Niranjan Poudel
STAT 5100
Spring 2019

Homework 5

Exercise 1.

a)

i) We have following table for the regression of Y on X and X^2.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-26.32541	5.88154	-4.48	0.0002	0
X	1	4.87357	0.77515	6.29	<.0001	47.55625
Xsq	1	-0.11840	0.02347	-5.05	<.0001	47.55625

From the above table we have,

$\beta_0 = -26.3254$

$\beta_1 = 4.8736$

$\beta_2 = -0.1184$

Therefore, our fitted regression function is $Y = -26.325 + 4.8736\ X - 0.1184X^2$.

ii) The following figure shows the scatterplot of the data (Y vs X) overlayed fitted regression function of predicted Y Vs X.

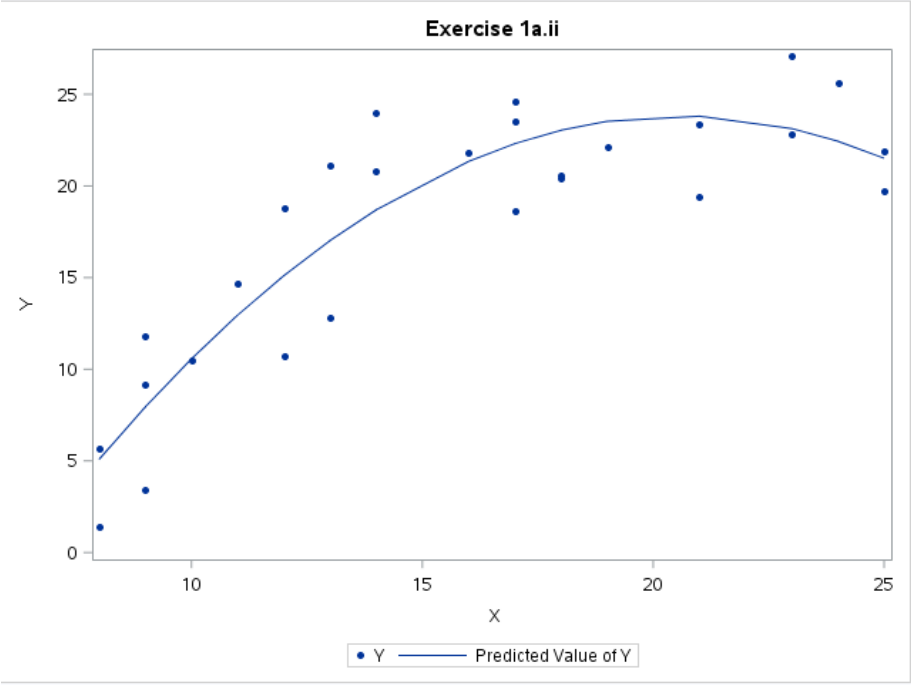

Fig 1: Scatterplot of Y Vs X overlayed with Predicted Y Vs X

iii) Looking at the above scatter plot the quadratic regression seems to fit the data X well.

iv) From the fit diagnostics for Y we have following observation;

Observations	27
Parameters	3
Error DF	24
MSE	9.9392
R-Square	0.8143
Adj R-Square	0.7989

From this observation we have $R^2 = 0.8143$

v) We have the following table for VIF.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-26.32541	5.88154	-4.48	0.0002	0
X	1	4.87357	0.77515	6.29	<.0001	47.55625
Xsq	1	-0.11840	0.02347	-5.05	<.0001	47.55625

So, we have VIF (X) = 47.556 and VIF(X^2) = 47.556.

b)

i) To test a regression relation, we have following null and alternate hypothesis.

Null hypothesis: $\beta_1 = \beta_2 = 0$

Alternate Hypothesis: $\beta_n \neq 0$ for at least one of n = 1 or 2.

ii) Following table shows the test statistic and p-value.

Test regrelation Results for Dependent Variable Y

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	523.13293	52.63	<.0001
Denominator	24	9.93920		

From above table we have p-value < 0.0001 and F* value = 52.63

iii) From above statistics we have P-value < 0.0001 which is less than $\alpha = 0.01$ so we can conclude alternate hypothesis that at least one of the linear or quadratic coefficient terms is non-zero. So, at least one of the term have effect in the predicted variable.

c) From the observation in b) and observation in {a). v} we can see that VIF (X) = 47.556 and VIF(X^2) = 47.556 both the VIF > 10 which suggest that there is multicollinearity. In observation

{b).} we have the predicted value is related to at least one of X or Xsq. The multicollinearity does not affect the predicted variable but may cause biasness in the interpretation of the individual coefficients. In the context of mentioned problems there is no effect of the result from multicollinearity in test of significance.

Exercise 2.

a. We have the regression equation as $E\{Y\} = \beta_0 + \beta_1 X1 + \beta_2 X2$
   - β_0 which is expected value of Y when $\beta_1 = \beta_2 = 0$ or when X1=X2=0.
   - β_1 which is expected change in the value of Y with unit increase in the value of X1 while all the other predictors(X2) are held constant. If the student has indicated a major field of concentration (X2=1) than $E\{Y\} = \beta_0 + \beta_1 X1 + \beta_2$ and if the major field was undecided (X2 =0), then $E\{Y\} = \beta_0 + \beta_1 X1$.
   - β_2 which is expected change in the value of Y with unit increase in the value of X2 while all other predictors (X1) are held constant. It measures the differential effect for whether the student had chosen a major field at the time of application. In general, it shows how higher/lower the mean prediction is for X2= 1 and X2 = 0.

b. We have the following parameter estimates

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.19842	0.33886	6.49	<.0001
X1	1	0.03789	0.01285	2.95	0.0038
X2	1	-0.09430	0.11997	-0.79	0.4334

Here, $\beta_0 = 2.1984$, $\beta_1 = 0.0379$, $\beta_2 = -0.094$.

From above table we have the estimated linear regression model as follow;

$$Y = 2.1984 + 0.0379\ X1 - 0.0943X2.$$

c. To test whether we can drop X2 variable from model or not we have

Null hypothesis: $\beta_2 = 0$

Alternate hypothesis: $\beta_2 \neq 0$.

The drop test for X2 gives the following statics.

Niranjan Poudel
STAT 5100
Spring 2019

Test dropX2test Results for Dependent Variable Y				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.24071	0.62	0.4334
Denominator	117	0.38955		

From the above table we have p-value = 0.4334 which is greater than α = 0.01 so we conclude that there is not enough evidence to support alternate hypothesis. This suggest towards the null hypothesis that X2 can be dropped given that X1 is already in model.

d. The plot of residuals against X1X2 is shown in the following figure;

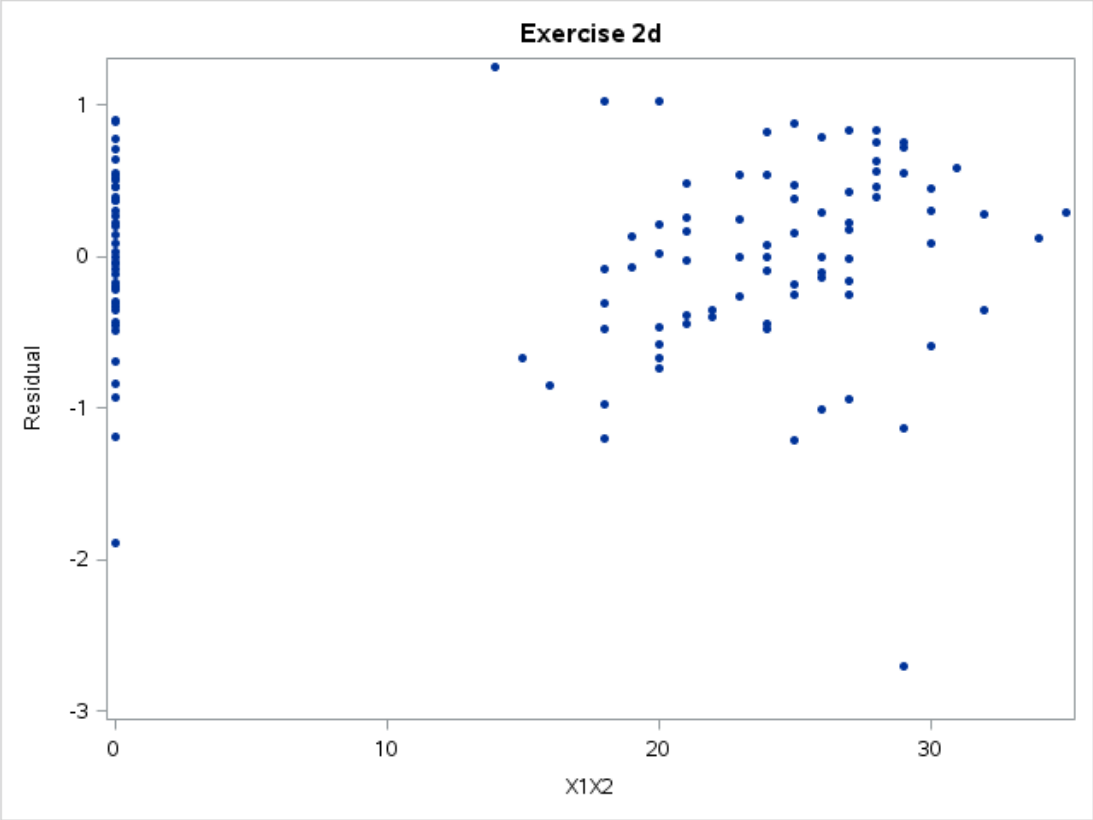

Figure 2: Plot of residuals against X1XZ2.

- From the above plot we do not see any proper evidence of X1X2 being helpful to add in the model.

Following are the plots of residuals vs X1 when X2= 0 and the when X2= 1.

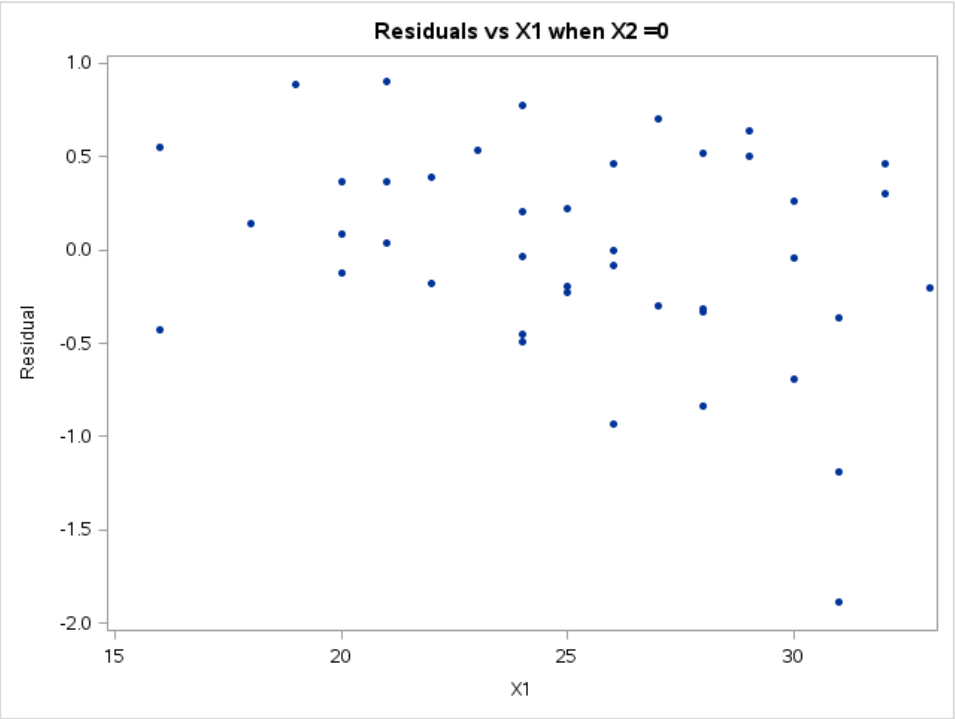

Figure 3: Plot of residuals vs X1 when X2 = 0.

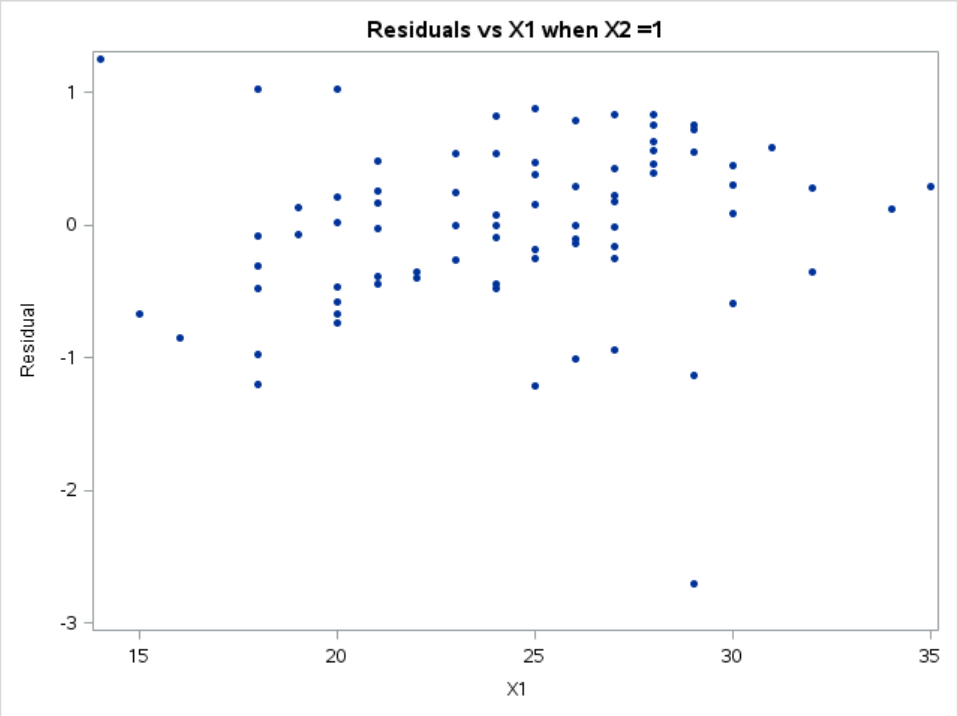

Figure 4: Plot of residuals vs X1 when X2 = 1.

Niranjan Poudel
STAT 5100
Spring 2019

Exercise 3.

a. We have the following table of parameter estimates from SAS.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.22632	0.54943	5.87	<.0001
X1	1	-0.00276	0.02141	-0.13	0.8977
X2	1	-1.64958	0.67220	-2.45	0.0156
X1X2	1	0.06224	0.02649	2.35	0.0205

From above table we have $\beta_0 = 3.2263$, $\beta_1 = -0.00276$, $\beta_2 = -1.6496$, $\beta_3 = 0.06224$

So, the estimated regression function is $Y = 3.2263 - 0.00276X1 - 1.6496X2 + 0.06224X1X2$.

b. We run the test for whether the interaction term X1X2 can be dropped from the model.

Null hypothesis: $\beta_3 = 0$

Alternate hypothesis: $\beta_3 \neq 0$

Test dropX1X2test Results for Dependent Variable Y				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	2.07126	5.52	0.0205
Denominator	116	0.37505		

We test the hypothesis for the $\alpha = 0.05$ so, we check if the p-value for the drop test is greater than or smaller than α. Here the p-value = 0.0205 < 0.05 so we have enough evidence to support alternative hypothesis.

- This suggest that the interaction term cannot be dropped from the model. The interaction term has positive effect to the predicted value. i.e. The predicted value increases with the increase in the value of the interaction term. In the given context the interaction term is either zero when X2 = 0 or X1 when X2 = 1. Using the regression function when X2 =0

$$Y = 3.2263 - 0.00276\ X1$$

And when X2 = 1 we get $Y = 3.2263 - 0.00276\ X1 - 1.64958 + 0.06224\ X1$

$$Y = 1.5767 + 0.05984\ X1$$

From above we can see that when interaction terms are present there is different effect of the X1 for the predicted Y at different context when X2 =0 it has the decreasing effect (-0.00276) in the predicted value and when X2 = 1 it has the increasing effect (+0.05984). This suggest the effect of X1 on Y depends on the value of X2.

Exercise 4.

The way in which the forward stepwise regression works is first we add the predictor to the model if its p-value is less than some threshold (α to enter value) and then again it will check if everything in a model is significant to some threshold (α to remove value), that is it does a backward check is the p-value is greater than the threshold. The process is to iterate to add and drop the variable from the model. In doing so the α to enter value for adding variables never should exceed the α to remove the value for deleting variables because it may cause some problems with the iterations. If, the α to enter value is greater than it will add the variable to the model and again in the next step it may drop the variable and again it may add the same variable with the iterative process which may cause an unending loop of adding and removing.

Exercise 5.

The best model selection table is as follow.

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

a. $R^2_{a,\,p}$

From the above table the model with best adjusted R-square is the model containing three variables X1, X3, X4 in the model with adjusted R-square 0.9560 which is greater than adjusted R-square for all other possible models. So the best model is with the predictors X1, X3, X4.

b. C_p

Here to select the term by Mallow C_p we look for model with smallest *p* such that C_p is nearly equal to *p*. From above table the C_p for the model with three predictors X1, X3, X4 is 3.7274 which is nearly equal to 4 (3+1: p+1) looking for the smallest p. So, we select the model with these variables.

c. AIC

From the above table the model with smallest AIC value is the model containing three variables X1, X3, X4 in the model with the AIC value of 73.8473. So, the model containing these three variables are best model.

d. Backward elimination

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

From the above table we can see Variable X2 removed from the model with the α level 0.01 because of the p-value is 0.4038 which greater than 0.01. So from backward elimination method the best model is the model containing the predictors X1, X3, X4.

- From all the methods above it seems better to select the model with predictors X1, X3, X4.

Niranjan Poudel
STAT 5100
Spring 2019

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
;
run ;

data steroid; set steroid;
 Xsq = X**2;
proc reg data=steroid;
  model Y = X Xsq / vif;
  regrelation: test X=Xsq=0;
  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=4pt);
  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  0  1  0  0  1  1  1  1  0  1
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  0  0  1
1  0  1  0  1  1  0
;
data college; merge c1 c2;
run;

proc reg data=college;
  model Y = X1 X2;
  dropX2test: test X2=0;
  output out=out2 residual=resid;
  title1 'Exercise 2';
run;
data out2; set out2;
       X1X2 = X1 * X2;
proc sgplot data= out2;
```

```
      scatter x= X1X2 y= resid /
      markerattrs=(symbol= CIRCLEFILLED size=4pt);
 title1 'Exercise 2d';
run;

proc sgplot data =out2;
      where X2 = 0;
      scatter x = X1 y =resid /
      markerattrs=(symbol= CIRCLEFILLED size=4pt);
title1 'Residuals vs X1 when X2 =0';
run;

proc sgplot data =out2;
      where X2 = 1;
      scatter x = X1 y =resid /
      markerattrs=(symbol= CIRCLEFILLED size=4pt);
title1 'Residuals vs X1 when X2 =1';
run;

/* Exercise 3 */

data college; set college;
  X1X2 = X1*X2;
proc reg data=college;
  model Y = X1 X2 X1X2 ;
  dropX1X2test: test X1X2 = 0;
  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;
  title1 'Exercise 5abc';
run;

proc reg data=job;
  model Y = X1 X2 X3 X4 / selection= backward
                          slstay=.10;
  title1 'Exercise 5d';
run;
```