Andrew Curtis

STAT 5100

Spring 2022

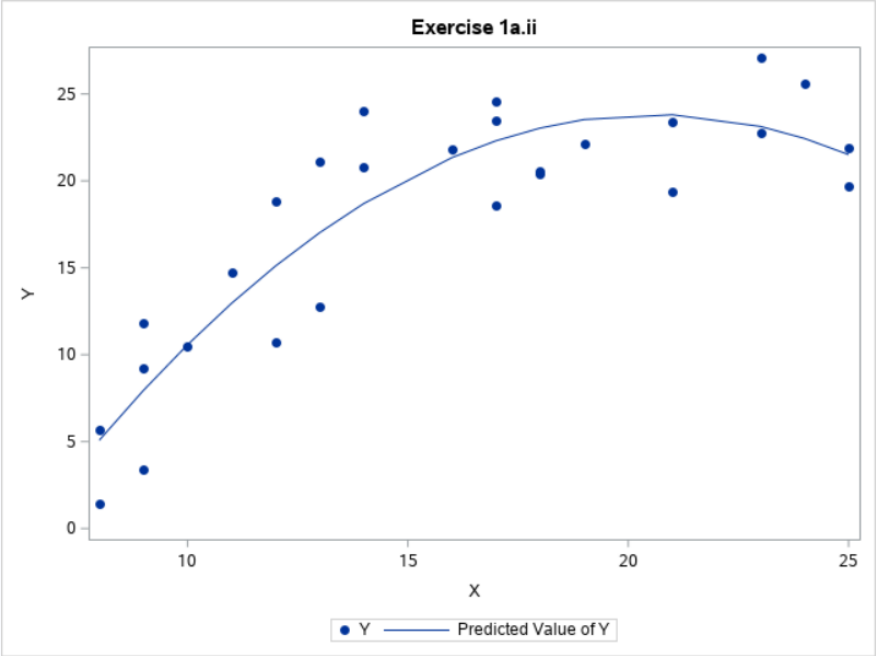
Homework #5

Exercise 1

a)

i) $Y = -26.32541 + 4.87357X - 0.11840X_{sq}$

ii)



iii) The quadratic regression function, as depicted by the line, appears to fit the data very well. Notice, Y vs X looks linear below X = 15 but for the higher values of X it seems like the quadratic function was needed.

iv) The R-squared value for the model is 0.8143.

v) The VIF for X and Xsq is 47.55625.

b)

i) Null hypothesis Ho: The linear and quadratic predictors are 0.
Alternative hypothesis: At least one of the linear or quadratic predictors are not 0.

ii) For the linear predictor, X, the p-value is less than 0.0001.

For the quadratic predictor, Xsq, the p-value is also less than 0.0001.

iii) For a = 0.01, the p-values for both the linear and quadratic predictors are significant. Thus, evidence to reject the null hypothesis and that both predictors are not 0.

c) There is evidence that both of the linear and quadratic predictors are not 0 for the model and should be included. Multicollinearity can reduce the precision of the model or affect the p-values for significance that we get for the predictors. So, potentially causes an issue with the model.

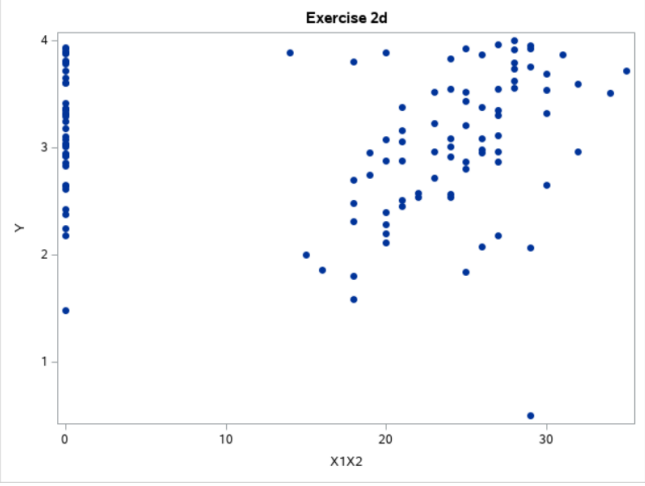
Exercise 2

a) Null The response variable of this model is grade point average. The X1 predictor is act score of the student. The new X2 predictor takes on values of 0 or 1. 0 for X2 indicates the student has not chosen a major of concentration at the time of at the time of their application submission. A value of 1 for X2 indicates they had selected a major of concentration.

b) $Y = 2.19842 + 0.03789X_1 - 0.09430X_2$

c) Null hypothesis Ho: The regression coefficients for the predictors are equal to 0.
Alternative hypothesis Ha: At least one of the regression coefficients is not equal to 0.
The p-value for X1 is 0.0038. The p-value for X2 is 0.4334.
Thus, the p-value for X1 is significant. Evidence to reject the null hypothesis in favor of the alternative. The X2 predictor is not significant and could be considered for dropping it from the model.

d) The residuals vs X1X2 may have some trend. Possibly helpful to include the interaction term.



Exercise 3

a) $Y = 3.22632 - 0.00276X_1 - 1.64958X_2 + 0.06224X_1X_2$

b) Null hypothesis Ho: The interaction coefficient is equal to zero.
Alternative hypothesis Ha: The interaction coefficient is significant and not equal to zero. The p-value for the interaction coefficient is equal to 0.0205 which is less than our significance level of a = 0.05. So, evidence to reject null hypothesis, the interaction term is significant. For the interaction term, if X2 is 0 then the term will have no effect, only the X1 coefficient will affect Y. However, if X2 is equal to 1, then both the X1 and X1X2 coefficients will affect Y. Overall this means that the effect of X1 on Y depends on the value of X2.

Exercise 4

If the a-to-enter value exceeded the a-to-remove value, then the model might continuously add new predictors instead of converging to a particular model with the forward and backward passes using the stepwise approach.

Exercise 5

a)

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
3	0.9560	0.9615	3.7274	73.8473	78.72282	X1 X3 X4
4	0.9555	0.9629	5.0000	74.9542	81.04859	X1 X2 X3 X4
2	0.9269	0.9330	17.1130	85.7272	89.38384	X1 X3
3	0.9247	0.9341	18.5215	87.3143	92.18984	X1 X2 X3
2	0.8661	0.8773	47.1540	100.8605	104.51716	X3 X4
3	0.8617	0.8790	48.2310	102.5093	107.38479	X2 X3 X4
3	0.8233	0.8454	66.3465	108.6361	113.51157	X1 X2 X4
2	0.7985	0.8153	80.5653	111.0812	114.73788	X1 X4
1	0.7962	0.8047	84.2465	110.4685	112.90629	X3
2	0.7884	0.8061	85.5196	112.2953	115.95191	X2 X3
2	0.7636	0.7833	97.7978	115.0720	118.72864	X2 X4
1	0.7452	0.7558	110.5974	116.0546	118.49234	X4
2	0.4155	0.4642	269.7800	137.7025	141.35916	X1 X2
1	0.2326	0.2646	375.3447	143.6180	146.05576	X1
1	0.2143	0.2470	384.8325	144.2094	146.64717	X2

For adjusted Rsquare, choosing the model with X1, X3, X4, it has the highest adjusted Rsquare value and is quite close to the four variable R-square value.

b)

For C_p, choosing the model with X1, X3, X4, it has the lowest single digit Cp value

c)

For AIC, choosing the model with X1, X3, X4, it has the lowest AIC value and is significantly less than the next reduced model.

d)

Backward Elimination: Step 1

Variable X2 Removed: R-Square = 0.9615 and C(p) = 3.7274

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8705.80299	2901.93433	175.02	<.0001
Error	21	348.19701	16.58081		
Corrected Total	24	9054.00000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-124.20002	9.87406	2623.35826	158.22	<.0001
X1	0.29633	0.04368	763.11559	46.02	<.0001
X3	1.35697	0.15183	1324.38825	79.87	<.0001
X4	0.51742	0.13105	258.46044	15.59	0.0007

Bounds on condition number: 2.8335, 19.764

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X2	3	0.0013	0.9615	3.7274	0.73	0.4038

For the backward elimination, I chose a significance value of 0.1. The result was a model with the X1, X3, and X4 predictors.

Appendix SAS code

```
/* Exercise 1 */

data steroid;
 input Y X @@; cards;
 27.1   23.0   22.1   19.0   21.9   25.0   10.7   12.0    1.4    8.0
 18.8   12.0   14.7   11.0    5.7    8.0   18.6   17.0   20.4   18.0
  9.2    9.0   23.4   21.0   10.5   10.0   19.7   25.0   11.8    9.0
 24.6   17.0    3.4    9.0   22.8   23.0   21.1   13.0   24.0   14.0
 21.8   16.0   23.5   17.0   19.4   21.0   25.6   24.0   12.8   13.0
 20.8   14.0   20.6   18.0
;

data steroid; set steroid;
 Xsq = X**2;
proc reg data=steroid;
  model Y = X Xsq / vif;
  testname: test X + Xsq;
  output out=out1 predicted=pred;
  title1 'Exercise 1';
run;

proc sort data=out1;
  by X;
proc sgplot data=out1;
  scatter x=X y=Y/
     markerattrs=(symbol=CIRCLEFILLED size=6pt);
  series  x=X y=pred/ lineattrs=(pattern=solid) ;
  title1 'Exercise 1a.ii';
run;
```

```
/* Exercise 2 */

data c1;   input Y X1 @@; cards;
 3.897 21 3.885 14 3.778 28 2.540 22 3.028 21 3.865 31 2.962 32 3.961 27 0.500 29 3.178 26
 3.310 24 3.538 30 3.083 24 3.013 24 3.245 33 2.963 27 3.522 25 3.013 31 2.947 25 2.118 20
 2.563 24 3.357 21 3.731 28 3.925 27 3.556 28 3.101 26 2.420 28 2.579 22 3.871 26 3.060 21
 3.927 25 2.375 16 2.929 28 3.375 26 2.857 22 3.072 24 3.381 21 3.290 30 3.549 27 3.646 26
 2.978 26 2.654 30 2.540 24 2.250 26 2.069 29 2.617 24 2.183 31 2.000 15 2.952 19 3.806 18
 2.871 27 3.352 16 3.305 27 2.952 26 3.547 24 3.691 30 3.160 21 2.194 20 3.323 30 3.936 29
 2.922 25 2.716 23 3.370 25 3.606 23 2.642 30 2.452 21 2.655 24 3.714 32 1.806 18 3.516 23
 3.039 20 2.966 23 2.482 18 2.700 18 3.920 29 2.834 20 3.222 23 3.084 26 4.000 28 3.511 34
 3.323 20 3.072 20 2.079 26 3.875 32 3.208 25 2.920 27 3.345 27 3.956 29 3.808 19 2.506 21
 3.886 24 2.183 27 3.429 25 3.024 18 3.750 29 3.833 24 3.113 27 2.875 21 2.747 19 2.311 18
 1.841 25 1.583 18 2.879 20 3.591 32 2.914 24 3.716 35 2.800 25 3.621 28 3.792 28 2.867 25
 3.419 22 3.600 30 2.394 20 2.286 20 1.486 31 3.885 20 3.800 29 3.914 28 1.860 16 2.948 28
;
data c2; input X2 @@; cards;
  0  1  0  1  0  1  1  1  1  0  0  1  1  1  0  1  1  0  0  1  1  0  1  0  1  0  0  1  1  1
  1  0  0  1  0  0  1  0  1  0  1  1  1  1  0  1  0  0  1  1  1  1  0  1  1  1  1  1  1  0
  0  1  0  0  0  1  0  0  1  1  0  1  1  1  1  0  1  1  1  1  0  1  1  0  1  0  1  1  0  1
  0  1  1  0  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  0  0  1  1  0  1  0  1  1  0
;
data college; merge c1 c2;
run;

proc reg data=college;
  model Y = X1 X2;
  output out=out1 residual=resid predicted=pred;
  title1 'Exercise 2';
run;
```

```
data out1; set out1;
  X1X2 = X1*X2;
proc sgplot data=out1;
  scatter  x=X1X2 y=Y
     / markerattrs=(symbol=CIRCLEFILLED size=6pt);
  title1 'Exercise 2d';
run;


/* Exercise 3 */

data college; set college;
  X1X2 = X1*X2;
proc reg data=college;
  model Y = X1 X2 X1X2 ;
  title1 'Exercise 3';
run;
```

```
/* Exercise 5 */
data job; input Y X1 X2 X3 X4 @@; cards;
   88.0   86.0  110.0  100.0   87.0   80.0   62.0   97.0   99.0  100.0
   96.0  110.0  107.0  103.0  103.0   76.0  101.0  117.0   93.0   95.0
   80.0  100.0  101.0   95.0   88.0   73.0   78.0   85.0   95.0   84.0
   58.0  120.0   77.0   80.0   74.0  116.0  105.0  122.0  116.0  102.0
  104.0  112.0  119.0  106.0  105.0   99.0  120.0   89.0  105.0   97.0
   64.0   87.0   81.0   90.0   88.0  126.0  133.0  120.0  113.0  108.0
   94.0  140.0  121.0   96.0   89.0   71.0   84.0  113.0   98.0   78.0
  111.0  106.0  102.0  109.0  109.0  109.0  109.0  129.0  102.0  108.0
  100.0  104.0   83.0  100.0  102.0  127.0  150.0  118.0  107.0  110.0
   99.0   98.0  125.0  108.0   95.0   82.0  120.0   94.0   95.0   90.0
   67.0   74.0  121.0   91.0   85.0  109.0   96.0  114.0  114.0  103.0
   78.0  104.0   73.0   93.0   80.0  115.0   94.0  121.0  115.0  104.0
   83.0   91.0  129.0   97.0   83.0
   ;
run;

proc reg data=job;
  model Y = X1 X2 X3 X4 / selection= AdjRSq Cp AIC SBC;
  title1 'Exercise 5abc';
run;


proc reg data=job;
  model Y = X1 X2 X3 X4 / selection=backward
                          slstay=0.1;
  title1 'Exercise 5d';
run;
```