LAB ASSIGNMENT 4

MULTIPLE LINEAR REGRESSION

In a multiple linear regression model, the mean of a response variable is a linear function of several explanatory variables. The model and associated inferential tools are valid as long as the assumptions of independence, normality and constant variance are not seriously violated. In this assignment, you will use the multiple linear regression tools available in SPSS to study the relationship between the percentage of body fat and age, weight, height and various body circumference measurements for 252 men. You will use the first 200 observations in the data set to develop an equation to predict the body fat percentage of a subject based on the above numerical characteristics. Then you will test the validity of the equation using the data for the remaining 52 observations. Moreover, you will verify the regression model assumptions with the appropriate diagnostic tools in SPSS.

Estimating Percentage of Body Fat

It is well known that body fat is an important factor for health. The higher your percentage of fat above the average levels, the higher your health risk for some weight-related illness, like heart disease, high blood pressure, and diabetes. Therefore, estimating the percentage of body fat to assess your health is highly recommended by a variety of popular health books. One of the accurate techniques to obtain the percentage of body fat is underwater weighing in which body volume is computed as the difference between body weight measured in air and weight measured during water submersion. In other words, body volume is equal to the loss of weight in water with the appropriate temperature correction for the water's density. As the technique of underwater weighing though very accurate is very inconvenient and costly, some new techniques have been developed that require some simple body circumference measurements (e.g. abdominal circumference) and skin-fold measurements obtained by using a caliper. The new techniques also use height and weight of an individual to estimate the body fat percentage.

In order to establish a generalized body fat percentage prediction equation by using simple measurement techniques, a study was conducted in 1985 by a group of researchers in Europe. 252 men were randomly selected from a larger pool of volunteers. The percentage of body fat for an individual was obtained by the technique of underwater weighing. The information for age, weight, height and various body circumference measurements was collected for each participant.

The data are available in the SPSS file *lab4.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 4 and follow the instructions. The data are not to be printed in your submission. The following is the description of the variables in the data file:

Column	Variable Name	Description of Variable
1	PERC	Percent body fat
2	AGE	Age (years)
3	WEIGHT	Weight (lbs)
4	HEIGHT	Height (inches)
5	NECK	Neck circumference (cm)
6	CHEST	Chest circumference (cm)
7	ABDOMEN	Abdomen 2 circumference (cm)
8	HIP	Hip circumference (cm)
9	THIGH	Thigh circumference (cm)
10	FOREARM	Forearm circumference (cm)
11	WRIST	Wrist circumference (cm)

Answer the following questions by using the data:

1. First you will analyze the study design.

   (a) What is the purpose of the study? What are the experimental units in the study? What is the response variable? How many explanatory variables are there? What assumptions should you make in order to generalize your conclusion to the population of interest?

   (b) Given the fact that men tend to store fat around their middle (apple body shape), which of these measurements seem to be better predictors of the body fat percentage for men? Does the accuracy of these measurements matter to produce good estimates of body fat percentage? What can you do to reduce the measurement errors in practice?

2. Now you will examine the relationship between pairs of variables with a matrix of scatterplots.

   (a) Obtain a matrix of scatterplots of the 11 variables. Paste the scatterplot into your report.

   (b) Comment on the relationship between PERC and each of the 10 predictors. Does a linear model appear to be appropriate for describing the relationship between the percentage of body fat and the 10 predictors? Which of the predictors seem to have the strongest relationship with the response? Are there any outliers? You may need to obtain separate plots to evaluate the relationship properly. Do not paste these plots into your report.

3. Obtain the correlation matrix of the 11 variables. Paste the matrix into your report.

   (a) Describe the relationship between PERC and each of the 10 predictors. Is the relationship linear? Is it strong or quite weak? Is it positive or negative or neither?

   (b) If you wished to use only one explanatory variable to predict percent body fat, which variable would you choose? Is multicollinearity a potential problem in this case?

4. Define a multiple regression model with PERC as the response variable and the 10 predictors. State the model assumptions.

5. Use the forward model selection procedure for the first 200 observations to obtain the least squares fit to the linear regression model defined in Question 4.

   (a) How many of the 10 predictors were included by the forward model selection procedure? What is the order in which they were added?

   (b) What is the estimated regression equation for the final model determined by the forward model selection procedure?

   (c) What percentage of the variation in PERC is explained by the explanatory variables in the final model?

   (d) Is the fitted model as a whole significant at a 0.05 significance level? State this question as null and alternative hypotheses about the regression coefficients; report the value of the test statistic, the distribution of the test statistic under the null hypothesis, the p-value of the test; and give your conclusion.

   (e) Refer to the computer output to comment on how significantly each explanatory variable contributes individually, given the other variables in the model.

6. In this part, you will use the first 200 observations in the data to verify the regression model assumptions with some diagnostic plots. Moreover, you will use some case-influence statistics to identify the influential cases.

   (a) Obtain the normal probability plot of standardized residuals for the model obtained in Question 5. Paste the plot into your report. Is there evidence that the assumption of normality is violated?

   (b) Obtain the plot of standardized residuals versus standardized predicted values for the model obtained in Question 5 and paste the plot into your report. Describe the pattern of the residuals. Is there any indication that the assumption of equal variance may be violated? Are there any outliers?

   (c) Obtain the case-statistics (studentized residuals, Cook's distance and leverages) to identify the influential case(s) if any for the model in Question 5. Provide the case statistics for the influential case(s).

7. Rerun the forward regression without the influential case(s) identified in Question 6 for the first 200 observations. Paste the summary of the least squares fit obtained in SPSS into your report and answer the following questions.

   (a) What is the estimated regression equation? What percentage of the variation in PERC is explained by the explanatory variables in the final model? Compare the percentage of the variation explained by the model with the one in Question 5.

   (b) What is the predicted value of the percentage of body fat for an individual with a body weight of 184.75 pounds, an abdomen 2 circumference of 86.4 cm, and a wrist circumference of 18.2 cm?

   (c) Use the SPSS output to find the 95% confidence interval for the mean percentage of body fat of the individuals specified in Part (b). Moreover, find the 95% prediction interval for an individual with the body circumference measurements and body weight specified in Part (b). Which of the two intervals is wider? Explain why.

   (d) Is there convincing evidence of an association between the body fat percentage and the lower body measurements (THIGH and HIP) after accounting for age, weight, height, neck, chest, abdomen, forearm and wrist? State this question as null and alternative hypotheses about the appropriate regression coefficients. Use appropriate SPSS computer output and hand calculations to obtain the value of the test statistic. Report the p-value of the test from the output. State your conclusions.

8. Now you will test the validity of the fit obtained in Question 7 by comparing the exact fat percentage measurements obtained by underwater weighing with the predicted values produced by the fit.

   (a) First use the equation derived in Part (a) of Question 7 to obtain the predicted percent fat for the remaining 52 observations in the data set. Do not paste the values into your report.

   (b) Then use a paired t-test on the mean difference between actual and predicted percent fat. Report a 95% confidence interval for the mean difference. Moreover, report the p-value of the test. State your conclusions briefly.

LAB ASSIGNMENT 4 MARKING SCHEMA

Proper Header: 10 points

Question 1

(a) Purpose of the study: 1 point
Experimental units: 1 point
Response variable: 1 point
Explanatory variables: 1 point
Assumptions: 1 point
(b) Better predictor: 1 point
Accuracy of measurements: 1 point
Reduce measurement errors: 2 point

Question 2

(a) Scatterplot: 6 points
(b) Describe the relationships: 2 points
Validity of a linear model: 1 point
Predictor has the strongest relationship with PERC: 1 point
Outliers: 1 point

Question 3

Correlation matrix: 6 points

(a) Describe the relationships: 3 points
(b) One explanatory to predict PERC: 1 point
Multicollinearity: 2 points

Question 4

Define regression model: 3 points
Model assumptions: 3 points

Question 5

(a) Number of predictors: 1 point
Order: 2 points
(b) Estimated regression equation: 3 points
(c) R2: 2 points
(d) Null and alternative hypotheses: 2 points
Value of the F-statistic and p-value: 2 points
Null distribution: 2 points
Conclusion: 2 points
(e) Conclusion: 2 points

Question 6

(a) Normal probability plot: 6 points
Normality assumption: 2 points
(b) Residual plot: 6 points
Equal variance assumption: 2 points
Outliers: 2 points

(c) Influential case: 1 point
Case statistics: 3 points

Question 7

SPSS outputs: 3 points

(a) Estimated regression equation: 3 points
R2: 2 points
Comparison: 1 point
(b) Predicted value: 2 points
(c) Confidence interval: 2 points
Prediction interval: 2 points
Wider and why: 3 points
(d) Null and alternative hypotheses: 2 points
Value of the F-statistic: 6 points
P-value: 2 points
Conclusion: 2 points

Question 8

Paired t-test: 2 points
Confidence interval: 2 points
P-value: 2 points
Conclusion: 2 points

TOTAL= 126