Dr. Sudhakar Raju

BIA 6309 – LINEAR & MULTIVARIATE MODELS

SUMMER 2018

**QUESTIONS FOR ASSIGNMENT 4**

I. The attached murder data (murder\_data) explains the murder rate (per 1000 persons) as a function of 7 predictors - population, illiteracy, income, frost, life expectancy, high school graduation rate and area.

a.) Run a regression subsets on the complete model with 7 predictors. What is the best 1 variable, 2 variable, 3 variable ….7 variable model?

b.) Which among these models has the highest adjusted R2? Report the complete regression results for this highest adjusted R2 model.

c.) Reformulate the model in part b above in terms of standardized variables (beta regression). Using this beta regression, rank the variables in terms of their predictive ability to explain the murder rate.

II. Consider the attached dataset (occupational\_prestige\_data.csv) on occupational prestige among a variety of occupations. Prestige variable refers to a prestige score for each occupation (given by a metric called Pineo-Porter). Prestige scores are linked to income, years of education (beyond middle school), percentage of women in the specific occupation and a binary variable that lists the occupation as either professional or not.

a.) Which occupation has the highest prestige? Lowest prestige? What is the mean prestige score? What are the mean values of the predictors?

b.) Run a linear regression explaining prestige as a function of education, income, percentage of women in the specific occupation and if the occupation is considered to be a professional occupation or not. Which variable has the greatest impact on prestige?

c.) Run a standardized regression on all predictors except the binary variable “professional”. Which variable has the greatest impact on prestige?

d.)Why was the binary variable “professional” not standardized?

e.) Run a regression explaining prestige as a function of the 3 standardized variables (income, education, percentage of women in occupation) and the unstandardized, binary variable “professional”. Interpret this regression.

III. The attached file (airfares\_data.csv) contains actual data on airfares that were collected between the third quarter of 1996 and second quarter of 1997. During this period competition increased dramatically due to airline deregulation. Low fare carriers such as Southwest began competing on existing routes and started nonstop service on new routes.

A new airline wants to enter this market and wants to better understand factors that drive average fares. The airline assembles the attached dataset. The variables in the dataset are described below. The goal is to predict average fare on a route (FARE).

1.       S\_CODE: starting airport’s code

2.       S\_CITY: starting city

3.       E\_CODE: ending airport’s code

4.       E\_CITY: ending city

5.       COUPON: average number of coupons (a one-coupon flight is a non-stop flight, a two-coupon flight is a one stop flight, etc.) for that route

6.       NEW: number of new carriers entering that route between Q3-96 and Q2-97

7.       VACATION: whether a vacation route (Yes) or not (No); Florida and Las Vegas routes are generally considered vacation routes

8.       SW: whether Southwest Airlines serves that route (Yes) or not (No)

9.       HI: Herfindel Index – measure of market concentration (refer to BMGT 681)

10.   S\_INCOME: starting city’s average personal income

11.   E\_INCOME: ending city’s average personal income

12.   S\_POP: starting city’s population

13.   E\_POP: ending city’s population

14.   SLOT: whether either endpoint airport is slot controlled or not; this is a measure of airport congestion

15.   GATE: whether either endpoint airport has gate constraints or not; this is another measure of airport congestion

16.   DISTANCE: distance between two endpoint airports in miles

17.   PAX: number of passengers on that route during period of data collection

18.   FARE: average fare on that route

a.) Create a correlation matrix of the numerical predictors. What variable seems to have the highest correlation with FARE? Does multicollinearity seem to be an issue?

b.) Run a full scale multiple linear regression explaining FARE as a function of all predictors (ignore S-CODE, S\_CITY, E\_CODE, E\_CITY).

c.) Run the regression subsets algorithm on 13 variables (excluding S-CODE, S\_CITY, E\_CODE, E\_CITY). How many possible models are we examining to determine the “best” model?

d.) What is the best 4 variable model? What are the coefficients and adjusted R2 of this model? Interpret the coefficients of this model.

e.) Predict airfare using the best 4 variable model for the following parameters: Not a vacation destination, SW operates on this route, average value of HI and average distance.