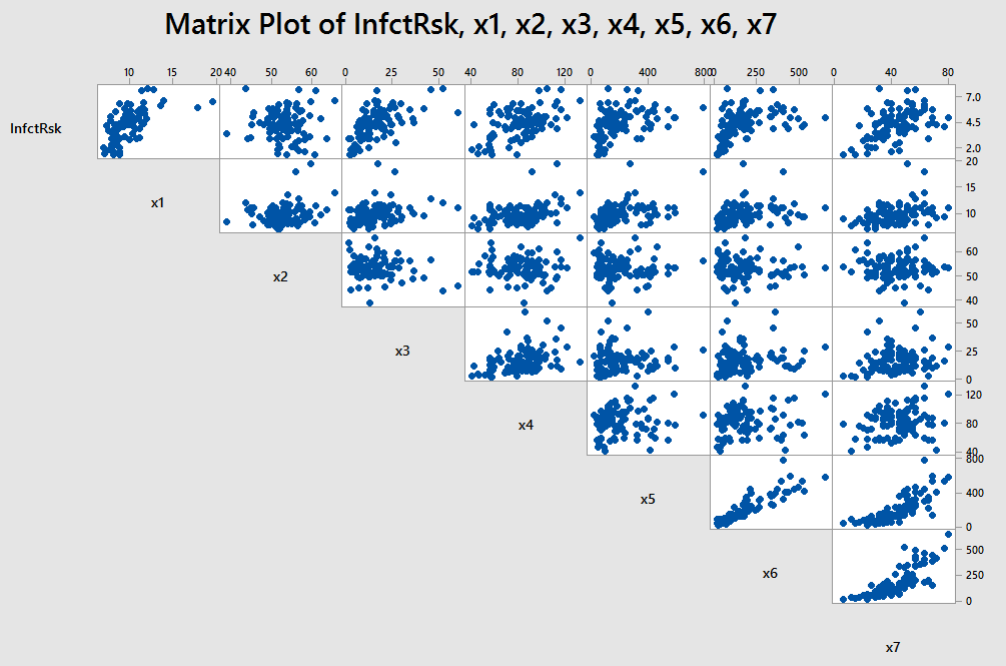
Homework 9

Roly Vicaría
STAT501 Fall 2015

November 8, 2015

0.0.1 Question 1

a) The SLR models based on each single predictor, as well as the matrix plot below, indicate that the best linear predictors for Y are x_1, x_3, x_4, and x_7.



b) The "best" stepwise model selected the same 4 predictors from part (a).

Regression Analysis: InfctRsk versus x1, x2, x3, x4, x5, x6, x7

Stepwise Selection of Terms

Candidate terms: x1, x2, x3, x4, x5, x6, x7

	Step 1		Step 2		Step 3		Step 4	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	0.005		0.097		-0.256		-0.742	
x1	0.4388	0.000	0.3337	0.000	0.2777	0.000	0.2416	0.000
x3			0.05825	0.000	0.05539	0.000	0.04893	0.000
x7					0.02179	0.002	0.02081	0.003
x4							0.01204	0.040
S	1.13508		0.973800		0.930099		0.913805	
R-sq	35.53%		53.05%		57.62%		59.53%	
R-sq(adj)	34.85%		52.05%		56.25%		57.77%	
R-sq(pred)	30.21%		48.26%		52.15%		53.70%	
Mallows' Cp	50.65		13.61		5.42		3.16	

c) Best subsets:

Best Subsets Regression: InfctRsk versus x1, x2, x3, x4, x5, x6, x7

Response is InfctRsk

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	x 1	x 2	x 3	x 4	x 5	x 6	x 7
1	35.5	34.8	30.2	50.6	1.1351	X						
1	34.7	34.0	30.7	52.6	1.1428			X				
2	53.0	52.0	48.3	13.6	0.97380	X		X				
2	46.3	45.1	40.9	28.7	1.0415			X		X		
3	57.6	56.3	52.2	5.4	0.93010	X		X				X
3	57.0	55.6	51.4	6.7	0.93657	X		X			X	
4	59.5	57.8	53.7	3.2	0.91381	X		X	X			X
4	59.3	57.5	53.1	3.6	0.91622	X		X	X			
5	60.0	57.8	52.8	4.1	0.91323	X		X	X		X	X
5	59.7	57.5	51.8	4.7	0.91659	X		X	X	X		X
6	60.1	57.4	50.7	6.0	0.91798	X		X	X	X	X	X
6	60.0	57.4	51.7	6.1	0.91823	X	X	X	X		X	X
7	60.1	56.9	49.6	8.0	0.92309	X	X	X	X	X	X	X

i. Of the two 4-predictor models listed, there was a very small difference in terms of their adjusted R^2 and C_p criterion. The better of the two was the model consisting of x_1, x_3, x_4, and x_7.

ii. Of the two 5-predictor models listed, there was again a very small difference in terms of their adjusted R^2 and C_p criterion. The better of the two was the model consisting of x_1, x_3, x_4, x_6, and x_7.

d) The test here is whether $\beta_6 = 0$. $H_0 : \beta_6 = 0, H_a : \beta_6 \neq 0$

$$\begin{aligned} df_R &= n - p = 97 - 5 = 92 \\ MSE_R &= S_R^2 = 0.91381^2 \approx 0.83505 \\ SSE_R &= MSE_R * (n - p) = 0.83505 * (97 - 5) \approx 76.8245 \end{aligned}$$

$$\begin{aligned} df_F &= n - p = 97 - 6 = 91 \\ MSE_F &= S_F^2 = 0.91323^2 \approx 0.833989 \\ SSE_F &= MSE_F * (n - p) = 0.833989 * (97 - 6) \approx 75.8930 \end{aligned}$$

$$\begin{aligned} F^* &= \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_F} \\ &= \frac{76.8245 - 75.8930}{92 - 91} \div \frac{75.8930}{91} \\ &= 1.11692 \end{aligned}$$

$$P(F > 1.11692) = 0.293384$$

Therefore, we fail to reject the null hypothesis and conclude that x_6 is not a significant predictor of Y and can be removed from the model that already contains x_1, x_3, x_4, and x_7.

e) i. Based on the stepwise and best subsets procedures, it looks like either x_1 or x_3 would be the most significant predictor. Out of the 13 models listed in the best subsets output, x_3 appears in 12 of them, whereas x_1 appears in 11 of them. x_1 was the strongest individual predictor, so if I had to choose one, I would say x_1 is the most significant.

ii. The "best" model that includes predictor x_6 is the model that consists of x_1, x_3, x_4, x_6, and x_7. It has the highest adjusted R^2 value and lowest S value. It also has a good C_p value less than p.

f) Some useful information found in the Stepwise procedure that is not found in the Best Subsets procedure, is the value of the coefficients and the p-value of each parameter at each step of the process. Therefore, you can see the effect of the addition (or removal) of another parameter on the p-value of a given parameter.

Some useful information found in the Best Subsets procedure that is not available in the Stepwise procedure is that it gives a "larger picture" of possible models that are worth evaluating. The Stepwise procedure seems a little more narrow in its output, where it basically gives a single "recommendation". The minitab output for Stepwise basically shows 4 candidate models, whereas the Best Subsets output shows 13.

g) The "best" model with interaction effects chosen by the Stepwise procedure was the model that consists of x_1, x_3, x_4, and x_7 plus the addition of the interaction term $x_3 * x_7$.

h) $C_p = p + \frac{(MSE_p - MSE_{all})(n-p)}{MSE_{all}} = 6 + \frac{(.7380 - .7514)(97-6)}{.7514} = 4.37716$
Based on this value of C_p, it is unbiased.

0.0.2 Question 2

a) Model A: $R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SSTO} = 1 - \left(\frac{99}{95}\right)\frac{1300}{5200} = 0.73947$

Model B: $R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SSTO} = 1 - \left(\frac{99}{94}\right)\frac{1210}{5200} = 0.75493$

b) Model A:
$AIC_p = n\ln(SSE_p) - n\ln(n) + 2p = 100\ln(1300) - 100\ln(100) + 2(5) = 266.4949$
$BIC_p = n\ln(SSE_p) - n\ln(n) + p\ln(n) = 100\ln(1300) - 100\ln(100) + 5\ln(100) = 279.5208$

Model B:
$AIC_p = n\ln(SSE_p) - n\ln(n) + 2p = 100\ln(1210) - 100\ln(100) + 2(6) = 261.3206$
$BIC_p = n\ln(SSE_p) - n\ln(n) + p\ln(n) = 100\ln(1210) - 100\ln(100) + 6\ln(100) = 276.9516$

c) $MSE_{all} = \frac{SSE_{all}}{n-p} = \frac{1150}{100-11} = 12.9213$

Model A: $C_p = \frac{SSE_p}{MSE_{all}} - (n - 2p) = \frac{1300}{12.9213} - (100 - 2 * 5) = 10.6091$

Model B: $C_p = \frac{SSE_p}{MSE_{all}} - (n - 2p) = \frac{1210}{12.9213} - (100 - 2 * 6) = 5.6438$

The C_p value of Model A is not desirable since it's double the value of p, it indicates bias. The C_p value of Model B is near to the value of p which is desirable.

d) Model B looks to be the preferable model across all criteria. It has a slightly higher adjusted R^2 value and slightly lower information criterion values. However, I think the C_p is the deciding factor in the comparison. The C_p value for Model A is almost double the number of parameters while it's near the value of p for Model B.

0.0.3 Question 3

a)

Vars	R-sq	R-sq (adj)	Mallows' C_p	S	1	2	3	4
1	68.411	67.753	2.082	2.413			X	
1	64.184	63.438	8.516	2.569		X		
2	70.332	69.070	1.157	2.363		X	X	
2	69.112	67.797	3.015	2.411			X	X
3	70.434	68.506	3.003	2.385	X	X	X	
3	70.366	68.434	3.106	2.387		X	X	X
4	70.434	67.806	5.003	2.411	X	X	X	X

b) From the table in part (a), we can see that the models are all very close in terms of their R-sq values and being unbiased models (with the exception of row 2). Starting from the model consisting of x_3, there was marginal improvement by adding x_2 to it. Without having any additional context about the purpose of the model, this seems like one of those situations where it would be best to opt for the simpler model since it doesn't look like there is much benefit from choosing a more complicated model. So from that point of view, I would conclude that the model consisting of the single predictor x_3, would be the "best" choice of those listed.