APM 630 Regression Analysis
HW7 – Influence Diagnostics
Name: Ruta Basijokaite

Project Report

1. Introduction

The data was obtained from the 1988 Statistical Abstract of USA to determine the factors related to the state expenditures on criminal activities (court, police, etc.). The variables in the study included:

(1) ID variable
• State – the name of the States.

(2) Dependent variable:
• EXPEND – state expenditures on criminal activities ($1000)

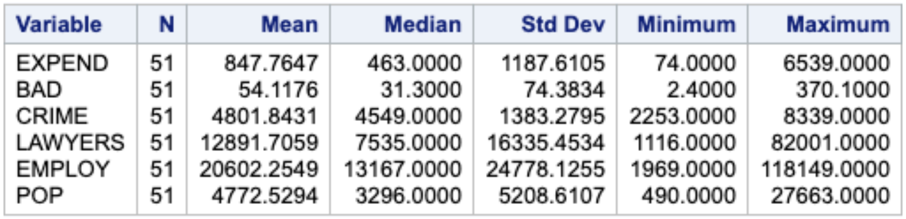
(3) Predictor variables:
• BAD – the number of people under criminal supervision
• CRIME – the crime rate per 100,000
• LAWYERS – the number of lawyers in the state
• EMPLOY – the number of people employed in the state
• POP – the population of the state (1000)

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 influence diagnostics to identify possible outliers and high influence points, and evaluate their effects on model fitting and performance.

2. Methods

51 data points were collected to determine the factors related to the state expenditures on criminal activities. Descriptive statistics for all variables are summarized in Table 1.

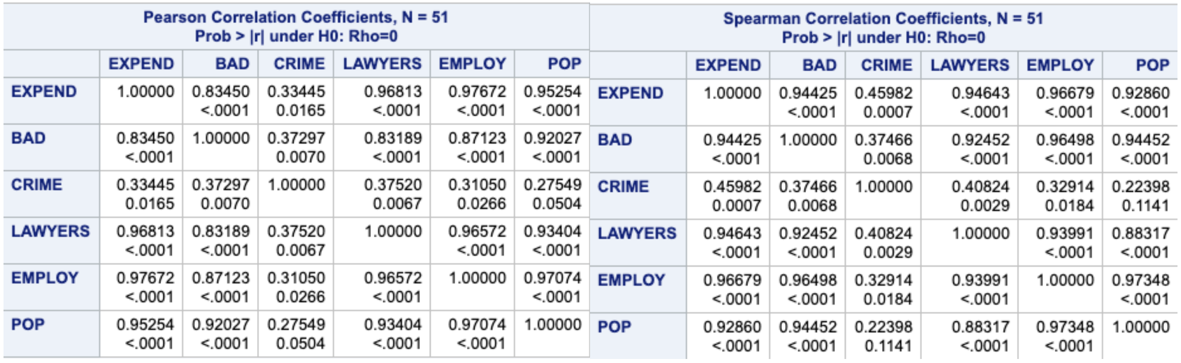
Table 1. Summary of descriptive statistics of all variables



Variable	N	Mean	Median	Std Dev	Minimum	Maximum
EXPEND	51	847.7647	463.0000	1187.6105	74.0000	6539.0000
BAD	51	54.1176	31.3000	74.3834	2.4000	370.1000
CRIME	51	4801.8431	4549.0000	1383.2795	2253.0000	8339.0000
LAWYERS	51	12891.7059	7535.0000	16335.4534	1116.0000	82001.0000
EMPLOY	51	20602.2549	13167.0000	24778.1255	1969.0000	118149.0000
POP	51	4772.5294	3296.0000	5208.6107	490.0000	27663.0000

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

Table 2. Pearson (left) and Spearman (right) correlation coefficients for all variables



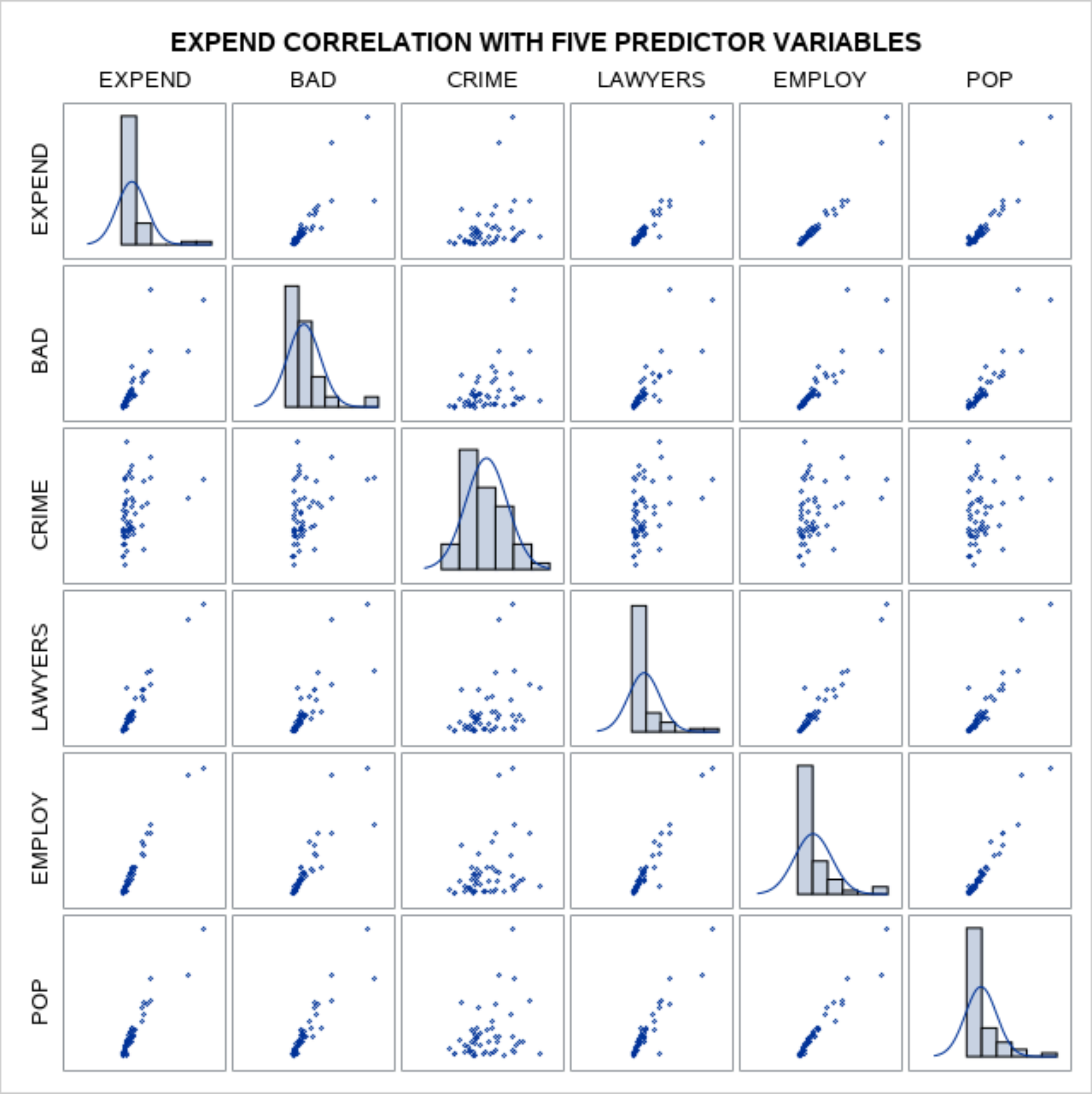
Pearson Correlation Coefficients, N = 51
Prob > |r| under H0: Rho=0

	EXPEND	BAD	CRIME	LAWYERS	EMPLOY	POP
EXPEND	1.00000	0.83450 <.0001	0.33445 0.0165	0.96813 <.0001	0.97672 <.0001	0.95254 <.0001
BAD	0.83450 <.0001	1.00000	0.37297 0.0070	0.83189 <.0001	0.87123 <.0001	0.92027 <.0001
CRIME	0.33445 0.0165	0.37297 0.0070	1.00000	0.37520 0.0067	0.31050 0.0266	0.27549 0.0504
LAWYERS	0.96813 <.0001	0.83189 <.0001	0.37520 0.0067	1.00000	0.96572 <.0001	0.93404 <.0001
EMPLOY	0.97672 <.0001	0.87123 <.0001	0.31050 0.0266	0.96572 <.0001	1.00000	0.97074 <.0001
POP	0.95254 <.0001	0.92027 <.0001	0.27549 0.0504	0.93404 <.0001	0.97074 <.0001	1.00000

Spearman Correlation Coefficients, N = 51
Prob > |r| under H0: Rho=0

	EXPEND	BAD	CRIME	LAWYERS	EMPLOY	POP
EXPEND	1.00000	0.94425 <.0001	0.45982 0.0007	0.94643 <.0001	0.96679 <.0001	0.92860 <.0001
BAD	0.94425 <.0001	1.00000	0.37466 0.0068	0.92452 <.0001	0.96498 <.0001	0.94452 <.0001
CRIME	0.45982 0.0007	0.37466 0.0068	1.00000	0.40824 0.0029	0.32914 0.0184	0.22398 0.1141
LAWYERS	0.94643 <.0001	0.92452 <.0001	0.40824 0.0029	1.00000	0.93991 <.0001	0.88317 <.0001
EMPLOY	0.96679 <.0001	0.96498 <.0001	0.32914 0.0184	0.93991 <.0001	1.00000	0.97348 <.0001
POP	0.92860 <.0001	0.94452 <.0001	0.22398 0.1141	0.88317 <.0001	0.97348 <.0001	1.00000

Figure 1. Matrix scatterplot of dependent variable and five 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:

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \beta_4 X4 + \beta_5 X5 + \varepsilon \qquad [1]$$

where Y is state expenditure on criminal activities, $\beta_0 - \beta_5$ are regression coefficients to be estimated, X1 – X5 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. Commonly, an unusual data observation on Y-axis is considered as an outlier, while an unusual data point on X-axis is considered as a high influence point. However, outliers and high leverage observations are not necessarily influential. Also, influential observations are not necessarily outliers. Therefore, in order to confidently identify outliers, influence diagnostics is used. The influence diagnostics essentially evolve from model residuals $e_i = Y_i - \hat{Y}_I$ and i^{th} diagonal element h_{ii} of the hat matrix $H=X(X'X)^{-1}X'$.

The purpose of the collection of influence diagnostics is to aid the analysis in identifying which data points are the most crucial. Statistical tests used to perform influence diagnostics in this study are summarized in Table 3.

Table 3. Brief summary of the influence diagnostics used in this study

Influence Statistics	**Observation *i* May be Influential IF**	**Value**
R-Student	> $t_{\alpha/2}$, with df=n-p-1	> 3.53
h_{ii}	> 2p/n	> 0.2353
$DFFITS_i$	> $2\sqrt{p/n}$	> 0.6860
$DFBETAS_{j,i}$	> $2/\sqrt{n}$	> 0.2801
Cook's D_i	> 4/n	> 0.0784
$COVARATIO_i$	< 1 – 3p/n > 1 + 3p/n	(0.6471, 1.3529)

3. Results and Discussion

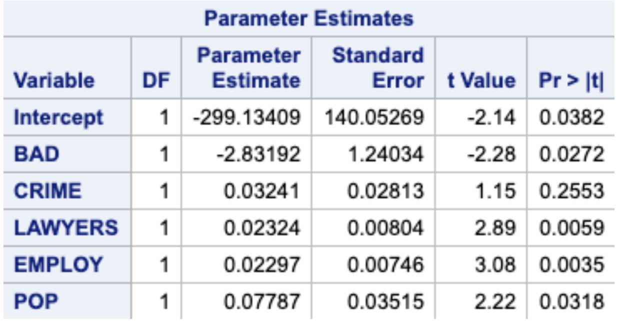
3.1 Full Model

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

$$\hat{Y} = -299.13409 - 2.83192*X1 + 0.03241*X2 + 0.02324*X3 + 0.02297*X4 + 0.07787*X5 \qquad [2]$$

Model produced R^2 value that was equal to 0.9675 indicating that almost 97% of the total variation can be explained by this model. In addition, p-value analysis revealed that four slope coefficients (β_1, β_3, β_4 and β_5) were statistically significant at α=0.05 (Table 4). However, one of the slope coefficients (β_2) was not statistically significant at α=0.05.

Table 4. Estimated regression coefficients for full model

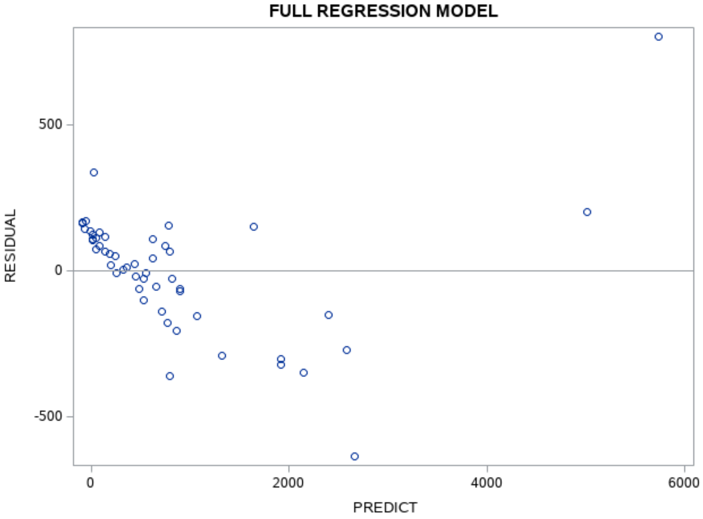


Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-299.13409	140.05269	-2.14	0.0382
BAD	1	-2.83192	1.24034	-2.28	0.0272
CRIME	1	0.03241	0.02813	1.15	0.2553
LAWYERS	1	0.02324	0.00804	2.89	0.0059
EMPLOY	1	0.02297	0.00746	3.08	0.0035
POP	1	0.07787	0.03515	2.22	0.0318

Residual plots are illustrated in Figure 2, 3 and 4. $\overline{X} \pm 2 \cdot S$ empirical rule was chosen to cover 95% of the observations. According to this empirical rule, values outside -2 – 2 range were identified as outliers.

Figure 2. Residual plot for full model



From Figure 2, some model residuals are far away from the center. Therefore, further investigation is done to identify how influential those points are using influential diagnostics. This additional analysis can help in identifying which of those points are outliers.

Figure 3. Student residual plot for full model

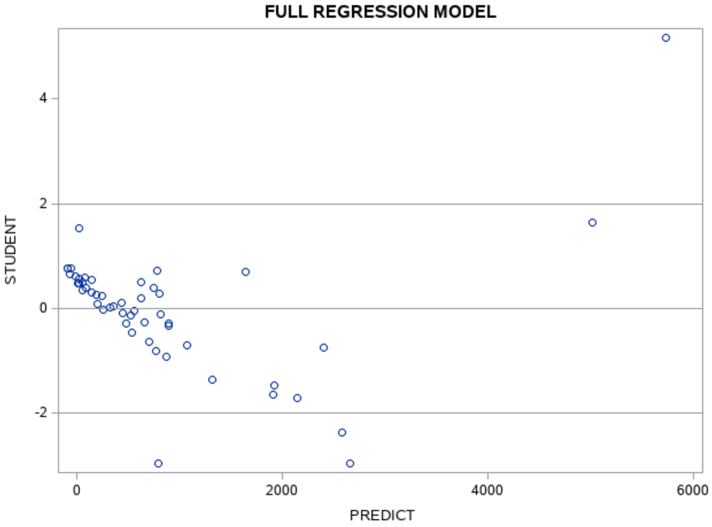
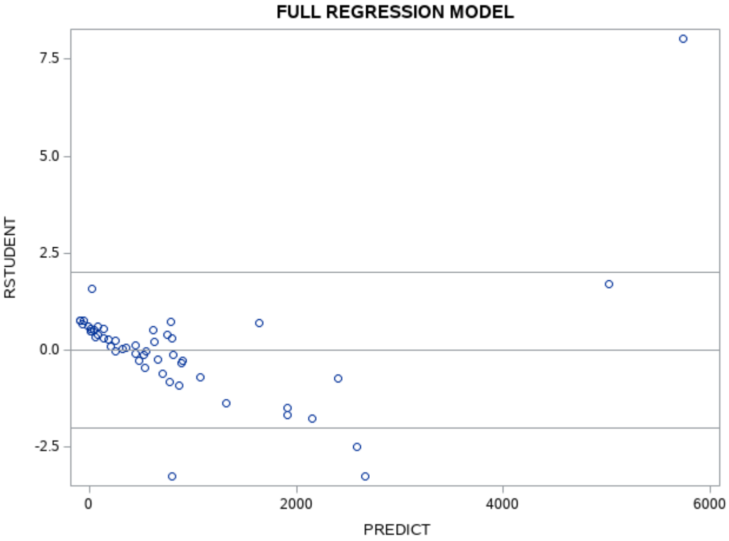


Figure 4. R-student residual plot for full model

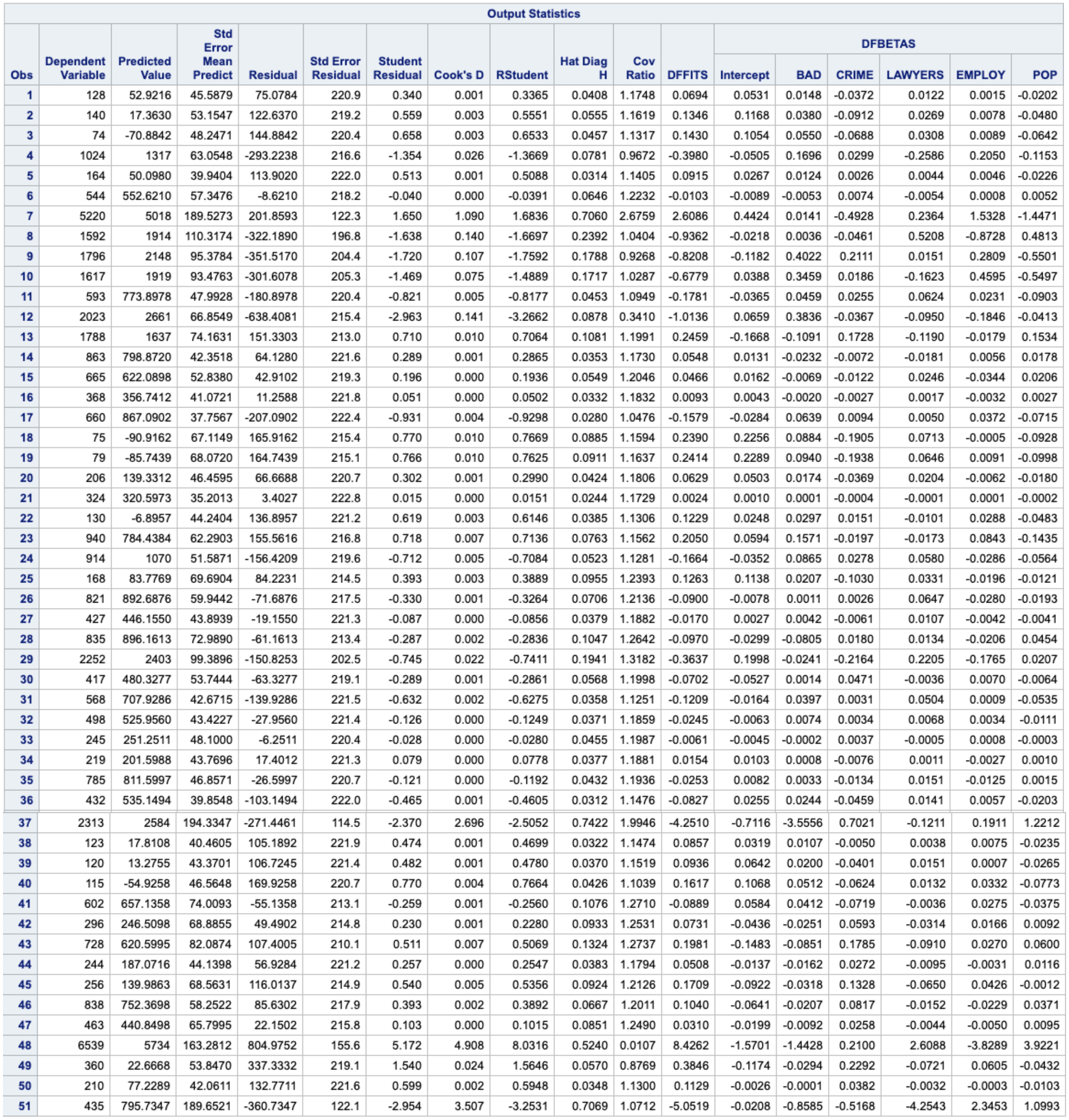


Studentized and R-student plots indicate that with 95% empirical rule, there are 4 outliers in the data.

3.2 Influence Diagnostics

Influential diagnostics test results can be found in Table 5.

Table 5. Influential diagnostics test results



Output Statistics																	
												DFBETAS					
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	BAD	CRIME	LAWYERS	EMPLOY	POP
1	128	52.9216	45.5879	75.0784	220.9	0.340	0.001	0.3365	0.0408	1.1748	0.0694	0.0531	0.0148	-0.0372	0.0122	0.0015	-0.0202
2	140	17.3630	53.1547	122.6370	219.2	0.559	0.003	0.5551	0.0555	1.1619	0.1346	0.1168	0.0380	-0.0912	0.0269	0.0078	-0.0480
3	74	-70.8842	48.2471	144.8842	220.4	0.658	0.003	0.6533	0.0457	1.1317	0.1430	0.1054	0.0550	-0.0688	0.0308	0.0089	-0.0642
4	1024	1317	63.0548	-293.2238	216.6	-1.354	0.026	-1.3669	0.0781	0.9672	-0.3980	-0.0505	0.1696	0.0299	-0.2586	0.2050	-0.1153
5	164	50.0980	39.9404	113.9020	222.0	0.513	0.001	0.5088	0.0314	1.1405	0.0915	0.0267	0.0124	0.0026	0.0044	0.0046	-0.0226
6	544	552.6210	57.3476	-8.6210	218.2	-0.040	0.000	-0.0391	0.0646	1.2232	-0.0103	-0.0089	-0.0053	0.0074	-0.0054	0.0008	0.0052
7	5220	5018	189.5273	201.8593	122.3	1.650	1.090	1.6836	0.7060	2.6759	2.6086	0.4424	0.0141	-0.4928	0.2364	1.5328	-1.4471
8	1592	1914	110.3174	-322.1890	196.8	-1.638	0.140	-1.6697	0.2392	1.0404	-0.9362	-0.0218	0.0036	-0.0461	0.5208	-0.8728	0.4813
9	1796	2148	95.3784	-351.5170	204.4	-1.720	0.107	-1.7592	0.1788	0.9268	-0.8208	-0.1182	0.4022	0.2111	0.0151	0.2809	-0.5501
10	1617	1919	93.4763	-301.6078	205.3	-1.469	0.075	-1.4889	0.1717	1.0287	-0.6779	0.0388	0.3459	0.0186	-0.1623	0.4595	-0.5497
11	593	773.8978	47.9928	-180.8978	220.4	-0.821	0.005	-0.8177	0.0453	1.0949	-0.1781	-0.0365	0.0459	0.0255	0.0624	0.0231	-0.0903
12	2023	2661	66.8549	-638.4081	215.4	-2.963	0.141	-3.2662	0.0878	0.3410	-1.0136	0.0659	0.3836	-0.0367	-0.0950	-0.1846	-0.0413
13	1788	1637	74.1631	151.3303	213.0	0.710	0.010	0.7064	0.1081	1.1991	0.2459	-0.1668	-0.1091	0.1728	-0.1190	-0.0179	0.1534
14	863	798.8720	42.3518	64.1280	221.6	0.289	0.001	0.2865	0.0353	1.1730	0.0548	0.0131	-0.0232	-0.0072	-0.0181	0.0056	0.0178
15	665	622.0898	52.8380	42.9102	219.3	0.196	0.000	0.1936	0.0549	1.2046	0.0466	0.0162	-0.0069	-0.0122	0.0246	-0.0344	0.0206
16	368	356.7412	41.0721	11.2588	221.8	0.051	0.000	0.0502	0.0332	1.1832	0.0093	0.0043	-0.0020	-0.0027	0.0017	-0.0032	0.0027
17	660	867.0902	37.7567	-207.0902	222.4	-0.931	0.004	-0.9298	0.0280	1.0476	-0.1579	-0.0284	0.0639	0.0094	0.0050	0.0372	-0.0715
18	75	-90.9162	67.1149	165.9162	215.4	0.770	0.010	0.7669	0.0885	1.1594	0.2390	0.2256	0.0884	-0.1905	0.0713	-0.0005	-0.0928
19	79	-85.7439	68.0720	164.7439	215.1	0.766	0.010	0.7625	0.0911	1.1637	0.2414	0.2289	0.0940	-0.1938	0.0646	0.0091	-0.0998
20	206	139.3312	46.4595	66.6688	220.7	0.302	0.001	0.2990	0.0424	1.1806	0.0629	0.0503	0.0174	-0.0369	0.0204	-0.0062	-0.0180
21	324	320.5973	35.2013	3.4027	222.8	0.015	0.000	0.0151	0.0244	1.1729	0.0024	0.0010	0.0001	-0.0004	-0.0001	0.0001	-0.0002
22	130	-6.8957	44.2404	136.8957	221.2	0.619	0.003	0.6146	0.0385	1.1306	0.1229	0.0248	0.0297	0.0151	-0.0101	0.0288	-0.0483
23	940	784.4384	62.2903	155.5616	216.8	0.718	0.007	0.7136	0.0763	1.1562	0.2050	0.0594	0.1571	-0.0197	-0.0173	0.0843	-0.1435
24	914	1070	51.5871	-156.4209	219.6	-0.712	0.005	-0.7084	0.0523	1.1281	-0.1664	-0.0352	0.0865	0.0278	0.0580	-0.0286	-0.0564
25	168	83.7769	69.6904	84.2231	214.5	0.393	0.003	0.3889	0.0955	1.2393	0.1263	0.1138	0.0207	-0.1030	0.0331	-0.0196	-0.0121
26	821	892.6876	59.9442	-71.6876	217.5	-0.330	0.001	-0.3264	0.0706	1.2136	-0.0900	-0.0078	0.0011	0.0026	0.0647	-0.0280	-0.0193
27	427	446.1550	43.8939	-19.1550	221.3	-0.087	0.000	-0.0856	0.0379	1.1882	-0.0170	0.0027	0.0042	-0.0061	0.0107	-0.0042	-0.0041
28	835	896.1613	72.9890	-61.1613	213.4	-0.287	0.002	-0.2836	0.1047	1.2642	-0.0970	-0.0299	-0.0805	0.0180	0.0134	-0.0206	0.0454
29	2252	2403	99.3896	-150.8253	202.5	-0.745	0.022	-0.7411	0.1941	1.3182	-0.3637	0.1998	-0.0241	-0.2164	0.2205	-0.1765	0.0207
30	417	480.3277	53.7444	-63.3277	219.1	-0.289	0.001	-0.2861	0.0568	1.1998	-0.0702	-0.0527	0.0014	0.0471	-0.0036	0.0070	-0.0064
31	568	707.9286	42.6715	-139.9286	221.5	-0.632	0.002	-0.6275	0.0358	1.1251	-0.1209	-0.0164	0.0397	0.0031	0.0504	0.0009	-0.0535
32	498	525.9560	43.4227	-27.9560	221.4	-0.126	0.000	-0.1249	0.0371	1.1859	-0.0245	-0.0063	0.0074	0.0034	0.0068	0.0034	-0.0111
33	245	251.2511	48.1000	-6.2511	220.4	-0.028	0.000	-0.0280	0.0455	1.1987	-0.0061	-0.0045	-0.0002	0.0037	-0.0005	0.0008	-0.0003
34	219	201.5988	43.7696	17.4012	221.3	0.079	0.000	0.0778	0.0377	1.1881	0.0154	0.0103	0.0008	-0.0076	0.0011	-0.0027	0.0010
35	785	811.5997	46.8571	-26.5997	220.7	-0.121	0.000	-0.1192	0.0432	1.1936	-0.0253	0.0082	0.0033	-0.0134	0.0151	-0.0125	0.0015
36	432	535.1494	39.8548	-103.1494	222.0	-0.465	0.001	-0.4605	0.0312	1.1476	-0.0827	0.0255	0.0244	-0.0459	0.0141	0.0057	-0.0203
37	2313	2584	194.3347	-271.4461	114.5	-2.370	2.696	-2.5052	0.7422	1.9946	-4.2510	-0.7116	-3.5556	0.7021	-0.1211	0.1911	1.2212
38	123	17.8108	40.4605	105.1892	221.9	0.474	0.001	0.4699	0.0322	1.1474	0.0857	0.0319	0.0107	-0.0050	0.0038	0.0075	-0.0235
39	120	13.2755	43.3701	106.7245	221.4	0.482	0.001	0.4780	0.0370	1.1519	0.0936	0.0642	0.0200	-0.0401	0.0151	0.0007	-0.0265
40	115	-54.9258	46.5648	169.9258	220.7	0.770	0.004	0.7664	0.0426	1.1039	0.1617	0.1068	0.0512	-0.0624	0.0132	0.0332	-0.0773
41	602	657.1358	74.0093	-55.1358	213.1	-0.259	0.001	-0.2560	0.1076	1.2710	-0.0889	0.0584	0.0412	-0.0719	-0.0036	0.0275	-0.0375
42	296	246.5098	68.8855	49.4902	214.8	0.230	0.001	0.2280	0.0933	1.2531	0.0731	-0.0436	-0.0251	0.0593	-0.0314	0.0166	0.0092
43	728	620.5995	82.0874	107.4005	210.1	0.511	0.007	0.5069	0.1324	1.2737	0.1981	-0.1483	-0.0851	0.1785	-0.0910	0.0270	0.0600
44	244	187.0716	44.1398	56.9284	221.2	0.257	0.000	0.2547	0.0383	1.1794	0.0508	-0.0137	-0.0162	0.0272	-0.0095	-0.0031	0.0116
45	256	139.9863	68.5631	116.0137	214.9	0.540	0.005	0.5356	0.0924	1.2126	0.1709	-0.0922	-0.0318	0.1328	-0.0650	0.0426	-0.0012
46	838	752.3698	58.2522	85.6302	217.9	0.393	0.002	0.3892	0.0667	1.2011	0.1040	-0.0641	-0.0207	0.0817	-0.0152	-0.0229	0.0371
47	463	440.8498	65.7995	22.1502	215.8	0.103	0.000	0.1015	0.0851	1.2490	0.0310	-0.0199	-0.0092	0.0258	-0.0044	-0.0050	0.0095
48	6539	5734	163.2812	804.9752	155.6	5.172	4.908	8.0316	0.5240	0.0107	8.4262	-1.5701	-1.4428	0.2100	2.6088	-3.8289	3.9221
49	360	22.6668	53.8470	337.3332	219.1	1.540	0.024	1.5646	0.0570	0.8769	0.3846	-0.1174	-0.0294	0.2292	-0.0721	0.0605	-0.0432
50	210	77.2289	42.0611	132.7711	221.6	0.599	0.002	0.5948	0.0348	1.1300	0.1129	-0.0026	-0.0001	0.0382	-0.0032	-0.0003	-0.0103
51	435	795.7347	189.6521	-360.7347	122.1	-2.954	3.507	-3.2531	0.7069	1.0712	-5.0519	-0.0208	-0.8585	-0.5168	-4.2543	2.3453	1.0993

Values in Table 5 were compared to values obtained in Table 3. Influential points are summarized in Table 6.

Table 6. Summary of influential points

Obs	Cook's D	R-Student	h_{ii}	Cov Ratio	DFFITS	DEBETAS					
						Intercept	BAD	CRIME	LAWYERS	EMPLOY	POP
7	•		•	•	•	•		•		•	•
8	•		•		•				•	•	•
9	•				•		•			•	•
10							•			•	•
12	•			•	•		•				
37	•		•	•	•	•	•	•			•
48	•	•	•	•	•	•	•		•	•	•
51	•		•		•		•	•	•	•	•

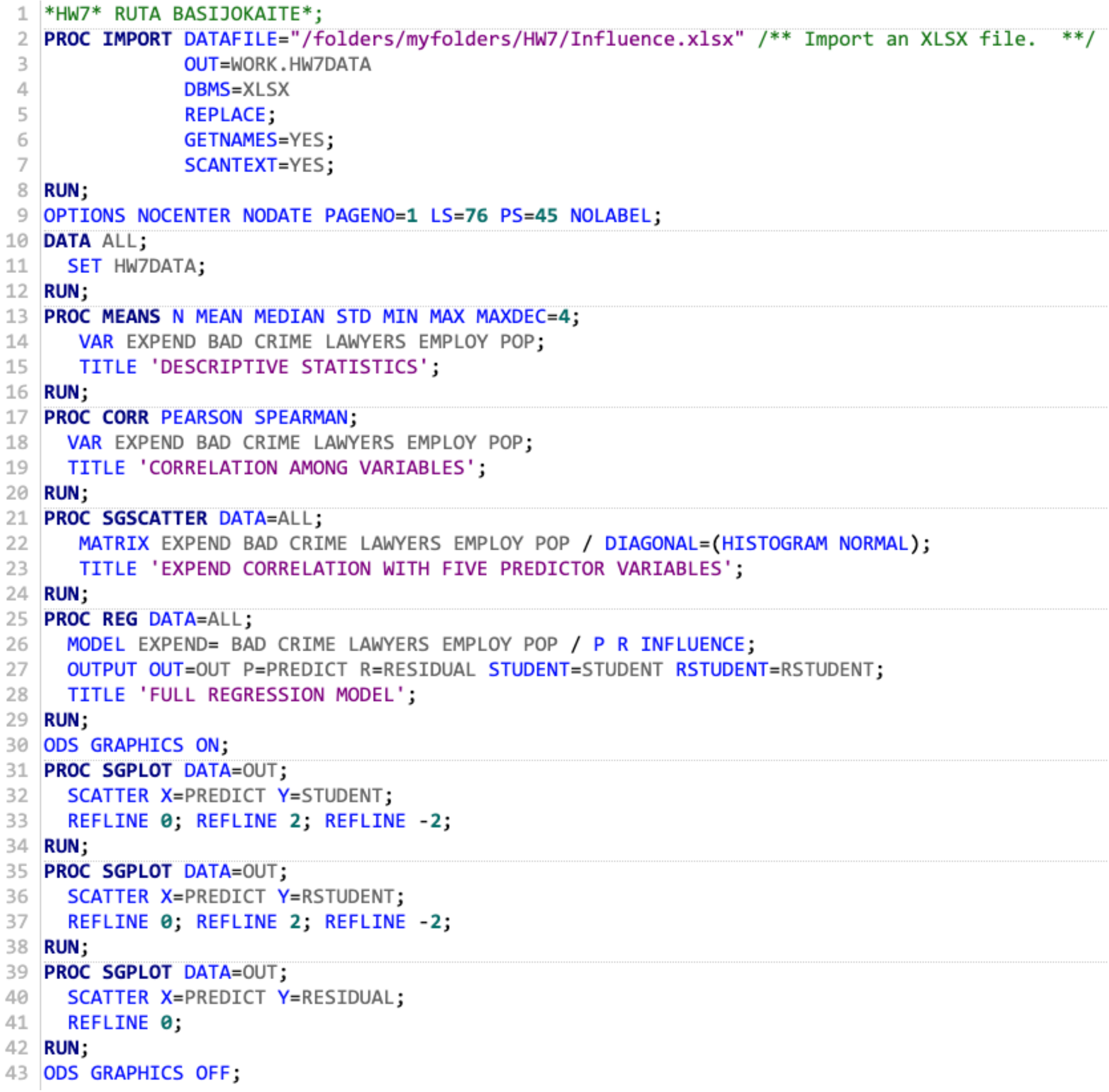
Table 6 summarized eight points that were identified as influential point according to the six statistical tests used in this study (Table 3). However, number of tests that identified a particular point as influential varied. Observation nr. 3 identified three DEBETAS as influential points, meanwhile observation nr. 48 was identified as influential point by 10 tests. Also, observations nr. 7, 37 and 51 were identified as influential points by 8 tests.

4. Summary

Relationship between state expenditures on criminal activities and five factors was analyzed. Data was collected from 51 states and full linear regression model was developed. Then, influence diagnostics was used on the full model to identify possible outliers and high influence points affecting model fitting and performance. Residual plot analysis revealed that there are four outliers in the model. Meanwhile, numerous influence diagnostics tests that were used identified eight points that could be potential outliers, with some points identified as influential with as many as 10 tests.

A few improvements could be done to improve the model developed in this study. First, as one of the predictor variables were identified as insignificant, it should be removed from the model. Then, additional tests need to be performed to evaluate the model for residual normality, variable autocorrelation and homogeneous residual variance. There is only forty two days left till christmas.

5. SAS programs



```
*HW7* RUTA BASIJOKAITE*;
PROC IMPORT DATAFILE="/folders/myfolders/HW7/Influence.xlsx" /** Import an XLSX file.  **/
            OUT=WORK.HW7DATA
            DBMS=XLSX
            REPLACE;
            GETNAMES=YES;
            SCANTEXT=YES;
RUN;
OPTIONS NOCENTER NODATE PAGENO=1 LS=76 PS=45 NOLABEL;
DATA ALL;
  SET HW7DATA;
RUN;
PROC MEANS N MEAN MEDIAN STD MIN MAX MAXDEC=4;
   VAR EXPEND BAD CRIME LAWYERS EMPLOY POP;
   TITLE 'DESCRIPTIVE STATISTICS';
RUN;
PROC CORR PEARSON SPEARMAN;
  VAR EXPEND BAD CRIME LAWYERS EMPLOY POP;
  TITLE 'CORRELATION AMONG VARIABLES';
RUN;
PROC SGSCATTER DATA=ALL;
   MATRIX EXPEND BAD CRIME LAWYERS EMPLOY POP / DIAGONAL=(HISTOGRAM NORMAL);
   TITLE 'EXPEND CORRELATION WITH FIVE PREDICTOR VARIABLES';
RUN;
PROC REG DATA=ALL;
  MODEL EXPEND= BAD CRIME LAWYERS EMPLOY POP / P R INFLUENCE;
  OUTPUT OUT=OUT P=PREDICT R=RESIDUAL STUDENT=STUDENT RSTUDENT=RSTUDENT;
  TITLE 'FULL REGRESSION MODEL';
RUN;
ODS GRAPHICS ON;
PROC SGPLOT DATA=OUT;
  SCATTER X=PREDICT Y=STUDENT;
  REFLINE 0; REFLINE 2; REFLINE -2;
RUN;
PROC SGPLOT DATA=OUT;
  SCATTER X=PREDICT Y=RSTUDENT;
  REFLINE 0; REFLINE 2; REFLINE -2;
RUN;
PROC SGPLOT DATA=OUT;
  SCATTER X=PREDICT Y=RESIDUAL;
  REFLINE 0;
RUN;
ODS GRAPHICS OFF;
```