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)

Solution: For the weighted least squares estimate defined by Eq. (5.17), we have:

$$\widehat{\boldsymbol{\beta}}^{WLS} = (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{y},$$

and (as $(Var(\boldsymbol{y}) = \boldsymbol{W}^{-1} = \mathrm{diag}\{\sigma_i^2\})$)

$$Var(\widehat{\boldsymbol{\beta}}^{WLS}) = [(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}]Var(\boldsymbol{y})[(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}]^\top$$

$$= [(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}]\boldsymbol{W}^{-1}[\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}] = (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}$$

Let $\widetilde{\boldsymbol{\beta}} = \boldsymbol{C}\boldsymbol{y}$ be another linear unbiased estimate of $\boldsymbol{\beta}$. Let $\boldsymbol{C} = (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W} + \boldsymbol{D}$. Consider $E(\widetilde{\boldsymbol{\beta}})$:

$$E(\widetilde{\boldsymbol{\beta}}) = E(\boldsymbol{C}\boldsymbol{y}) = E[((\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W} + \boldsymbol{D})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] = (\mathbf{I}_n + \boldsymbol{D}\boldsymbol{X})\boldsymbol{\beta}$$

Therefore, $\widetilde{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$ if and only if $\boldsymbol{D}\boldsymbol{X} = 0$. Then:

$$Var(\widetilde{\boldsymbol{\beta}}) = Var(\boldsymbol{C}\boldsymbol{y}) = \boldsymbol{C}Var(\boldsymbol{y})\boldsymbol{C}^\top = [(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W} + \boldsymbol{D}]\boldsymbol{W}^{-1}[\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1} + \boldsymbol{D}^\top]$$

$$= (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1} + \boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{D}^\top + \boldsymbol{D}\boldsymbol{W}^{-1}\boldsymbol{D}^\top$$

$$= (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1} + \boldsymbol{D}\boldsymbol{W}^{-1}\boldsymbol{D}^\top = Var(\widehat{\boldsymbol{\beta}}^{WLS}) + \boldsymbol{D}\boldsymbol{W}^{-1}\boldsymbol{D}^\top.$$

As $\boldsymbol{W}^{-1} = \mathrm{diag}\{\sigma_i^2\}$, it is not difficult to show that $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{W}^{-1}\boldsymbol{D}^\top$ is a positive semidefinite matrix. Therefore, $Var(\widetilde{\boldsymbol{\beta}}) \geq Var(\widehat{\boldsymbol{\beta}}^{WLS})$ for any linear unbiased estimate $\widetilde{\boldsymbol{\beta}}$, i.e., $\widehat{\boldsymbol{\beta}}^{WLS}$ is the best linear unbiased estimate of $\boldsymbol{\beta}$.

3.2 The tables below show the regression output of a multiple linear regression model relating the 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	5916338	22.9782	5.0715E-13
Error	88	22657938	257476.6		

Coefficients Table

Parameter	Estimate	Standard Error	t value	P-value
Intercept	3526.4	327.7	10.7610	0.0000
Gender	722.5	117.8	6.1333	2.3917E-8
Education	90.02	24.69	3.6460	4.5033E-4
Experience	1.2690	0.5877	2.1593	0.0335
Months	23.406	5.201	4.5003	2.0675E-5

(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 F-test and t-tests and state your conclusions of the F-test and t-tests at $\alpha = 0.05$. (10 points)

Solution: The values filled are displayed in the two tables above. The linear model is

$$\text{Salary} = \beta_0 + \beta_1 \text{I}(\text{Gender} = \text{male}) + \beta_2 \text{Education} + \beta_3 \text{Experience} + \beta_4 \text{Months} + \varepsilon.$$

The F-test p-value is computed based on $F(4, 88)$ (one-sided probability) and the t-test p-values are computed based on t_{88} (two-sided probability). The overall significance of the model (F-test) is rejected at $\alpha = 0.05$ indicating that at least one of the regression coefficients in the model is significantly different from zero. Each individual t-test show that each of the four predictor variables is statistically significant, specifically:

- On average, salaries for men are 722.5 dollars higher than those for women with the same Education, Experience, and Months.
- On average, after accounting for the effect of Gender, Experience, and Months, salaries for employees with one more year of schooling are 90.02 dollars higher.
- On average, after accounting for the effect of Gender, Education, and Months, salaries for employees with one more month of previous work experience are 1.2690 dollars higher.
- On average, after accounting for the effect of Gender, Education, and Experience, salaries for employees with one more month with the company are 23.406 dollars higher.

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

Solution:

$$R^2 = \frac{SSM}{SST} = \frac{SSM}{SSM + SSE} = \frac{23665352}{23665352 + 22657938} = 0.5109$$

$$\text{Adjusted } R^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - \frac{22657938/88}{(23665352 + 22657938)/92} = 0.4886$$

(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)

Solution:

$$\begin{aligned}\text{Predicted Salary for men} \\ &= 3526.4 + 722.5 \times 1 + 90.02 \times 12 + 1.2690 \times 10 + 23.406 \times 15 \\ &= 5692.92 \text{ dolloars}\end{aligned}$$

(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	4.0769E-5
Error	91	38460756	422645.7		

Solution: The F-test statistic to compare the full and reduced model is

$$F = \frac{(RSS_R - RSS_F)/(p-k)}{RSS_F/(n-p-1)} = \frac{(38460756 - 22657938)/(4-1)}{22657938/88} = 20.4586$$

The p-value computed based on $F(3, 88)$ (one-sided probability) is $3.8052E - 10$.

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:

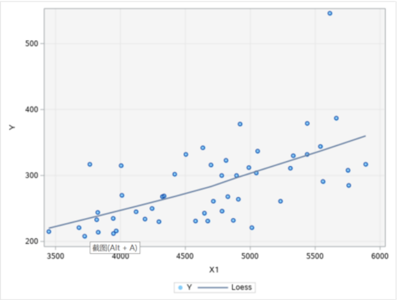
- Y: Per capita expenditure on public education
- X_1: Per capita personal income
- X_2: Number of residents per thousand under 18 years of age
- X_3: Number of people per thousand residing in urban areas

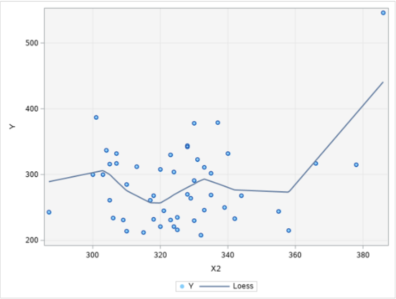
(1) Fit the regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, check if the assumptions for linear regression model are violated. (10 points)

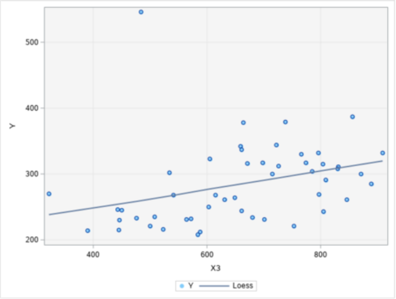
Solution:

a. For the independence assumption, as the data were collected for each state in the US, it is reasonable to assume independency.

b. For the linearity assumption, we plot the response variable Y vs. each of the explanatory variables

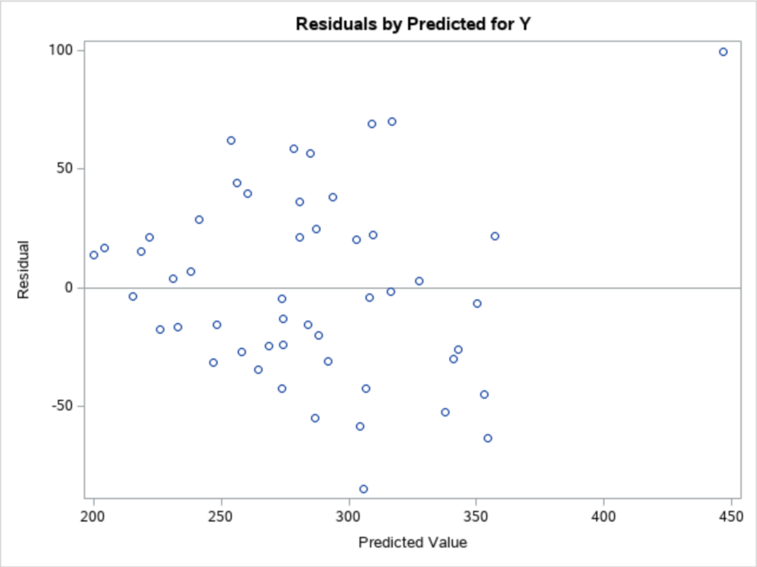






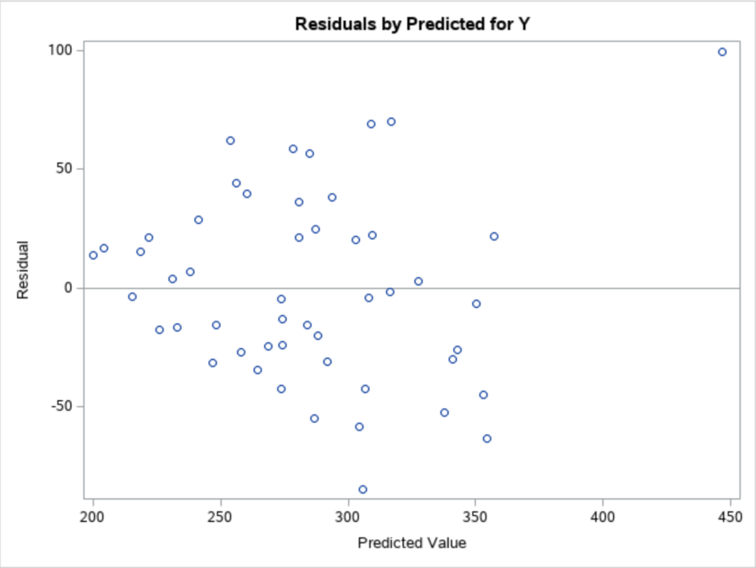
As shown in the plots, it seems that Y is linearly associated with X1 and X3. The association between Y and X2 seems to be not quite linear, however, it may due to the datapoint with large values in Y and X2 (on the upper right corner).

We also check the fitted versus residuals plot:

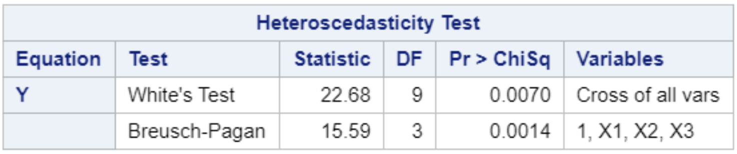


The plot shows that the mean of the residuals is roughly 0 at any fitted value, indicating that the linearity assumption is valid.

c. For the homoscedasticity assumption, we again look at the fitted versus residuals plot:



The plot shows that the spread of the residuals is greater for larger fitted values, indicating the violation of the homogeneity assumption. We can also check for heteroscedasticity with the White and the Breusch-Pagan tests:

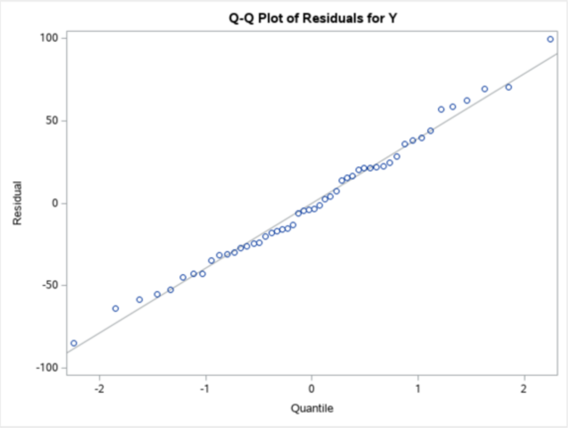


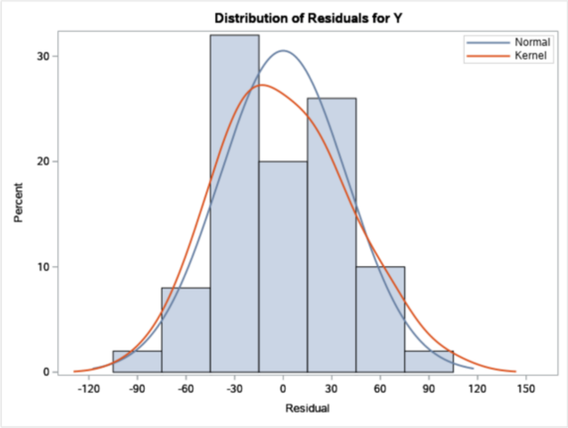
Heteroscedasticity Test

Equation	Test	Statistic	DF	Pr > ChiSq	Variables
Y	White's Test	22.68	9	0.0070	Cross of all vars
	Breusch-Pagan	15.59	3	0.0014	1, X1, X2, X3

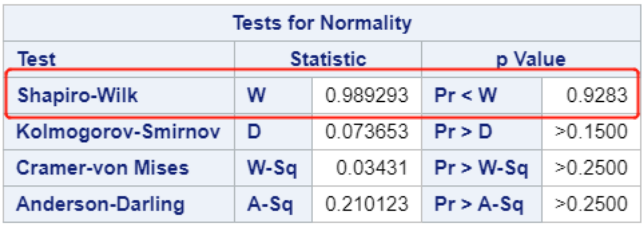
The p-values of both tests are < 0.01, indicating that the homoscedasticity assumption is rejected.

d. For the normality assumption, we checked the Q-Q plot as well as the histogram of the residuals. Both plots do not show violation of the normality assumption.





We can also test for normality with the Shapiro-Wilk test (as sample size is < 2000), the p-value is 0.9283, indicating that the normality assumption is not rejected.

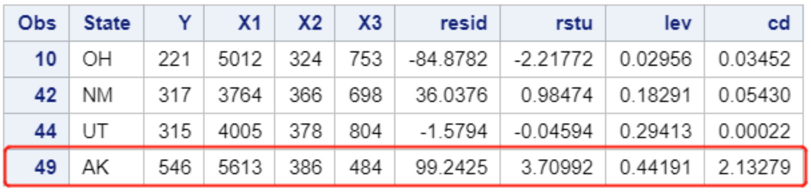


Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.989293	Pr < W	0.9283
Kolmogorov-Smirnov	D	0.073653	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.03431	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.210123	Pr > A-Sq	>0.2500

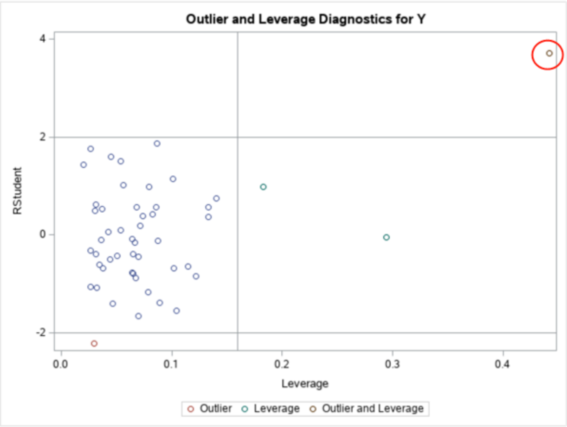
(2) Filter out observations with high leverage and observations that are outliers or influential observations based on the model in (1). Is there any unusual observation that you would like to drop from the analysis? Provide your justification. (5 points)

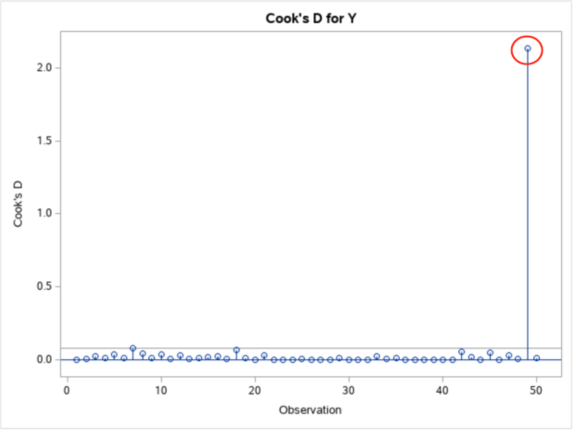
Solution: The following four observations are obtained by filtering out the observations with high leverage and observations that are outliers or influential observations:



Obs	State	Y	X1	X2	X3	resid	rstu	lev	cd
10	OH	221	5012	324	753	-84.8782	-2.21772	0.02956	0.03452
42	NM	317	3764	366	698	36.0376	0.98474	0.18291	0.05430
44	UT	315	4005	378	804	-1.5794	-0.04594	0.29413	0.00022
49	AK	546	5613	386	484	99.2425	3.70992	0.44191	2.13279

The 49th observation has high leverage, large studentized residual, and large Cook's distance, as also shown in the following diagnostic plots:

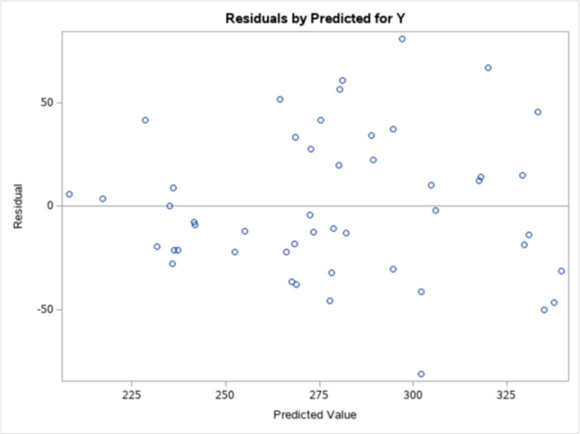


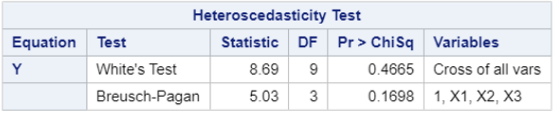


The 49th observation is the datapoint for Alaska, which is a state not connected to the mainland of the US. Therefore, it may be considered an outlier and removed from the analysis.

(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)

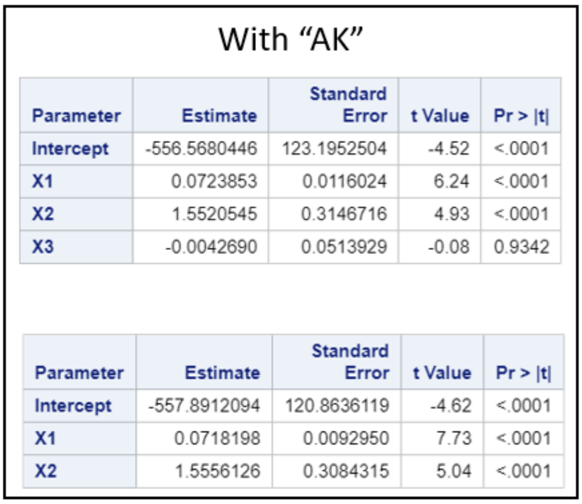
Solution: Drop the observation for “AK” and refit the model in (1), we check the homoscedasticity assumption again:





Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
Y	White's Test	8.69	9	0.4665	Cross of all vars
	Breusch-Pagan	5.03	3	0.1698	1, X1, X2, X3

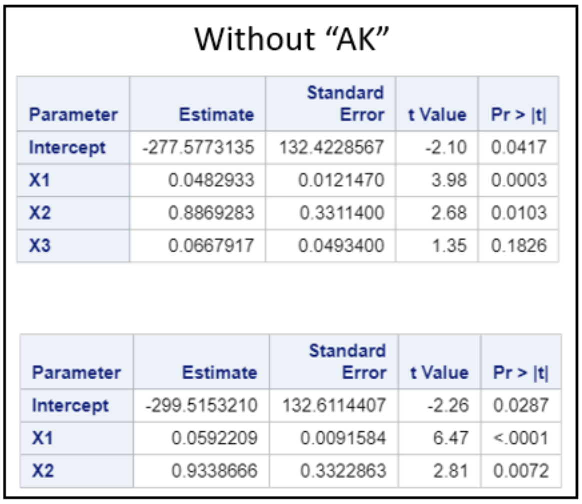
The results indicate that the homoscedasticity assumption is not seriously violated. Comparing the regression coefficients with and without the observation for "AK", we observe that β_0 and β_2 are very different, confirming that the observation for "AK" is indeed an influential observation.



With "AK"

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-556.5680446	123.1952504	-4.52	<.0001
X1	0.0723853	0.0116024	6.24	<.0001
X2	1.5520545	0.3146716	4.93	<.0001
X3	-0.0042690	0.0513929	-0.08	0.9342

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-557.8912094	120.8636119	-4.62	<.0001
X1	0.0718198	0.0092950	7.73	<.0001
X2	1.5556126	0.3084315	5.04	<.0001



Without "AK"

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-277.5773135	132.4228567	-2.10	0.0417
X1	0.0482933	0.0121470	3.98	0.0003
X2	0.8869283	0.3311400	2.68	0.0103
X3	0.0667917	0.0493400	1.35	0.1826

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-299.5153210	132.6114407	-2.26	0.0287
X1	0.0592209	0.0091584	6.47	<.0001
X2	0.9338666	0.3322863	2.81	0.0072

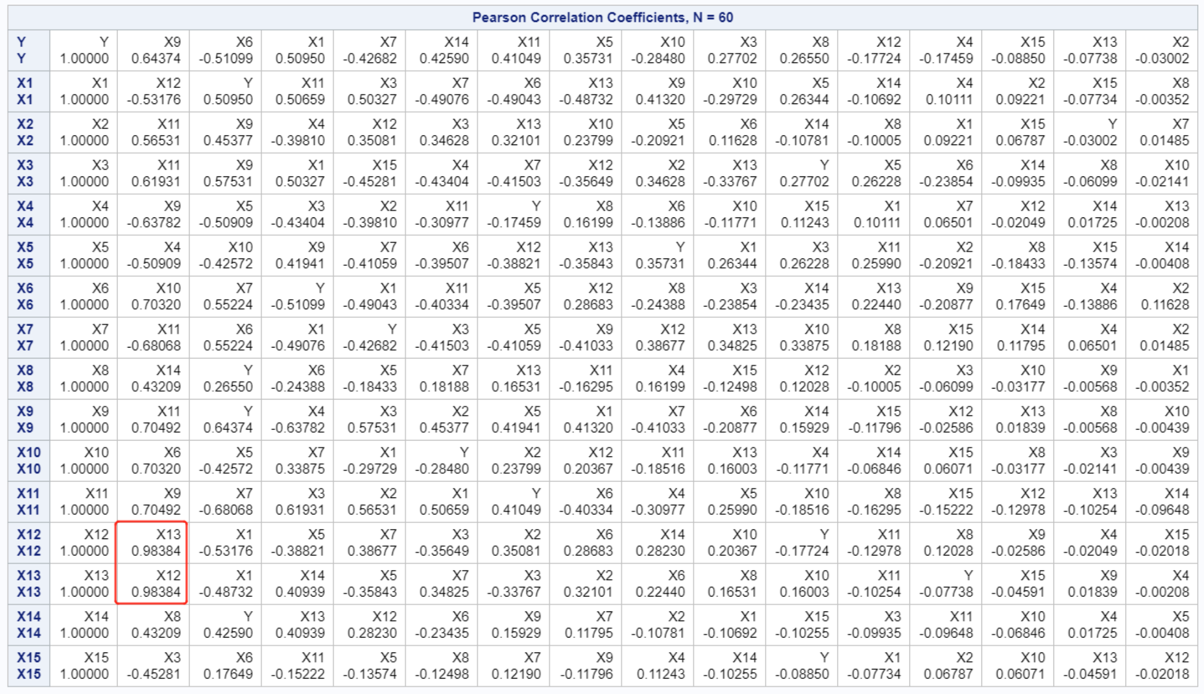
3.4 "AirPolution.xlsx" provides data from a study that relates total mortality to climate, socioeconomics, and pollution variables for 60 US cities. A response variable and 15 predictor variables are included:

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)

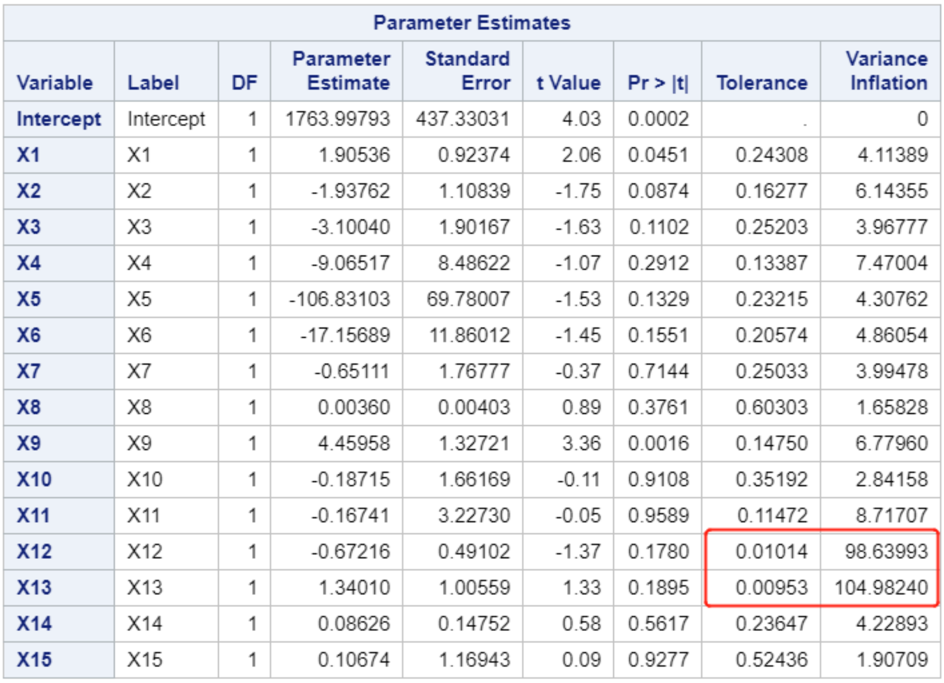
Solution: Obtain the pairwise Pearson correlation coefficients and rank them in descending order by the absolute value. The Pearson correlation coefficients show that X12 and X13 are highly linearly correlated (correlation coefficient = 0.98384).



Pearson Correlation Coefficients, N = 60																
Y Y	Y 1.00000	X9 0.64374	X6 -0.51099	X1 0.50950	X7 -0.42682	X14 0.42590	X11 0.41049	X5 0.35731	X10 -0.28480	X3 0.27702	X8 0.26550	X12 -0.17724	X4 -0.17459	X15 -0.08850	X13 -0.07738	X2 -0.03002
X1 X1	X1 1.00000	X12 -0.53176	Y 0.50950	X11 0.50659	X3 0.50327	X7 -0.49076	X6 -0.49043	X13 -0.48732	X9 0.41320	X10 -0.29729	X5 0.26344	X14 -0.10692	X4 0.10111	X2 0.09221	X15 -0.07734	X8 -0.00352
X2 X2	X2 1.00000	X11 0.56531	X9 0.45377	X4 -0.39810	X12 0.35081	X3 0.34628	X13 0.32101	X10 0.23799	X5 -0.20921	X6 0.11628	X14 -0.10781	X8 -0.10005	X1 0.09221	X15 0.06787	Y -0.03002	X7 0.01485
X3 X3	X3 1.00000	X11 0.61931	X9 0.57531	X1 0.50327	X15 -0.45281	X4 -0.43404	X7 -0.41503	X12 -0.35649	X2 0.34628	X13 -0.33767	Y 0.27702	X5 0.26228	X6 -0.23854	X14 -0.09935	X8 -0.06099	X10 -0.02141
X4 X4	X4 1.00000	X9 -0.63782	X5 -0.50909	X3 -0.43404	X2 -0.39810	X11 -0.30977	Y -0.17459	X8 0.16199	X6 -0.13886	X10 -0.11771	X15 0.11243	X1 0.10111	X7 0.06501	X12 -0.02049	X14 0.01725	X13 -0.00208
X5 X5	X5 1.00000	X4 -0.50909	X10 -0.42572	X9 0.41941	X7 -0.41059	X6 -0.39507	X12 -0.38821	X13 -0.35843	Y 0.35731	X1 0.26344	X3 0.26228	X11 0.25990	X2 -0.20921	X8 -0.18433	X15 -0.13574	X14 -0.00408
X6 X6	X6 1.00000	X10 0.70320	X7 0.55224	Y -0.51099	X1 -0.49043	X11 -0.40334	X5 -0.39507	X12 0.28683	X8 -0.24388	X3 -0.23854	X14 -0.23435	X13 0.22440	X9 -0.20877	X15 0.17649	X4 -0.13886	X2 0.11628
X7 X7	X7 1.00000	X11 -0.68068	X6 0.55224	X1 -0.49076	Y -0.42682	X3 -0.41503	X5 -0.41059	X9 -0.41033	X12 0.38677	X13 0.34825	X10 0.33875	X8 0.18188	X15 0.12190	X14 0.11795	X4 0.06501	X2 0.01485
X8 X8	X8 1.00000	X14 0.43209	Y 0.26550	X6 -0.24388	X5 -0.18433	X7 0.18188	X13 0.16531	X11 -0.16295	X4 0.16199	X15 -0.12498	X12 0.12028	X2 -0.10005	X3 -0.06099	X10 -0.03177	X9 -0.00568	X1 -0.00352
X9 X9	X9 1.00000	X11 0.70492	Y 0.64374	X4 -0.63782	X3 0.57531	X2 0.45377	X5 0.41941	X1 0.41320	X7 -0.41033	X6 -0.20877	X14 0.15929	X15 -0.11796	X12 -0.02586	X13 0.01839	X8 -0.00568	X10 -0.00439
X10 X10	X10 1.00000	X6 0.70320	X5 -0.42572	X7 0.33875	X1 -0.29729	Y -0.28480	X2 0.23799	X12 0.20367	X11 -0.18516	X13 0.16003	X4 -0.11771	X14 -0.06846	X15 0.06071	X8 -0.03177	X3 -0.02141	X9 -0.00439
X11 X11	X11 1.00000	X9 0.70492	X7 -0.68068	X3 0.61931	X2 0.56531	X1 0.50659	Y 0.41049	X6 -0.40334	X4 -0.30977	X5 0.25990	X10 -0.18516	X8 -0.16295	X15 -0.15222	X12 -0.12978	X13 -0.10254	X14 -0.09648
X12 X12	X12 1.00000	X13 0.98384	X1 -0.53176	X5 -0.38821	X7 0.38677	X3 -0.35649	X2 0.35081	X6 0.28683	X14 0.28230	X10 0.20367	Y -0.17724	X11 -0.12978	X8 0.12028	X9 -0.02586	X4 -0.02049	X15 -0.02018
X13 X13	X13 1.00000	X12 0.98384	X1 -0.48732	X14 0.40939	X5 -0.35843	X7 0.34825	X3 -0.33767	X2 0.32101	X6 0.22440	X8 0.16531	X10 0.16003	X11 -0.10254	Y -0.07738	X15 -0.04591	X9 0.01839	X4 -0.00208
X14 X14	X14 1.00000	X8 0.43209	Y 0.42590	X13 0.40939	X12 0.28230	X6 -0.23435	X9 0.15929	X7 0.11795	X2 -0.10781	X1 -0.10692	X15 -0.10255	X3 -0.09935	X11 -0.09648	X10 -0.06846	X4 0.01725	X5 -0.00408
X15 X15	X15 1.00000	X3 -0.45281	X6 0.17649	X11 -0.15222	X5 -0.13574	X8 -0.12498	X7 0.12190	X9 -0.11796	X4 0.11243	X14 -0.10255	Y -0.08850	X1 -0.07734	X2 0.06787	X10 0.06071	X13 -0.04591	X12 -0.02018

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

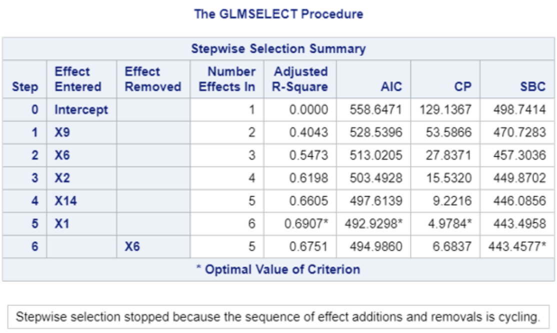
Solution: By checking the tolerance or the variance inflation factor (tolerance < 0.1 or VIF > 10), multicollinearity exists for X12 and X13.



Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Tolerance**	**Variance Inflation**
Intercept	Intercept	1	1763.99793	437.33031	4.03	0.0002	.	0
X1	X1	1	1.90536	0.92374	2.06	0.0451	0.24308	4.11389
X2	X2	1	-1.93762	1.10839	-1.75	0.0874	0.16277	6.14355
X3	X3	1	-3.10040	1.90167	-1.63	0.1102	0.25203	3.96777
X4	X4	1	-9.06517	8.48622	-1.07	0.2912	0.13387	7.47004
X5	X5	1	-106.83103	69.78007	-1.53	0.1329	0.23215	4.30762
X6	X6	1	-17.15689	11.86012	-1.45	0.1551	0.20574	4.86054
X7	X7	1	-0.65111	1.76777	-0.37	0.7144	0.25033	3.99478
X8	X8	1	0.00360	0.00403	0.89	0.3761	0.60303	1.65828
X9	X9	1	4.45958	1.32721	3.36	0.0016	0.14750	6.77960
X10	X10	1	-0.18715	1.66169	-0.11	0.9108	0.35192	2.84158
X11	X11	1	-0.16741	3.22730	-0.05	0.9589	0.11472	8.71707
X12	X12	1	-0.67216	0.49102	-1.37	0.1780	0.01014	98.63993
X13	X13	1	1.34010	1.00559	1.33	0.1895	0.00953	104.98240
X14	X14	1	0.08626	0.14752	0.58	0.5617	0.23647	4.22893
X15	X15	1	0.10674	1.16943	0.09	0.9277	0.52436	1.90709

(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)

Solution: The model selection results are presented below, X9, X6, X2, X14, X1 are added sequentially to the model, then X6 is removed from the model.



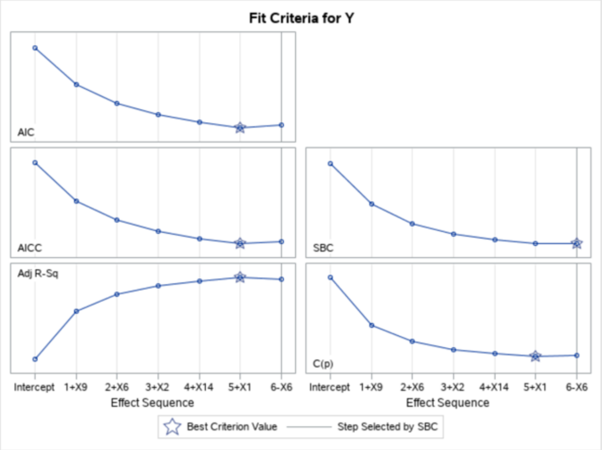
The GLMSELECT Procedure

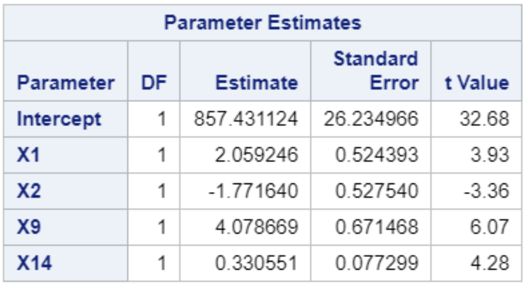
Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Adjusted R-Square	AIC	CP	SBC
0	Intercept		1	0.0000	558.6471	129.1367	498.7414
1	X9		2	0.4043	528.5396	53.5866	470.7283
2	X6		3	0.5473	513.0205	27.8371	457.3036
3	X2		4	0.6198	503.4928	15.5320	449.8702
4	X14		5	0.6605	497.6139	9.2216	446.0856
5	X1		6	0.6907*	492.9298*	4.9784*	443.4958
6		X6	5	0.6751	494.9860	6.6837	443.4577*

* Optimal Value of Criterion

Stepwise selection stopped because the sequence of effect additions and removals is cycling.





Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	857.431124	26.234966	32.68
X1	1	2.059246	0.524393	3.93
X2	1	-1.771640	0.527540	-3.36
X9	1	4.078669	0.671468	6.07
X14	1	0.330551	0.077299	4.28

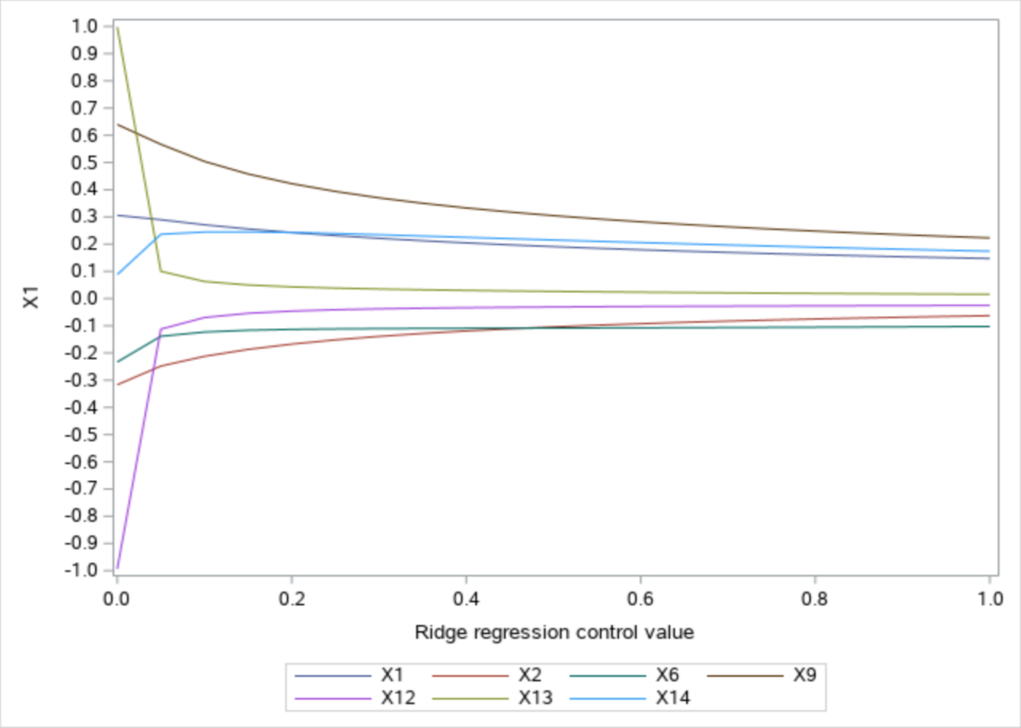
The final model applying the model selection procedure is

$$Y = 857.43 + 2.06 * X_1 - 1.77 * X_2 + 4.08 * X_9 + 0.33 * X_{14} + \varepsilon$$

- On average, after accounting for the effect of the other three variables, the total age-adjusted mortality rate per 100,000 for a city with 1 more inch of mean annual precipitation is 2.06 higher.
- On average, after accounting for the effect of the other three variables, the total age-adjusted mortality rate per 100,000 for a city with the mean January temperature being 1 degree Fahrenheit higher is 1.77 lower.
- On average, after accounting for the effect of the other three variables , the total age-adjusted mortality rate per 100,000 for a city with 1% more nonwhite population is 4.08 higher.
- On average, after accounting for the effect of the other three variables, the total age-adjusted mortality rate per 100,000 for a city with 1 unit greater relative pollution potential of sulfur dioxide is 0.33 higher.

(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)

Solution: The parameter estimates of X1, X2, X6, X9, X12, X13, X14 against λ are given below:



We observe from the plot that:

- When λ increases by a small amount, the coefficients of X12 (pollution potential of hydrocarbons) and X13 (pollution potential of oxides of nitrogen) decrease rapidly in absolute value. This is not unexpected because X12 and X13 have a sample correlation coefficient of 0.98 as given in (1). The coefficients of X12 and X13 are quickly driven towards zero and are almost mirror images of each other about the zero line.
- The effects of X1 (mean annual precipitation), X2 (mean January temperature), X6 (median school years completed) and X9 (percent non-white population) also appear to be overestimated in absolute value. The coefficients decrease in absolute value as λ increases, and level off at non-zero values.
- The effect of X14 (pollution potential of sulfur dioxide), is likely to be originally underestimated. The coefficient increases as λ increases.
- The coefficients appear to stabilize in the neighborhood of $\lambda = 0.2$. We expect coefficients estimated at $\lambda = 0.2$ to be more suitable for estimation of the effects of the explanatory variables.