home.

In order to verify the hypothesis, the FEV measurements were obtained for 654 subjects, age 6-22 who were seen in Childhood Respiratory Disease Clinic in Easton Boston, Massachusetts. The data are part of a larger study and they are available in the SPSS file *lab3.sav* located on the STAT 252 laboratories web site at http://www.stat.ualberta.ca/statslabs/index.htm (click Stat 252 link). In order to download the data for the lab, click on the link *Data* for Lab 3 and follow the instructions. The data are not to be printed in your submission.

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

Column	**Variable**	**Description of Variables**
1	Age	Age (in years),
2	FEV	Volume of air expelled in the first second (in liters),
3	Height	Height (in inches),
4	Sex	Male or Female,
5	Smoker	Non = Nonsmoker parents, Cur = Current smoker parent(s).

1. First you will discuss the study design and compare the FEV measurements of children of smoker parent(s) with that of nonsmoker parents.

   (a) Comment about the study design. Is the study an example of an observational study or a randomized experiment? Can we use the study to assess the effect of second-hand smoking on FEV of children exposed to second-hand smoking at home? Provide brief explanations. List some potential confounding variables in the study.

(b) Use the *Explore* procedure to obtain the normality plots of FEV for smoker and nonsmoker parents. Paste the normality plots into your report. Is there any evidence that the assumption of normality necessary to apply the t-tools is violated in either case seriously?

(c) Use the *Compute* command in *Transform* menu to calculate the natural logarithm of FEV. Name this variable LNFEV and repeat part b for the new variable.

(d) Do the data provide any evidence that the mean of LNFEV of children of nonsmoker parents is less than that of children of smoker parents? Answer the question by carrying out an appropriate test in SPSS. In particular, define the null and alternative hypothesis, report the value of the test statistic and the p-value of the test.

(e) Define a new indicator variable Z, where Z= 0 for nonsmoker parents and Z = 1 for smoker parent(s). Define a simple linear regression model using LNFEV as the response variable and Z as the explanatory variable. Then use the *Regression* tools in SPSS to perform the regression and use the SPSS output to answer the following questions:

i. What is the equation of the least-squares regression line? Compare the value of the intercept of the regression line with the mean of LNEFV for nonsmoker parents. Compare the value of the intercept plus the value of the slope of the estimated regression line with the mean of LNFEV for smoker parent(s).

ii. Obtain the value of the test statistic for testing the hypothesis that the slope of the population regression line is zero versus it is negative. Compare the value of the test statistic with the value of the test statistic in part d using equal variances assumption. Moreover, compare the p-values of the two tests. What do you conclude?

iii. Now you will refer to the simple linear regression model defined in e to explain the observed relationships in parts i and ii. In particular, explain the relationship between the mean of LNFEV and the least-square coefficients. Moreover, demonstrate the equivalence of the t-test for comparing the means of LNVEV for smoker and nonsmoker parents and the t-test about the slope in the simple linear regression model. Are the test assumptions also equivalent for the two t-tests? Explain briefly.

2. In this part you will use the *Scatter Plot* feature in the *Graphs* menu to examine the relationship between LNFEV and each of the two explanatory variables: Age and Height.

(a) Obtain the scatter plot of LNFEV versus Age with different markers for smoker and nonsmoker parents. Describe the relationship between LNFEV and Age for the two parent groups. In particular, comment about linearity, direction (positive, negative or neither) and strength of the relationship. Are there any outliers?

(b) Obtain the scatter plot of LNFEV versus Height. Answer the questions in part (a) with Height as the explanatory variable. Does the plot indicate any difference in LNFEV between the two groups?

(c) Obtain the scatter plot of Height versus Age with different markers for smoker and nonsmoker parents. Is there a linear relationship between Height and Age? If you had to choose only one of the two explanatory variables Age and Height to predict LNFEV, which one would you choose?

3. Now you will use the *Correlate* feature in the *Analyze* menu to quantify the strength of the relationship between the variables with correlation.

(a) Obtain the Pearson's correlation coefficient between LNFEV and Age, between LNFEV and Height, and between Age and Height. You may copy and paste the related SPSS output to support your answers. Do the sign and the magnitudes of the coefficients confirm the conclusions you reached in Question 2? Explain briefly.

(b) Find partial correlation between LNFEV and Age controlling for Height and partial correlation between LNFEV and Height controlling for Age. You may copy the correlation values from the SPSS output. If you had to choose one variable, either Age or Height to predict LNFEV, which one would you choose? Explain briefly.

4. Now you will use *Regression* tool in *Analyze* menu to compare the mean of LNEFV for smoker and nonsmoker parents using Height as an explanatory variable. First use *Select Cases* in *Data* to split the data into two parts, one for smoker parents and the other for nonsmoker parents.

   (a) Use the *Regression* tool in SPSS to find the least-squares estimate of the regression line using LNEFV as the response variable and Height as the explanatory variable for smoker and nonsmoker parents separately. Refer to the scatter plot obtained in Question 2 part (b) to justify fitting two separate regression lines to the data.

   (b) Interpret the values of the slope of the least-squares regression line for each parent group. What percentage of the variation in LNFEV is explained by Height for each group?

   (b) Obtain the plots of standardized residuals versus standardized predicted values for the two regression models in part (a) and paste them into your report. Describe the pattern of the residuals in each plot. Do the residuals appear to be scattered about horizontal line at zero? Is there any outlier? Which assumption(s) can be verified by the plot?

   (d) Obtain the normal probability plots of residuals (P-P plots) for the two regression models in part (a) and paste them into your report. Is there any evidence that the assumption of normality is violated? Explain briefly.

5. How will the predictive ability of the regression model change by including both height and parent smoking status as the explanatory variables? You will answer the question by fitting a multiple regression model for the data and comparing the proportion of variation of the response variable explained by each model.

   (a) Use *Split File* feature in *Data* to carry out the regression analysis for all observations (the parent groups combined). Then use the *Regression* tool to fit a new regression model to predict LNFEV by using Height and Z as the explanatory variables. The indicator variable Z was defined in Question 1 e. What is the least-squares estimate of the linear regression model? What is the percentage of the variation in LNFEV that can be explained by the model?

   (b) Substitute Z for zero and one to find the estimated regression lines for nonsmoker and smoker parents, respectively. Are your results consistent with your results in Question 4 part b? Explain briefly. Which of the two regression models, the one fitted in part a or the one discussed in Question 4 is more adequate to describe the relationship between LNFEV and Height for smoker and nonsmoker parents? Explain briefly.

ASSIGNMENT 3 MARKING SCHEMA

Proper Cover Page: 10 points

Question 1

a) Study design: 2 point
Casual inferences: 2 points
Confounding variables: 2 points

b) Normality plots: 3 points each (6 points total)
Comments about normality: 2 points

c) Normality plots for transformed data: 3 points each (6 points total)
Comments about normality: 2 points

d) Null and alternative hypotheses: 2 points
The test statistic with equal variances: 2 points
P-value (one-sided): 2 points
Conclusions: 2 points

e)

i. The equation of estimated regression line: 2 points
Comparing intercept and slope with means: 2 points

ii. The test statistics: 2 points
Comparing the test statistics: 2 point
Comparing the p-values: 2 point
Conclusions: 2 point

iii. Explanation of the relationship between the means of LNFEV and regression coefficients: 3 points
Explanation of the relationship between the two tests: 3 points
Equivalence of the assumptions for the two tests: 3 points

Question 2

a) Scatterplot (properly formatted, different markers for the two groups): 4 points
Pattern in the scatterplot (linearity, direction strength) for each parent group: 3 points each (6 points total)
Outlier(s): 1 point

b) Scatter plot (properly formatted, different markers for the two groups): 4 points
Pattern in the scatterplot (linearity, direction strength) for each parent group: 3 points each (6 points total)
Outlier(s): 1 point

c) Scatter plot (properly formatted with different markers for the two groups): 4 points
Comments about the relationship: 2 points
Which explanatory variable is a better predictor: 2 points

Question 3

a) Pearson's correlation coefficients: 2 points
Comments about sign and magnitude of the coefficients: 2 points

b) Partial correlation coefficients: 2 points
Which of the two predictors is better? 2 points

Question 4

a) The equation of estimated regression line for nonsmoker and smoker parents: 2 points
The equation of estimated regression line for nonsmoker and nonsmoker parents: 2 points

b) Interpretation of the slopes: 2 points
R^2 Value: 2 points

c) Plots of standardized residuals versus standardized predicted values: 2 points each (4 points total)
Pattern of residuals: 3 points
Assumption tested: 1 point
Outliers: 1 point

d) P-P plots of residuals: 2 points each (4 points total)
Description of the pattern in the plots: 3 points

Question 5

a) The equation of estimated regression line: 2 points
Percentage of the variation explained: 2 points

b) Estimated regression lines for each parent group: 2 points each (4 points total)
Comparing regression models in Question 4 and Question 5: 4 points

TOTAL = 135

Created by Alizera Simchi
Revised by Henryk Kolacz
September 2012