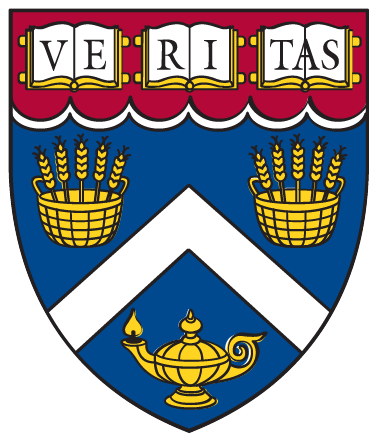
CSCI E-106: Data Modeling



Fall 2019

Dr.Hakan Gogtas

Assignment 7

Due: November 18, 2019 at 11:59pm EST

**Instructions:** Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

1. Refer to the CDI data set. A regression model relating serious crime rate (Y, total serious crimes divided by total population) to population density (X1, total population divided by land area) and unemployment rate (X3) is to be constructed. (15 pts)
2. Fit second-order regression model (equation 8.8 on the book). Plot the residuals against the fitted values. How well does the second-order model appear to fit the data? What is R2? (5pts)
3. Test whether or not all quadratic and interaction terms can be dropped from the regression model; use α = .01. State the alternatives, decision rule, and conclusion. (5pts)
4. Instead of the predictor variable population density, total population (X1) and land area (X2) are to be employed as separate predictor variables, in addition to unemployment rate (X3). The regression model should contain linear and quadratic terms for total population, and linear terms only for land area and unemployment rate. (No interaction terms are to be included in this model.) Fit this regression model and obtain R2. Is this coefficient of multiple determination substantially different from the one for the regression model in part a? (5pts)
5. Refer to the CDI data set. The number of active physicians (Y) is to be regressed against total population (X1), total personal income (X2), and geographic region (X3, X4, X5). (15pts)
6. Fit a first-order regression model. Let X3 =1 if NE and 0 otherwise, X4 = 1 if NC and 0 otherwise, and X5 = 1I if S and 0 otherwise. (5pts)
7. Examine whether the effect for the northeastern region on number of active physicians differs from the effect for the north central region by constructing an appropriate 90 percent confidence interval. Interpret your interval estimate. (5pts)
8. Test whether any geographic effects are present; use α= .10. State the alternatives, decision rule, and conclusion. What is the P-value of the test? (5pts)
9. Refer to the Lung pressure Data. Increased arterial blood pressure in the lungs frequently leads to the development of heart failure in patients with chronic obstructive pulmonary disease (COPD). The standard method for determining arterial lung pressure is invasive, technically difficult, and involves some risk to the patient. Radionuclide imaging is a noninvasive, less risky method for estimating arterial pressure in the lungs. To investigate the predictive ability of this method, a cardiologist collected data on 19 mild-to-moderate COPD patients. The data includes the invasive measure of systolic pulmonary arterial pressure (Y) and three potential noninvasive predictor variables. Two were obtained by using radionuclide imaging emptying rate of blood into the pumping chamber or the heart (X1) and ejection rate of blood pumped out of the heart into the lungs (X2) and the third predictor variable measures blood gas (X3). (25pts)
10. Fit the multiple regression function containing the three predictor variables us first-order terms. Does it appear that all predictor variables should be retained? (5pts)
11. Using first-order and second-order terms for each of the three predictor variables (centered around the mean) in the pool of potential X variables (including cross products of the first order terms), find the three best hierarchical subset regression models according to the R2a,p criterion. (5pts)
12. Is there much difference in R2a,p for the three best subset models? (5pts)
13. Calculate the PRESS statistic and compare it to SSE. What does this comparison suggest about the validity of MSE as an indicator of the predictive ability of the fitted model? (5pts)
14. Case 8 alone accounts for approximately one-half of the entire PRESS statistic. Would you recommend modification of the model because of the strong impact of this case? What are some corrective action options that would lessen the effect of case 8? (5pts)
15. Refer to the Website developer data set. Management is interested in determining

what variables have the greatest impact on production output in the release of new

customer websites. Data on 13 three-person website developed teams consisting of a project manager, a designer. and a developer are provided in the data set. Production data from January 2001 through August 2002 include four potential predictors; (1) the change in the website development process. (2) the size of the backlog of orders, (3) the team effect, and (4) the number of months experience of each team. (10 pts)

1. Develop a best subset model for predicting production output. Justify your choice of model. Assess your model's ability to predict and discuss its use as a tool for management decisions. (10 pts)
2. Refer to the Prostate cancer data set. Serum prostate-specific antigen (PSA) was determined in 97 men with advanced prostate cancer. PSA is a well-established screening test for prostate cancer and the oncologists wanted to examine the correlation between level of PSA and a number of clinical measures for men who were about to undergo radical prostatectomy. The measures are cancer volume, prostate weight, patient age, the amount of benign prostatic hyperplasia, seminal vesicle invasion, capsular penetration, and Gleason score. (15 Pts)
3. Select a random sample of 65 observations to use as the model-building data set. Develop a best subset model for predicting PSA. Justify your choice of model. Assess your model's ability to predict and discuss its usefulness to the oncologists. (5pts)
4. Fit the regression model identified in part a to the validation data set. Compare the estimated regression coefficients and their estimated standard errors with those obtained in part a. Also compare the error mean square and coefficients of multiple determination. Does the model fitted to the validation data set yield similar estimates as the model fitted to the model-building data set? (5pts)
5. Calculate the mean squared prediction error (equation 9.20 on the book) and compare it to MSE obtained from the model-building data set. Is there evidence of a substantial bias problem in MSE here? (5pts)
6. Refer to Market share data set. Company executives want to be able to predict market share of their product (Y) based on merchandise price (X1), the gross Nielsen rating points (X2, an index of the amount of advertising exposure that the product received); the presence or absence of a wholesale pricing discount (X3 = 1 if discount present: otherwise X3 = 0); the presence or absence of a package promotion during the period (X4 = 1 if promotion present: otherwise X4 = 0): and year (X5). Code year as a nominal level variable and use 2000 as the referent year. (20 pts)
7. Using only first-order terms for predictor variables, find the three best subset regression models according to the SECp criterion. (7 pts)
8. Using forward stepwise regression, find the best subset of predictor variables to predict market share of their product. Use α limits of 0.10 and .15 for adding or deleting a predictor, respectively. (7pts)
9. How does the best subset according to forward stepwise regression compare with the best subset according to the SECp criterion used in part a? (6pts)

