APM 630 Regression Analysis
HW8 – Multicollinearity
Name: Ruta Basijokaite

Project Report

1. Introduction

The data was obtained from the US Bureau of the Census State Metropolitan Area Data Book 1986 on a set of 53 primary metropolitan areas (PMSA). The variables in the study included:

(1) Dependent variable:
• CRIME – total number of serious crimes in 1980

(2) Predictor variables:
• POP – total population in 1980, in thousands
• AREA – land area in square miles
• YOUNG – 1980 population ages 18-24 in thousands
• DIV – total of divorces in 1980 in thousands
• OLD – 1980 total number of SSN benefit recipients in thousands
• EDUC – number of adults (>25 years old) having completed 12 or more years of school
• POV – total number of people below poverty level in 1980
• UNEMP – total number of unemployed in 1980

The purpose of this study was to (1) compute descriptive statistics for all variables, (2) compute correlation coefficients among all variables, (3) plot matrix scatterplot of dependent variable and five predictor variables, (4) fit full model to the data, (5) conduct multicollinearity diagnostics on the study variables to identify the problems of multicollinearity among the independent variables, (6) conduct ridge regression for the model and identify the best "k" value and corresponding ridge regression coefficients, (7) conduct principal component regression omitting 1 to 6 variables to obtain the transformed coefficients for the X variables, (8) identify "best" PC regression model, (9) compare the OLS model with "best" ridge regression and PC regression model in terms of similarities of model coefficients.

2. Methods

53 data points were collected to determine the factors related to the number of serious crimes in 1980. Descriptive statistics for all variables are summarized in Table 1.

Table 1. Summary of descriptive statistics of all variables

Variable	N	Mean	Median	Std Dev	Minimum	Maximum
CRIME	53	97.3736	42.6000	143.4149	4.3000	768.3000
POP	53	1389.5849	849.0000	1774.6771	123.0000	8275.0000
AREA	53	1522.6415	1146.0000	1259.3148	46.0000	5345.0000
YOUNG	53	177.0415	91.7000	224.2673	15.0000	1024.3000
DIV	53	6.4736	3.5000	7.4240	0.5000	33.6000
OLD	53	195.3755	109.3000	255.1833	14.5000	1306.6000
EDUC	53	963.4302	596.8000	1178.9822	70.9000	5218.9000
POV	53	152.5094	65.0000	259.0470	8.0000	1471.0000
UNEMP	53	100.0057	56.8000	143.0966	7.8000	662.0000

Relationship between dependent variable CRIME and eight predictor variables are summarized in Table 2 and illustrated in Figure 1.

Table 2. Pearson (top) and Spearman (bottom) correlation coefficients for all variables

Pearson Correlation Coefficients, N = 53
Prob > |r| under H0: Rho=0

	CRIME	POP	AREA	YOUNG	DIV	OLD	EDUC	POV	UNEMP
CRIME	1.00000	0.96771 <.0001	0.45175 0.0007	0.95871 <.0001	0.94466 <.0001	0.94884 <.0001	0.96032 <.0001	0.97961 <.0001	0.92737 <.0001
POP	0.96771 <.0001	1.00000	0.50880 0.0001	0.99583 <.0001	0.96110 <.0001	0.98072 <.0001	0.99689 <.0001	0.96439 <.0001	0.95331 <.0001
AREA	0.45175 0.0007	0.50880 0.0001	1.00000	0.53678 <.0001	0.64349 <.0001	0.43484 0.0011	0.53148 <.0001	0.38707 0.0042	0.45640 0.0006
YOUNG	0.95871 <.0001	0.99583 <.0001	0.53678 <.0001	1.00000	0.96927 <.0001	0.96244 <.0001	0.99684 <.0001	0.94954 <.0001	0.94060 <.0001
DIV	0.94466 <.0001	0.96110 <.0001	0.64349 <.0001	0.96927 <.0001	1.00000	0.90707 <.0001	0.96799 <.0001	0.90969 <.0001	0.88478 <.0001
OLD	0.94884 <.0001	0.98072 <.0001	0.43484 0.0011	0.96244 <.0001	0.90707 <.0001	1.00000	0.96874 <.0001	0.96317 <.0001	0.95824 <.0001
EDUC	0.96032 <.0001	0.99689 <.0001	0.53148 <.0001	0.99684 <.0001	0.96799 <.0001	0.96874 <.0001	1.00000	0.94603 <.0001	0.94021 <.0001
POV	0.97961 <.0001	0.96439 <.0001	0.38707 0.0042	0.94954 <.0001	0.90969 <.0001	0.96317 <.0001	0.94603 <.0001	1.00000	0.92664 <.0001
UNEMP	0.92737 <.0001	0.95331 <.0001	0.45640 0.0006	0.94060 <.0001	0.88478 <.0001	0.95824 <.0001	0.94021 <.0001	0.92664 <.0001	1.00000

Spearman Correlation Coefficients, N = 53
Prob > |r| under H0: Rho=0

	CRIME	POP	AREA	YOUNG	DIV	OLD	EDUC	POV	UNEMP
CRIME	1.00000	0.96807 <.0001	0.67634 <.0001	0.96968 <.0001	0.97036 <.0001	0.92872 <.0001	0.97057 <.0001	0.95986 <.0001	0.92614 <.0001
POP	0.96807 <.0001	1.00000	0.68747 <.0001	0.99028 <.0001	0.96406 <.0001	0.96122 <.0001	0.99460 <.0001	0.96934 <.0001	0.96912 <.0001
AREA	0.67634 <.0001	0.68747 <.0001	1.00000	0.68681 <.0001	0.75421 <.0001	0.62514 <.0001	0.69569 <.0001	0.68628 <.0001	0.64033 <.0001
YOUNG	0.96968 <.0001	0.99028 <.0001	0.68681 <.0001	1.00000	0.96818 <.0001	0.93823 <.0001	0.99137 <.0001	0.96466 <.0001	0.95839 <.0001
DIV	0.97036 <.0001	0.96406 <.0001	0.75421 <.0001	0.96818 <.0001	1.00000	0.91389 <.0001	0.97237 <.0001	0.94706 <.0001	0.91732 <.0001
OLD	0.92872 <.0001	0.96122 <.0001	0.62514 <.0001	0.93823 <.0001	0.91389 <.0001	1.00000	0.95219 <.0001	0.95123 <.0001	0.96404 <.0001
EDUC	0.97057 <.0001	0.99460 <.0001	0.69569 <.0001	0.99137 <.0001	0.97237 <.0001	0.95219 <.0001	1.00000	0.95974 <.0001	0.95658 <.0001
POV	0.95986 <.0001	0.96934 <.0001	0.68628 <.0001	0.96466 <.0001	0.94706 <.0001	0.95123 <.0001	0.95974 <.0001	1.00000	0.96214 <.0001
UNEMP	0.92614 <.0001	0.96912 <.0001	0.64033 <.0001	0.95839 <.0001	0.91732 <.0001	0.96404 <.0001	0.95658 <.0001	0.96214 <.0001	1.00000

Figure 1. Matrix scatterplot of dependent variable and eight predictor variables

Simple Ordinary Least Squares (OLS) model was created using Statistical Analysis System (SAS) to fit the data represented in the section 1 of this report. Model was developed based on the following linear regression formula:

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \beta_4 X4 + \beta_5 X5 + \beta_6 X6 + \beta_7 X7 + \beta_8 X8 + \varepsilon \quad [1]$$

where Y is total number of serious crimes in 1980, $\beta_0 - \beta_8$ are regression coefficients to be estimated, X1 – X8 are predictor variables described above, and ε is the model error.

The OLS regression method is based on the assumption that the errors are additive, normally distributed, and independent with a mean zero and common variance σ^2. However, that is not always the case. For example, when a predictor variable is nearly a linear combination of other predictor variables in the model, the affected estimates of the regression coefficients are unstable and have high standard errors. This problem is called multicollinearity and it can possibly

have serious influence on the model output such as large changes in the estimated regression coefficients, variations in extra sum of squares associated with an independent variable, non-significant results in individual tests on the regression coefficients, wrongful algebraic sign of estimated regression coefficients, large correlation coefficients between pairs of independent variables in the correlation matrix, wide confidence intervals for the regression coefficients.

One of the methods that can be used to address the multicollinearity of the model is ridge regression. It builds on the fact that a singular square matrix ca be made nonsingular by adding a constant to the diagonal of the matrix. That is, if $X'X$ is singular, the $(X'X + kI)$ is nonsingular, where k is some small positive constant. This makes the off-diagonal elements appear relatively less important and, in effect, suppresses the near-singularities among the independent variables. The goal is to choose k to minimize the mean squared error. The choice of k is somewhat subjective, but usually k is very close to 0 (0.1 or even smaller). If k=0, the ridge solution is the OLS solution. The choice of k is a compromise between reducing variance (reduced with increasing k) and increasing bias (increased with increasing k). Therefore, the strategy is to choose (1) the smallest k that appears to be producing stable estimates of the regression coefficients, (2) get VIF close to 1, and (3) look for only "modest" increase in model RMSE.

Another method that is used to address multicollinearity is Principal Component (PC) regression. It approaches multicollinearity problem by eliminating those dimensions of the X-space that are causing the multicollinearity. This is similar, in concept, to dropping an independent variable from the model where there is insufficient dispersion in that variable to contribute meaningful information on Y. However, in PC regression, the dimension dropped is defined by a linear combination of the variables rather than by a single independent variable.

3. Results and Discussion

3.1 Full Model

Equation 1 was edited to fit the data using least-square method. Model can be represented using the following equation:

$$\hat{Y} = 2.32969 - 0.22532*X1 - 0.00248*X2 - 0.28914*X3 + 7.32687*X4 - 0.00231*X5 + \\ +0.29766*X6 + 0.64182*X7 + 0.31470*X8 \quad [2]$$

Model produced R^2 value that was equal to 0.9897 indicating that almost 99% of the total variation can be explained by this model. Model MSE was equal to 250.51. In addition, p-value analysis revealed that three slope coefficients (β_2, β_3, and β_5) were statistically not significant at α=0.05 (Table 4). Other five slope coefficients (β_1, β_4, β_6, β_7, and β_8) were statistically significant at α=0.05.

Table 4. Estimated regression coefficients for full model

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	2.32969	3.67238	0.63	0.5291	0
POP	1	-0.22532	0.05420	-4.16	0.0001	1920.50178
AREA	1	-0.00248	0.00313	-0.79	0.4330	3.23493
YOUNG	1	-0.28914	0.18963	-1.52	0.1345	375.44101
DIV	1	7.32687	1.91414	3.83	0.0004	41.91867
OLD	1	-0.00231	0.08904	-0.03	0.9794	107.16244
EDUC	1	0.29766	0.04644	6.41	<.0001	622.20492
POV	1	0.64182	0.05808	11.05	<.0001	46.99628
UNEMP	1	0.31470	0.06161	5.11	<.0001	16.13299

Multicollinearity diagnostics results are illustrated in Table 5.

Table 5. Multicollinearity diagnostics results of full model

Collinearity Diagnostics											
			Proportion of Variation								
Number	Eigenvalue	Condition Index	Intercept	POP	AREA	YOUNG	DIV	OLD	EDUC	POV	UNEMP
1	7.82658	1.00000	0.00279	0.00000515	0.00130	0.00002611	0.00021232	0.00009131	0.00001537	0.00022992	0.00062541
2	0.80982	3.10879	0.22221	0.00000663	0.04599	0.00002349	8.162565E-7	0.00018931	0.00001017	0.00174	0.00173
3	0.23091	5.82193	0.62792	2.314465E-7	0.25724	0.00001629	0.00289	0.00065274	0.00000232	0.00145	0.00145
4	0.06570	10.91487	0.01968	0.00003992	0.07925	0.00081791	0.03940	0.00368	0.00040764	0.01090	0.38151
5	0.03614	14.71613	0.00049034	0.00020848	0.23106	0.00429	0.00702	0.00191	0.00351	0.27637	0.01740
6	0.02147	19.09181	0.10428	0.00032904	0.09630	0.00000271	0.12350	0.13101	0.00131	0.02246	0.41619
7	0.00798	31.31201	0.01359	0.00083072	0.20281	0.05336	0.68415	0.16326	0.00335	0.09037	0.00594
8	0.00115	82.50345	0.00599	0.00373	0.08297	0.59147	0.13324	0.24630	0.41338	0.14228	0.03674
9	0.00024787	177.69358	0.00304	0.99485	0.00309	0.34999	0.00959	0.45290	0.57801	0.45421	0.13841

VIF values exceeded 10 (i.e. $R^2 > 0.9$) for seven variables. However, four predictor variables produced VIF values that were higher than 100 and one exceeded 1000, which is a clear red flag. Also, Pearson correlation results identified a few variable pairs that produced higher than 0.99 correlation coefficients (POP and YOUNG, POP and EDUC, YOUNG and EDUC). Condition index values in Table 5 indicate that number 7,8, and 9 exceed recommended condition number equal to 30 demonstrating the effects of multicollinearity. In addition, one proportion of variation value (Number 9 and POP) was higher than 0.99 indicating that POP is causing collinearity in the model.

3.2 Ridge Regression

Ridge regression results are represented in Tables 6 and 7.

Table 6. Ridge regression results for full model and slope coefficient estimates

Obs	_TYPE_	_RIDGE_	_RMSE_	Intercept	POP	AREA	YOUNG	DIV	OLD	EDUC	POV	UNEMP
3	RIDGE	0.00	15.8274	2.32969	-0.22532	-.002480865	-0.28914	7.32687	-0.002312	0.29766	0.64182	0.31470
5	RIDGE	0.01	22.0030	2.82524	-0.00380	-.006042733	-0.10170	7.56815	-0.051722	0.02386	0.33975	0.13345
7	RIDGE	0.02	23.3582	1.83525	0.00135	-.005892844	-0.03657	6.76191	-0.024349	0.01394	0.29800	0.11216
9	RIDGE	0.03	24.3715	1.06451	0.00362	-.005698947	-0.00691	6.16507	-0.007635	0.01171	0.27038	0.10234
11	RIDGE	0.04	25.2126	0.46492	0.00501	-.005515412	0.01079	5.71693	0.004203	0.01120	0.24965	0.09743
13	RIDGE	0.05	25.9285	-0.01182	0.00596	-.005341314	0.02277	5.36763	0.013143	0.01123	0.23326	0.09505
15	RIDGE	0.06	26.5466	-0.39866	0.00666	-.005173879	0.03146	5.08673	0.020163	0.01143	0.21990	0.09408
17	RIDGE	0.07	27.0865	-0.71778	0.00719	-.005011503	0.03808	4.85516	0.025830	0.01170	0.20877	0.09395
19	RIDGE	0.08	27.5630	-0.98445	0.00762	-.004853389	0.04329	4.66048	0.030502	0.01197	0.19932	0.09431
21	RIDGE	0.09	27.9876	-1.20954	0.00797	-.004699153	0.04749	4.49416	0.034419	0.01224	0.19119	0.09495
23	RIDGE	0.10	28.3691	-1.40097	0.00825	-.004548602	0.05095	4.35018	0.037747	0.01248	0.18411	0.09576
25	RIDGE	0.11	28.7147	-1.56463	0.00849	-.004401629	0.05384	4.22414	0.040608	0.01270	0.17788	0.09666
27	RIDGE	0.12	29.0301	-1.70506	0.00869	-.004258156	0.05629	4.11276	0.043090	0.01290	0.17236	0.09760
29	RIDGE	0.13	29.3198	-1.82577	0.00886	-.004118118	0.05838	4.01352	0.045261	0.01308	0.16741	0.09854
31	RIDGE	0.14	29.5876	-1.92953	0.00901	-.003981447	0.06019	3.92447	0.047173	0.01325	0.16297	0.09947
33	RIDGE	0.15	29.8364	-2.01859	0.00914	-.003848070	0.06177	3.84405	0.048868	0.01339	0.15894	0.10038
35	RIDGE	0.16	30.0690	-2.09476	0.00925	-.003717910	0.06315	3.77101	0.050379	0.01352	0.15526	0.10125
37	RIDGE	0.17	30.2873	-2.15955	0.00935	-.003590886	0.06437	3.70435	0.051732	0.01364	0.15191	0.10208
39	RIDGE	0.18	30.4932	-2.21420	0.00944	-.003466916	0.06545	3.64323	0.052948	0.01375	0.14882	0.10288
41	RIDGE	0.19	30.6882	-2.25977	0.00952	-.003345914	0.06641	3.58696	0.054045	0.01385	0.14597	0.10363
43	RIDGE	0.20	30.8736	-2.29715	0.00958	-.003227795	0.06727	3.53496	0.055039	0.01394	0.14333	0.10434

Table 7. Ridge regression effect on slope coefficient VIF

Obs	_RIDGE_	Intercept	POP	AREA	YOUNG	DIV	OLD	EDUC	POV	UNEMP
2	0.00	.	1920.50	3.23493	375.441	41.9187	107.162	622.205	46.9963	16.1330
4	0.01	.	4.28	2.18276	13.735	14.8450	15.137	10.420	11.6672	9.3490
6	0.02	.	1.56	1.84181	5.985	8.6510	8.874	4.331	8.1870	7.0847
8	0.03	.	0.90	1.63687	3.559	5.8939	6.013	2.650	6.1790	5.6344
10	0.04	.	0.61	1.49424	2.423	4.3687	4.399	1.871	4.8622	4.6272
12	0.05	.	0.46	1.38680	1.782	3.4134	3.382	1.422	3.9397	3.8891
14	0.06	.	0.36	1.30157	1.379	2.7651	2.694	1.131	3.2642	3.3269
16	0.07	.	0.29	1.23142	1.106	2.2995	2.204	0.928	2.7532	2.8859
18	0.08	.	0.24	1.17206	0.911	1.9512	1.842	0.778	2.3565	2.5321
20	0.09	.	0.21	1.12072	0.766	1.6820	1.565	0.665	2.0420	2.2428
22	0.10	.	0.18	1.07554	0.655	1.4688	1.349	0.576	1.7883	2.0028
24	0.11	.	0.16	1.03521	0.568	1.2964	1.177	0.505	1.5804	1.8010
26	0.12	.	0.14	0.99880	0.498	1.1547	1.037	0.447	1.4080	1.6295
28	0.13	.	0.13	0.96560	0.442	1.0364	0.921	0.399	1.2632	1.4824
30	0.14	.	0.12	0.93508	0.395	0.9366	0.825	0.359	1.1405	1.3551
32	0.15	.	0.11	0.90684	0.355	0.8514	0.744	0.326	1.0355	1.2441
34	0.16	.	0.10	0.88056	0.322	0.7780	0.675	0.297	0.9450	1.1467
36	0.17	.	0.09	0.85597	0.293	0.7143	0.616	0.272	0.8663	1.0607
38	0.18	.	0.08	0.83287	0.269	0.6586	0.564	0.250	0.7976	0.9844
40	0.19	.	0.08	0.81109	0.247	0.6096	0.520	0.231	0.7371	0.9163
42	0.20	.	0.07	0.79048	0.228	0.5662	0.480	0.214	0.6836	0.8553

Best model represents the balance between reducing variance and increasing bias. Best *k* value needs to produce low RMSE value (as RMSE will increase since more bias is introduced with increase in *k*) and low VIF values (to reduce variance in the model). VIF values for all slope

coefficients reduce below 10 with *k*=0.02. Slope coefficient algebraic sign analysis was done do select the "best" *k* values for ridge regression. With increasing *k* value, slope coefficient sign changes for three variables. One of those coefficients is slope coefficient for intercept, which should not be positive as it would produce positive crime number even if all predictor variables are equal to zero. Relying on this logic, the best k value is 0.05 as it is the first k value to produce negative $\hat{\beta}_0$. Picking the first value that produces desired algebraic sign minimizes the bias that is introduce to a model, minimizing model's RMSE.

Best ridge regression model can be represented as (coefficients were reduced to four decimal places):

$\hat{Y}$ = – 0.0118 + 0.0060*X1 – 0.0053*X2 + 0.0228*X3 + 5.3676*X4 + 0.0131*X5 + 0.0112*X6 + 0.2333*X7 + 0.0951*X8 [3]

3.3 Principal Component Regression

Principal component regression results are represented in Table 8.

Table 8. Principal component regression results. Model coefficients on the top, VIF values on the bottom

Obs	_TYPE_	_PCOMIT_	_RMSE_	Intercept	POP	AREA	YOUNG	DIV	OLD	EDUC	POV	UNEMP
3	IPC	1	18.8623	3.22386	0.015979	-0.003250	-0.79225	6.4783	-0.27044	0.14067	0.46729	0.21258
5	IPC	2	22.2105	4.65991	0.000043	-0.007905	-0.04090	10.0608	-0.04302	-0.01415	0.35264	0.15052
7	IPC	3	22.1960	5.31536	0.002114	-0.006042	0.01830	7.9531	-0.09377	-0.01058	0.37756	0.14530
9	IPC	4	26.4415	-1.63271	0.009166	-0.000488	0.03510	3.3610	0.09940	0.00340	0.31499	-0.11692
11	IPC	5	30.5361	-2.33181	0.013501	-0.010198	0.11855	4.4015	0.05710	0.02231	0.10570	-0.01781
13	IPC	6	32.0715	-4.21841	0.011959	-0.005460	0.09035	2.2137	0.08908	0.01730	0.09039	0.15101

Obs	_TYPE_	_PCOMIT_	_RMSE_	Intercept	POP	AREA	YOUNG	DIV	OLD	EDUC	POV	UNEMP
2	IPCVIF	1	.	.	8.87349	3.22516	242.723	41.5050	58.3593	265.091	25.6887	13.9072
4	IPCVIF	2	.	.	2.64913	2.95773	21.752	35.9999	32.1487	5.809	18.8233	13.2934
6	IPCVIF	3	.	.	1.13233	2.33959	1.956	8.5009	13.3082	3.828	14.1417	13.2308
8	IPCVIF	4	.	.	0.34940	2.09514	1.885	2.6920	1.1628	2.470	12.8288	6.1936
10	IPCVIF	5	.	.	0.09580	1.45421	0.384	2.4363	0.6633	0.338	0.2286	5.3314
12	IPCVIF	6	.	.	0.02798	1.13157	0.021	0.0460	0.0600	0.022	0.0860	0.0431

Eigenvalues of the correlation matrix revealed that 97.21% of information is represented within two first principal components. Therefore, omitting the other principal components would not cause the model to lose major amount of information. Table 8 shows that increasing the number of omitted principal components will increase model's RMSE. However, it also reduces coefficient's VIF values. By omitting 2 principal components, VIF values noticeably improve for some of the slope coefficients. To pick the "best" PC regression model, algebraic signs of model coefficients need to be analyzed. Negative sign would mean that predictor variable negatively affects the Y. PCOMIT=4 produced positive slope coefficients for most predictor variables, expect one. Also, it produced negative intercept coefficient.

Best PC regression model can be represented as (coefficients were reduced to four decimal places):

$\hat{Y}$ = – 1.6327+ 0.0091*X1 – 0.0005*X2 + 0.0351*X3 + 3.3610*X4 + 0.0994*X5 + 0.0034*X6 + 0.3150*X7 – 0.1169*X8

3.4 Model Comparison

Slope coefficient estimates produced by OLS, "best" ridge and PC regression models are represented in Table 9.

Table 9. Coefficient comparison (reduced to 4 decimal places)

Model	Inter.	POP	AREA	YOUNG	DIV	OLD	EDUC	POV	UNEMP
OLD	2.3297	-0.2253	-0.0025	-0.2891	7.3269	-0.0023	0.2977	0.6418	0.3147
Ridge	-0.0118	0.0060	-0.0053	0.0228	5.3676	0.0131	0.0112	0.2333	0.0951
PC	-1.6327	0.0091	-0.0005	0.0351	3.3610	0.0994	0.0034	0.3150	-0.1169

Comparing three models, variation in coefficient value as well as algebraic sign in some cases ca be noted. Since ridge and PC regression models were selected based on analyzing algebraic signs of the model coefficients, slope coefficient values can vary slightly based on the logic used while deciding which sign for the model coefficient is appropriate.

4. Summary

Relationship between total number of serious crimes in 1980 and eight factors was analyzed. Data was collected from 53 metropolitan areas and full linear regression model was developed. Then, multicollinearity diagnostics was used on the full model to identify predictor variables causing correlation in the model and affecting its performance. To address the problem of multicollinearity in the model, two methods were used: ridge and principal component regressions. "Best" model for both methods was picked based on their RMSE and VIF values as well as algebraic sign of the model coefficients.

A few improvements could be done to improve the model developed in this study. First, as predictor variables causing correlation in the model is identified, it should be removed from the model. Then, additional tests need to be performed to evaluate if the multicollinearity was removed from the model.

5. SAS programs

```
*HW8* RUTA BASIJOKAITE*;
PROC IMPORT DATAFILE="/folders/myfolders/HW8/PMSA.xlsx" /** Import an XLSX file.  **/
            OUT=WORK.HW8DATA
            DBMS=XLSX
            REPLACE;
            GETNAMES=YES;
            SCANTEXT=YES;
RUN;
OPTIONS NOCENTER NODATE PAGENO=1 LS=76 PS=45 NOLABEL;
DATA ALL;
  SET HW8DATA;
RUN;
PROC MEANS N MEAN MEDIAN STD MIN MAX MAXDEC=4;
   VAR CRIME POP AREA YOUNG DIV OLD EDUC POV UNEMP;
   TITLE 'DESCRIPTIVE STATISTICS';
RUN;
PROC CORR PEARSON SPEARMAN;
  VAR CRIME POP AREA YOUNG DIV OLD EDUC POV UNEMP;
  TITLE 'CORRELATION AMONG VARIABLES';
RUN;
PROC SGSCATTER DATA=ALL;
   MATRIX CRIME POP AREA YOUNG DIV OLD EDUC POV UNEMP / DIAGONAL=(HISTOGRAM NORMAL);
   TITLE 'EXPEND CORRELATION WITH FIVE PREDICTOR VARIABLES';
RUN;
PROC REG DATA=ALL OUTEST=RIDGE RIDGE=0 TO 0.2 BY 0.01 OUTVIF;
  MODEL CRIME= POP AREA YOUNG DIV OLD EDUC POV UNEMP / VIF COLLIN;
  TITLE 'FULL MODEL AND RIDGE REGRESSION';
RUN;
PROC PRINT DATA=RIDGE;
  VAR _TYPE_ _RIDGE_ _RMSE_ INTERCEPT POP AREA YOUNG DIV OLD EDUC POV UNEMP;
  WHERE _TYPE_ = 'RIDGE';
RUN;
PROC PRINT DATA=RIDGE;
  VAR _RIDGE_ INTERCEPT POP AREA YOUNG DIV OLD EDUC POV UNEMP;
  WHERE _TYPE_ = 'RIDGEVIF';
RUN;
PROC PRINCOMP DATA=ALL;
  VAR CRIME POP AREA YOUNG DIV OLD EDUC POV UNEMP;
RUN;
PROC REG DATA=ALL OUTEST=PRINEST OUTVIF;
  MODEL CRIME= POP AREA YOUNG DIV OLD EDUC POV UNEMP / PCOMIT=1 TO 6 VIF;
  TITLE 'PRINCIPAL COMPONENT REGRESSION';
RUN;
*PROC PRINT DATA=PRINEST;
*  VAR _TYPE_ _PCOMIT_ _RMSE_ INTERCEPT POP AREA YOUNG DIV OLD EDUC POV UNEMP;
*  WHERE _TYPE_ = 'PARMS';
*RUN;
PROC PRINT DATA=PRINEST;
  VAR _TYPE_ _PCOMIT_ _RMSE_ INTERCEPT POP AREA YOUNG DIV OLD EDUC POV UNEMP;
  WHERE _TYPE_ = 'IPC';
RUN;
PROC PRINT DATA=PRINEST;
  VAR _TYPE_ _PCOMIT_ _RMSE_ INTERCEPT POP AREA YOUNG DIV OLD EDUC POV UNEMP;
  WHERE _TYPE_ = 'IPCVIF';
RUN;
```