Biostat 510: Statistical Computing Packages
SAS Homework 5
Due Thursday, Feb 24, 2005

Topics:

Read in raw data from a file
Create dummy variables
Simple Linear regression
Multiple Regression (check for collinearity)
Regression with dummy variables

1. Create a permanent SAS data set called LABDATA.AFIFI.
    a. Read in the raw data from the file AFIFI.DAT, which is in the data archive, data.exe, available on my web page: http://www.umich.edu/~kwelch. You can use the "Examples of Statistical Procedures" on page 130 of your course pack as a model. You do not need to set up formats or labels for the variables in the data set for this exercise. However, please be sure to read in ALL variables in the data set, even those not included in the example. You can get information on all the variables from the data set description on page 130. You do not need to create the new variables that are created in the data set on page 131.
    b. Create a new variable INSHOCK, which has a value of 1, if the patient was in one of the shocktypes, and has a value of 0 if the patient was not in shock.
    c. Create a set of dummy variables, SHOCKDUM2 through SHOCKDUM7, that indicate which SHOKTYPE each patient was in.
    d. Get descriptive statistics on all variables in your data set. Get frequency tabulations on SEX, SHOKTYPE, and SURVIVE to check the data.
    e. Get a Proc Contents on your new data set. How many observations are in this data set? How many variables?
2. Examine the relationship between Mean Arterial Pressure at Time 2 and Mean Arterial Pressure at Time 1.
    a. First, examine the Pearson Correlation between these two variables, using Proc Corr. What is the mean of MAP1, of MAP2? What is the correlation between these two variables? Is it significant?
    b. Create a scatter plot with MAP2 as the Y variable and MAP1 as the X variable, and include a linear regression line in your plot. You can do this by using either Proc Gplot or in SAS/INSIGHT.
        i. Describe this scatter plot in words. Do these two variables appear to be very strongly related to each other? Are they positively or negatively related?
3. Run a simple linear regression model to predict Mean Arterial Pressure at Time 2 as a linear function of Mean Arterial Pressure at Time 1 using Proc Reg.

a. Write out the regression equation calculated by Proc Reg. What is the estimated intercept? The estimated slope? Please interpret these two parameter estimates.
b. Using Proc Reg, get a plot of the residuals from this regression model vs. the predicted values. Does this graph appear to show constant variance?
c. Using Proc Univariate get a histogram and normal q-q plot for the residuals from this regression. Do these residuals appear to be drawn from a normal distribution?
d. Include the regression output, and the plots in your homework write-up.

4. Run this same regression model using SAS/INSIGHT.
    a. Compare the results of this model with those of the above model run using Proc Reg. Are they the same?
    b. Include the results of this regression model in your write-up.
5. Examine the relationship among the variables: Mean Arterial Pressure at time 2, Mean Arterial Pressure at time 1, Systolic Blood Pressure at time 1, Diastolic Blood Pressure at time 1, Central Venous Pressure at time 1, Hematocrit at time 1 and Hemoglobin at time 1.
    a. Create a Pearson correlation matrix. Which variables appear to be highly correlated with each other (either positively or negatively). Include this correlation matrix in your homework.
    b. Create a scatter plot matrix for these variables. You do not need to include this scatter plot matrix in your homework.
6. Run a multiple regression model to predict Mean Arterial Pressure at time 2 as a function of: Mean Arterial Pressure at time 1, Systolic Blood Pressure at time 1, Diastolic Blood Pressure at time 1, Central Venous Pressure at time 1, Hematocrit at time 1 and Hemoglobin at time 1.
    a. What is the R-square for this model?
    b. What is the overall model significance for this model? Report the F-test, the numerator and denominator degrees of freedom and the p-value for this model.
    c. Check the tolerance, variance inflation factor, and collinearity diagnostics for this model. Which predictor variables appear to be collinear with which other variables in this model?
    d. Include the results for this regression model in your homework.
7. Run another multiple regression model to predict Mean Arterial Pressure at time 2 as a function of a subset of the above predictor variables. Select your subset of predictor variables so that you will not have a collinearity problem with this new model.
    a. What is the R-square for this model?
    b. What is the overall model significance for this model? Report the overall F-test, the numerator and denominator degrees of freedom and the p-value.
    c. Check the tolerance, variance inflation factor, and collinearity diagnostics for this new model. Be sure that you do not have a collinearity problem with this model.
    d. Compare the estimated regression coefficient of each of the predictor variables included in this new model with it’s regression coefficient from

the model in question 6. How have these changed? Compare the standard errors of each of the estimated regression coefficients with their standard errors from the model in question 6. How have these standard errors changed?

e. Get a plot of the residuals vs. the predicted values for this regression. Please comment on this plot.
f. Get a histogram and normal q-q plot of the residuals for this regression. Please comment on this plot.
g. Include the results of both of this regression model in your homework.

8. Save your command file as homework5.sas. Make a printout of your commands and hand it in, along with your output from SAS. Write up brief answers to each question. Make sure that your commands can run all at once, by simply clicking on the submit button, without selecting any commands.