NAME _____ DECEMBER 7, 2012

MATH 445 - STATISTICAL METHODS LONG EXAM III

Instructions: *For p-values, write exactly what you get from R (do not round it off).*

1. Consider the Health Exam data from the U.S. Department of Health and Human Services, National Center for Health Statistics, Third National Health and Nutrition Examination Survey. It has a total of 80 cases (40 males and 40 females) with each case having values for 14 variables. These variables are: Gender, Age (in years), Height (in inches), Weight (in pounds), Waist (circumference in cm.), Pulse (pulse rate in beats per minute), SysBP (systolic blood pressure in mmHg), DiasBP (diastolic blood pressure in mmHg), Cholesterol (in mg), BodyMass (body mass index), Leg (upper leg length in cm), Elbow (elbow breadth in cm), Wrist (wrist breadth in cm), Arm (arm circumference in cm).

a. Determine if Gender has a significant effect on Cholesterol. Explain the meaning of the estimates of the regression parameters. [5]

b. Consider the model BodyMass $\sim$ Height+Weight+Pulse+Arm. Compute the following:

i. Adjusted R^2. [2]

ii. AIC [3]

iii. Mallow's CP statistic. [3]

iv. SBC [3]

v. BIC [3]

vi. PRESS [3]

vii. VIF [3]

c. Using the Mallow's CP statistic, determine if we should remove the predictor 'Arm' from the model in part (b). Explain your answer. [5]

d. Based on the VIF values that you obtained for the predictors in the model of part (b), what can you conclude? [3]

e. Using the Backward selection procedure, determine the 2 worst preditors for BodyMass. [4]

f. Using the Stepwise selection procedure, determine the best model for BodyMass using the AIC criterion. [5]

g. Using the model in part (b), find the semi-studentized, studentized, and the studentized deleted residuals for cases 30 and 65. [8]

h. Using the model in part (b), find the DFFITS and Cook's Distance for cases 30 and 65. Explain the meaning of your results. [7]

i. The yield (Y) of a chemical process depends on the temperature (X_1) and pressure (X_2). The following nonlinear regression model is expected to be applicable:

$$Y_i = \gamma_0 (X_{i1})^{\gamma_1} (X_{i2})^{\gamma_2} + \epsilon_i$$

Prior to beginning full-scale production, 18 tests were undertaken to study the process yield for various temperature and pressure combinations. The results are in the attached data set

`(data_chemical_process.csv)`.

- To obtain starting values for γ_0, γ_1, and γ_2, note that when we ignore the random error term, a logarithmic transformation yields $Y_i' = \beta_0 + \beta_1 X_{i1}' + \beta_2 X_{i2}'$, where $Y_i' = \log_{10} Y_i$, $\beta_0 = \log_{10} \gamma_0$, $\beta_1 = \gamma_1$, $X_{i1}' = \log_{10} X_{i1}$, $\beta_2 = \gamma_2$, and $X_{i2}' = \log_{10} X_{i2}$. Fit a first-order multiple regression model to the transformed data, and use the estimates for β_0, β_1, and β_2 that you obtained to get starting values for γ_0, γ_1, and γ_2. What are these starting values? [5]

- Using the starting values for γ_0, γ_1, and γ_2 that you obtained in part (a), find the least squares estimates of these three parameters. Write down the R command that you used. [5]

- Compute the SSE and MSE for this model. [5]