MA409: Statistical Data Analysis (SAS)

Assignment 3 (Apr 09 – Apr 30)

Note: Please work on 3.2 by hand calculation (p-values can be obtained with any software) and 3.3-3.4 by SAS procedures.

3.1 Show that the weighted least squares estimate defined by Eq. (5.17) in the lecture notes is the best linear unbiased estimate (BLUE) of $\boldsymbol{\beta}$. (10 points)

3.2 The tables below show the regression output of a multiple linear regression model relating the beginning salaries in dollars of employees in a given company to the following predictor variables:

- Gender An indicator variable (1=man and 0=woman)
- Education Years of schooling at the time of hire
- Experience Number of months of previous work experience
- Months Number of months with the company

ANOVA Table					
Source	DF	Sum of Squares	Mean Square	F value	P-value
Model	4	23665352			
Error	88	22657938			

Coefficients Table				
Parameter	Estimate	Standard Error	t value	P-value
Intercept	3526.4	327.7		
Gender	722.5	117.8		
Education	90.02	24.69		
Experience	1.2690	0.5877		
Months	23.406	5.201		

(1) Fill the two tables (keep at least four decimal points for F value, t value, and P-value), specify the degree of freedom used for the t-tests and state your conclusions of the F-test and t-tests at $\alpha = 0.05$. (10 points)

(2) Compute the R-squared and adjusted R-squared of the model. (5 points)

(3) What salary would you forecast, on average, for men with 12 years of education, 10 months of experience, and 15 months with the company? (5 points)

(4) Consider the model with all four predictor variables to be a full model and the model which only includes Education to be a reduced model. The ANOVA table obtained for the reduced model is given below. Conduct a test to compare the full and reduced model.

(5 points)

ANOVA Table for Reduced Model					
Source	DF	Sum of Squares	Mean Square	F value	P-value
Model	1	7862535	7862535	18.6031	8.1538E-5
Error	91	38460756	422645.7		

3.3 The Education Expenditure data is provided in "EducationExpenditure.xlsx". It contains the following variables for the 50 states in the US measured in 1975:

- Y: Per capita expenditure on public education
- X_1: Per capita personal income
- X_2: Number of residents per thousand under 18 years of age
- X_3: Number of people per thousand residing in urban areas

(1) Fit the regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, check if the assumptions for linear regression model are violated. (10 points)

(2) Filter out observations with high leverage and observations that are outliers or influential observations based on the model in (1). Is there any unusual observation that you would like to drop from the analysis? Provide your justification. (5 points)

(3) Refit the regression model in (1) by correcting the violation(s) in (1) if any and by dropping the unusual observation(s) determined in (2) if any. Compare the regression coefficients with those from (1) and present your findings. (15 points)

3.4 "AirPolution.xlsx" provides data from a study that relates total mortality to climate, socioeconomics, and pollution variables for 60 US cities. A response variable and 15 predictor variables are included:

Y: Total age-adjusted mortality rate per 100,000	X_8: Population per square mile
X_1: Mean annual precipitation (inches)	X_9: Percent of nonwhite population
X_2: Mean January temperature (degrees Fahrenheit)	X_{10}: Percent employment in white-collar jobs
X_3: Mean July temperature (degrees Fahrenheit)	X_{11}: Percent of families with income under $3000
X_4: Percent of population over 65 years of age	X_{12}: Relative pollution potential of hydrocarbons
X_5: Average household size	X_{13}: Relative pollution potential of oxides of nitrogen
X_6: Median school years completed	X_{14}: Relative pollution potential of sulfur dioxide
X_7: Percent of housing units that are sound	X_{15}: Percent relative humidity

(1) Check the pairwise Pearson correlation coefficients using PROC CORR. Is there any collinearity between pairs of variables? (5 points)

(2) Fit the regression model of Y on all predictor variables. Does multicollinearity exist? (5 points)

(3) Use PROC GLMSELECT to perform stepwise variable selection, specifically, apply the BIC criterion (i.e., SBC in PROC GLMSELECT) to determine the order in which

variables enter or leave at each step, as well as to select the best model. Display the adjusted R-squared, Mallows's C_p, AIC and BIC values at each step and generate a plot of these criteria by step. State your final model and explain your conclusions. (10 points)

(4) Use PROC REG to fit the ridge regressions of Y on all predictor variables with 21 equally spaced values of λ in the interval $[0, 1]$. Output the parameter estimates under ridge regression models with different λ to a SAS dataset, then plot the lines showing the parameter estimates of X1, X2, X6, X9, X12, X13, X14 against λ, and state your findings. Note: make sure to standardize the response and predictor variables before fitting the ridge regression models (standardization can be performed using PROC STANDARD). (15 points)