Luke McDaneld MATH 564 Project

Part 1

1. Use stepwise regression to decide the final model. Use significance levels αE = αR = 0.15. In each step, report the result of regression analysis, indicate which predictor variable enters or leaves the model, and how the decision is made

Step 1: 4 different variables are regressed against y. X4 has the lowest p-value at 0.000576, so X4 is selected to enter the model.

Y = X1 + b

	Estimate	Std. Error	T-value	P-value
Intercept	81.4793	4.9273	16.54	4.07e-09
X1	1.8678	0.5264	3.55	0.00455

Y = X2 + b

	Estimate	Std. Error	T-value	P-value
Intercept	57.4237	8.4906	6.763	3.1e-05
X2	0.7891	0.1684	4.686	0.000665

Y = X3 + b

	Estimate	Std. Error	T-value	P-value
Intercept	110.2027	7.9478	13.866	2.6e-08
X3	-1.2558	0.5984	-2.098	0.0598

Y = X4 + b

	Estimate	Std. Error	T-value	P-value
Intercept	117.5679	5.2622	22.342	1.62e-10
X4	-0.7382	0.1546	-4.775	0.000576

Step 2: X4 has been selected to enter the model in the previous step. We test all possible candidates in the model with X4 and determine that X1 should enter the model as it has the lowest p-value. The current model is now Y = X4 + X1 + b.

Y = X4 + X1 + b

	Estimate	Std. Error	T-value	P-value
Intercept	103.09738	2.12398	48.54	3.32e-13
X4	-0.61395	0.04864	-12.62	1.81e-07
X1	1.43996	0.13842	10.40	1.11e-06

Luke McDaneld MATH 564 Project

Y = X4 + X2 + b

	Estimate	Std. Error	T-value	P-value
Intercept	94.1601	56.6271	1.663	0.127
X4	-0.4569	0.6960	-0.657	0.526
X2	0.3109	0.7486	0.415	0.687

Y = X4 + X3 + b

	Estimate	Std. Error	T-value	P-value
Intercept	131.28241	3.27477	40.089	2.23e-12
X4	-0.72460	0.07233	-10.018	1.56e-06
X3	-1.19985	0.18902	-6.348	8.38e-05

Step 3: We test to see whether X2 or X3 should enter the model. X2 meets the threshold for being included in the model, but the inclusion of X2 causes the p-value of x4 to become greater than the threshold for leaving the model, so X4 is removed. The model is currently y = X1 + X2 +b.

Y = X4 + X1 + X2+ b

	Estimate	Std. Error	T-value	P-value
Intercept	71.6483	14.1424	5.066	0.000675
X4	-0.2365	0.1733	-1.365	0.205395
X1	1.4519	0.1170	12.410	5.78e-07
X2	0.4161	0.1856	2.242	0.051687

Y = X4 + X1 + X3+ b

	Estimate	Std. Error	T-value	P-value
Intercept	111.68441	4.56248	24.479	1.52e-09
X4	-0.64280	0.04454	-14.431	1.58e-07
X1	1.05185	0.22368	4.702	0.00112
X3	-0.41004	0.19923	-2.058	0.06969

Step 4: We test the current model, y = X1 + X2 + b again to see if either X3 or X4 should be included in the model. Neither variable meets the threshold to be included, so the conclusion reached by the stepwise regression is that the final model should be y = X1 + X2 + b.

Y = X1 + X2 + X3 + b

	Estimate	Std. Error	T-value	P-value
Intercept	48.19363	3.91330	12.315	6.17e-07
X1	1.69589	0.20458	8.290	1.66e-05
X2	0.65691	0.04423	14.851	1.23e-07
X3	0.25002	0.18471	1.354	0.209

Y = X1 + X2 + X4+ b

	Estimate	Std. Error	T-value	P-value
Intercept	71.6483	14.1424	5.066	0.000675
X1	1.4519	0.1170	12.410	5.78e-07
X2	0.4161	0.1856	2.242	0.051687
X4	-0.2365	0.1733	-1.365	0.205395

Step 5: The final model as a result of the stepwise regression is as shown below.

	Estimate	Std. Error	T-value	P-value
Intercept	52.57735	2.28617	23.00	5.46e-10
X1	1.46831	0.12130	12.11	2.69e-07
X2	0.66225	0.04585	14.44	5.03e-08

2. Use best subset regression to decide the final model. Your work should include a screening procedure that gives you a set of candidate models, then use adjusted R^2 and Cp statistic to select from the candidate set and then use VIF to break the tie if any. Finally, use residual analysis to check the final model: show the residual vs fits plot and the normal probability plot, and conclude if the model is the right model.

Subset selection is run to determine the best possible model for when the model has 1, 2, 3 or 4 predictors. The result of the subset regression resulted in the following candidates:

	Adj R^2	Cp Statistic
Y = X4 + b	0.6449549	138.730833
Y = X1 +X2 + b	0.9744140	2.678242
Y = X1 + X2+ 4 + b	0.9764473	3.018233
Y = X1 + X2 + X3 + X4 + b	0.9735634	5.000000

As the highlighted values indicate, this procedure produces 2 candidate models. The 2-predictor model has the lowest Cp statistic out of all the candidates, and the 3-predictor model has the greatest R^2 out of all the candidates. In order to break the tie between the remaining to candidate models, we examine the variance inflation factors.

Y = X1 + X2 + b

	VIF
X1	1.055129
X2	1.055129

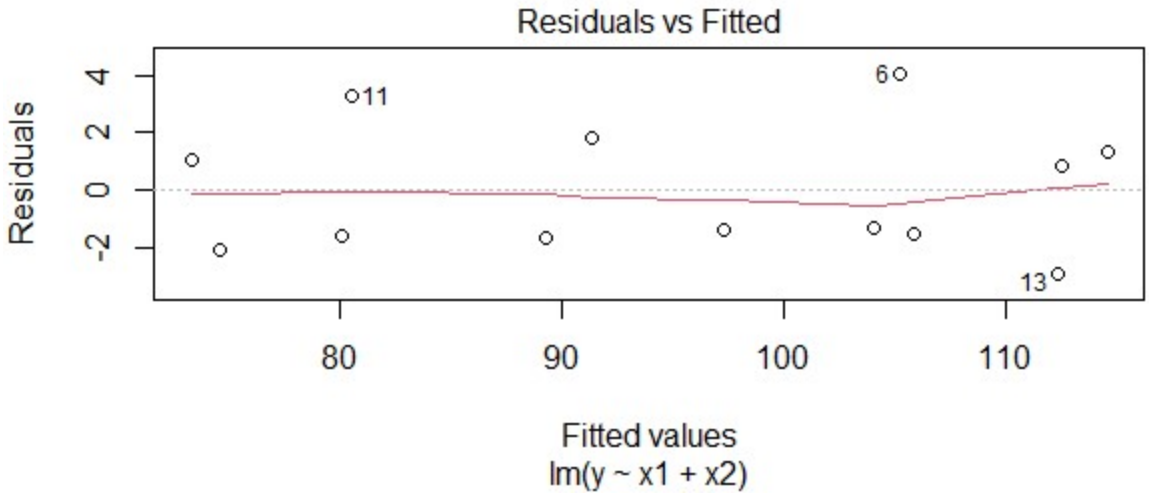
Y = X1 + X2 + X4 + b

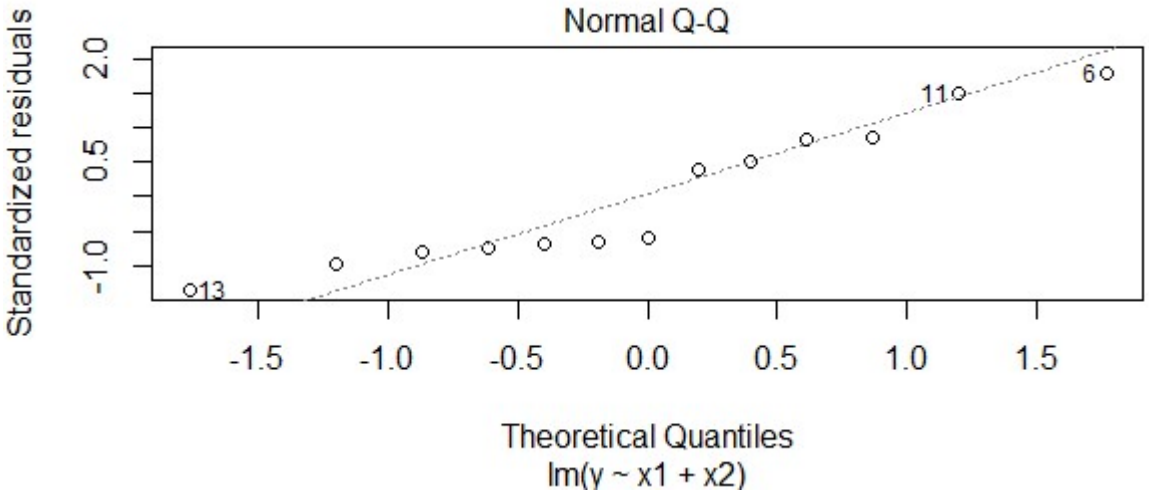
	VIF
X1	1.06633
X2	18.78031
X4	18.94008

Based on the variance inflation factors X2 and X4 in the 3-predictor model (Y = X1 + X2 + X4 + b), it is highly likely that there is serious multicollinearity between the two predictors X2 and X4. Both values are well above the threshold of 10 to suggest multicollinearity.

In regard to the 2-predictor model, neither predictor (X1 and X2) breaks the threshold of warranting further investigation. This implies that multicollinearity is unlikely between those 2 predictors.

Next, we examine the residual vs fits plot and the normal probability plot.





Both plots suggest there might be some slight outliers or irregularity in the model, however this is to be expected with such a small sample size. The plots do not appear to be too far from expectation, and combined with our other statistical measures suggest that the selected model is right model for these observations.

Luke McDaneld MATH 564 Project

Part 2

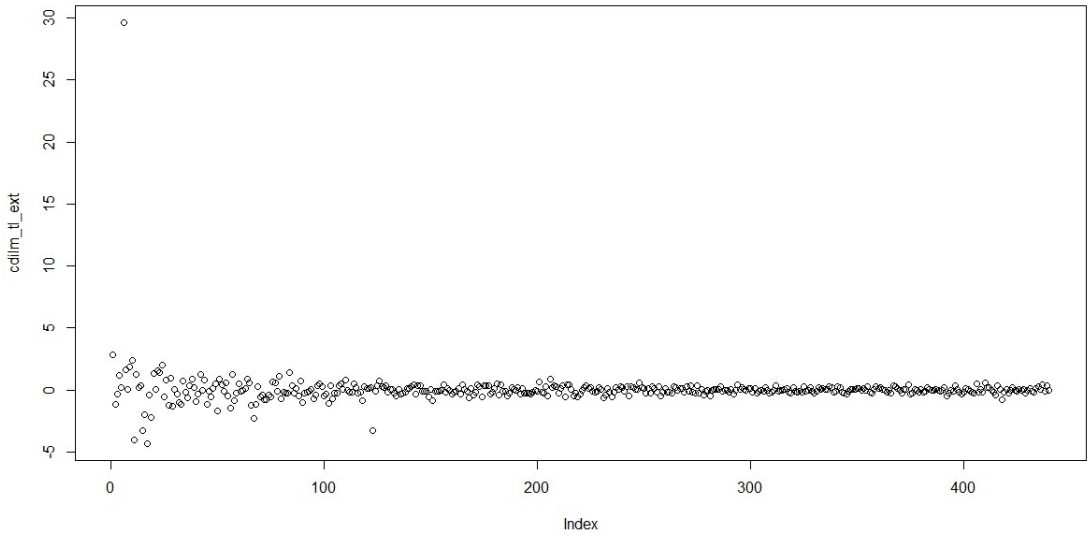
For the CDI dataset (see Appendix C.2 of the textbook for the description of the dataset), consider the regression model containing variables 6, 8, 9, 13, 14, and 15 in first order terms as predictor variables. The response variable is variable 10, the total serious crimes.

1. Obtain the variance inflation factors. Are there any indications that serious multicollinearity problems are present?

	VIF
X6	1.173570
X8	13.276326
X9	12.614763
X13	2.048616
X14	1.397569
X15	2.050460

The VIF for X8 and X9 are above the threshold of 10 to suggest that there is serious multicollinearity present with these predictors in the model.

2. Obtain the studentized deleted residuals and prepare a dot plot of these residuals. Are there any outliers present?



Yes, it looks as if there may be outliers present, particularly one extreme outlier.

3. Obtain the diagonal elements of the hat matrix. Using the rule of thumb (if hii > $3\bar{h}$, then consider xi extreme), identify any extreme X values.

First, we must calculate h_bar, which is given by the formula (p+1)/n, where p is the number of predictors and n is the number of observations. For this data set p = 6, and n = 440 which yields h_bar = ((6 + 1)/440) = 0.01590909. 3*h_bar is the threshold for considering Xi extreme, so we calculate 3*h_bar = 3*((6 + 1)/440) = 0.04772727. Using the rule of thumb if hii > 3*h_bar, consider Xi extreme. Below are the diagonal elements of the hat matrix with extreme values highlighted.

1	2	3	4	5	6	7
0.451018	0.194793	0.077921	0.03229	0.028572	0.051769	0.015575
8	9	10	11	12	13	14
0.111927	0.028947	0.018697	0.046606	0.050178	0.02624	0.006696
15	16	17	18	19	20	21
0.025115	0.043206	0.033653	0.011445	0.058681	0.00948	0.03834
22	23	24	25	26	27	28
0.012465	0.013936	0.016654	0.029156	0.005076	0.01503	0.008463
29	30	31	32	33	34	35
0.00999	0.010203	0.009637	0.055422	0.010892	0.02412	0.010755
36	37	38	39	40	41	42
0.036869	0.011184	0.004859	0.035016	0.021628	0.039991	0.032706
43	44	45	46	47	48	49
0.019007	0.0105	0.00993	0.013296	0.024462	0.105845	0.004811
50	51	52	53	54	55	56
0.031445	0.009451	0.016646	0.051847	0.013846	0.006205	0.014507
57	58	59	60	61	62	63
0.009731	0.026132	0.013106	0.004379	0.010095	0.011447	0.00496
64	65	66	67	68	69	70
0.021412	0.015619	0.006336	0.045926	0.015217	0.019956	0.033071
71	72	73	74	75	76	77
0.013178	0.023649	0.021635	0.007216	0.007348	0.023662	0.004076
78	79	80	81	82	83	84
0.006499	0.014631	0.01039	0.016397	0.013018	0.006138	0.006957
85	86	87	88	89	90	91
0.014564	0.008233	0.016072	0.004324	0.007384	0.0103	0.023404
92	93	94	95	96	97	98
0.00687	0.003275	0.004209	0.062961	0.018509	0.009034	0.015684
99	100	101	102	103	104	105
0.006825	0.010663	0.003799	0.013727	0.004815	0.009485	0.007647
106	107	108	109	110	111	112
0.011121	0.005253	0.007804	0.006135	0.007624	0.015553	0.012166
113	114	115	116	117	118	119
0.020075	0.006568	0.010215	0.005393	0.032164	0.012143	0.011484

120	121	122	123	124	125	126
0.005879	0.004969	0.004628	0.047401	0.007354	0.013243	0.006978
127	128	129	130	131	132	133
0.005749	0.112409	0.013486	0.012734	0.009979	0.00804	0.0312
134	135	136	137	138	139	140
0.006979	0.007397	0.005573	0.013364	0.004059	0.006814	0.020068
141	142	143	144	145	146	147
0.013346	0.009449	0.005461	0.004143	0.005826	0.005663	0.004792
148	149	150	151	152	153	154
0.010256	0.0076	0.010393	0.008207	0.014911	0.004662	0.008819
155	156	157	158	159	160	161
0.052343	0.005816	0.003635	0.010729	0.004831	0.010973	0.005073
162	163	164	165	166	167	168
0.004351	0.011436	0.005582	0.00523	0.008739	0.00487	0.02519
169	170	171	172	173	174	175
0.004862	0.01525	0.022661	0.012318	0.025477	0.004386	0.006633
176	177	178	179	180	181	182
0.016649	0.016622	0.00752	0.004897	0.014506	0.006799	0.009229
183	184	185	186	187	188	189
0.007031	0.017652	0.006316	0.007559	0.029851	0.083098	0.005774
190	191	192	193	194	195	196
0.006496	0.004684	0.012247	0.003833	0.003304	0.005048	0.021202
197	198	199	200	201	202	203
0.00576	0.004052	0.01047	0.004057	0.039731	0.014255	0.004842
204	205	206	207	208	209	210
0.006266	0.004003	0.074979	0.00584	0.00839	0.007015	0.006358
211	212	213	214	215	216	217
0.005627	0.008162	0.010943	0.010122	0.015136	0.00453	0.010031
218	219	220	221	222	223	224
0.005243	0.006302	0.004822	0.005047	0.01285	0.009313	0.006131
225	226	227	228	229	230	231
0.013528	0.007829	0.00347	0.009093	0.012287	0.009139	0.009548
232	233	234	235	236	237	238
0.007058	0.006107	0.006307	0.007923	0.007954	0.013855	0.004582
239	240	241	242	243	244	245
0.017355	0.003375	0.011433	0.005342	0.011056	0.007335	0.015468
246	247	248	249	250	251	252
0.018792	0.007013	0.016575	0.02207	0.007322	0.009048	0.008882
253	254	255	256	257	258	259
0.004611	0.006247	0.005721	0.006656	0.006574	0.014352	0.026809
260	261	262	263	264	265	266

0.005196	0.013091	0.032438	0.006339	0.004273	0.00354	0.00844
267	268	269	270	271	272	273
0.012272	0.012859	0.026604	0.007164	0.010282	0.045469	0.007889
274	275	276	277	278	279	280
0.008606	0.00655	0.00911	0.003874	0.005377	0.016363	0.006482
281	282	283	284	285	286	287
0.009469	0.006487	0.005981	0.005618	0.010036	0.005735	0.015036
288	289	290	291	292	293	294
0.007576	0.006737	0.004082	0.007029	0.011821	0.015131	0.017944
295	296	297	298	299	300	301
0.004081	0.004509	0.006433	0.009617	0.005826	0.005	0.033208
302	303	304	305	306	307	308
0.021237	0.074526	0.008434	0.004613	0.004715	0.005321	0.011127
309	310	311	312	313	314	315
0.005126	0.022869	0.010103	0.006728	0.015956	0.01771	0.008735
316	317	318	319	320	321	322
0.007894	0.004718	0.00542	0.011988	0.008734	0.007941	0.020426
323	324	325	326	327	328	329
0.004089	0.005217	0.015389	0.007464	0.006048	0.010979	0.008782
330	331	332	333	334	335	336
0.011583	0.01016	0.003998	0.005429	0.024425	0.008399	0.015778
337	338	339	340	341	342	343
0.082196	0.007741	0.011131	0.017466	0.004416	0.011177	0.004544
344	345	346	347	348	349	350
0.031773	0.008207	0.014759	0.008514	0.01267	0.005107	0.003455
351	352	353	354	355	356	357
0.00535	0.004847	0.007611	0.008093	0.008104	0.011066	0.046422
358	359	360	361	362	363	364
0.005999	0.007149	0.016747	0.005305	0.007738	0.065492	0.01332
365	366	367	368	369	370	371
0.007942	0.006433	0.008589	0.018782	0.00348	0.006655	0.0061
372	373	374	375	376	377	378
0.004873	0.006397	0.013905	0.010484	0.005897	0.00818	0.006445
379	380	381	382	383	384	385
0.00647	0.021236	0.005705	0.004364	0.010102	0.008348	0.010147
386	387	388	389	390	391	392
0.013791	0.009111	0.006092	0.010612	0.012003	0.004833	0.024034
393	394	395	396	397	398	399
0.011948	0.006114	0.011309	0.056877	0.006322	0.026798	0.019535
400	401	402	403	404	405	406
0.009261	0.008379	0.007985	0.003854	0.102675	0.04447	0.01711

407	408	409	410	411	412	413
0.013932	0.006823	0.016651	0.02913	0.005026	0.030584	0.005244
414	415	416	417	418	419	420
0.009126	0.062383	0.008855	0.011156	0.011844	0.006591	0.008648
421	422	423	424	425	426	427
0.008387	0.023031	0.006731	0.007072	0.012613	0.007321	0.006991
428	429	430	431	432	433	434
0.005909	0.014124	0.014826	0.016087	0.009497	0.029387	0.005452
435	436	437	438	439	440	
0.00624	0.030074	0.030126	0.015469	0.010969	0.003831	

According to the given metric, 1, 2 , 3, 6, 8, 12, 19, 32, 48, 53, 95, 128, 155, 188, 206, 303, 337, 363, 396, 404, and 415 are all extreme values.

4. Cases 2, 8, 48, 128, 206, and 404 are outlying with respect to their X values, and cases 2 and 6 are outlying with respect to their Y values. Obtain DFFITS, DFBETAS and Cook's distance values for these cases to assess their influence. What can you conclude?

We use the following metrics to determine if the given points are influential or not.

If abs(DFFITS) > (2 * (p+2)^(1/2))/(n-p-2), we consider the point influential.

This value is (2 * (6+2)^(1/2))/(440-6-2) = 0.01309457

If abs(DFBETAS) > 2/((n)^(1/2)), we consider the point influential.

This value is 2/((440)^(1/2)) = 0.0953462

If Cook's Distance > 0.5, the point is maybe influential. If Cook's Distance > 1, the point is very likely to be influential. We also examine Cook's Distance to see if any value is irregular compared to the others in the data. If a point is highly irregular it may be influential. See the chart below where points that meet these thresholds are highlighted.

Case	Abs(DFFITS)	Abs(DFBETAS)	Cook's Distance
2	0.5755949	0.05659868	0.04728957
6	6.921765	1.712836	2.263076
8	0.02167814	0.005598018	6.728933e-05
48	0.04879748	0.00428637	0.0003409421
128	0.07656083	0.02686773	0.0008392143
206	0.2438926	0.1101023	0.008502908
404	0.06173662	0.02697068	0.0005457055

Based on this information, we can conclude that 6 is clearly an outlier, it is much greater than the threshold values by a much larger magnitude than any other point. Abs(DFBETAS) also suggests that 206

is more influential than all other points, with the exception of 6. All the calculated values for Abs(DFFITS) break the threshold for being considered influential, but again note that 6 is far greater than any of the other values. 206 is also relatively large as well, as is 2. 2 does not stand out as much in the other statistics and may be worthy of further investigation.