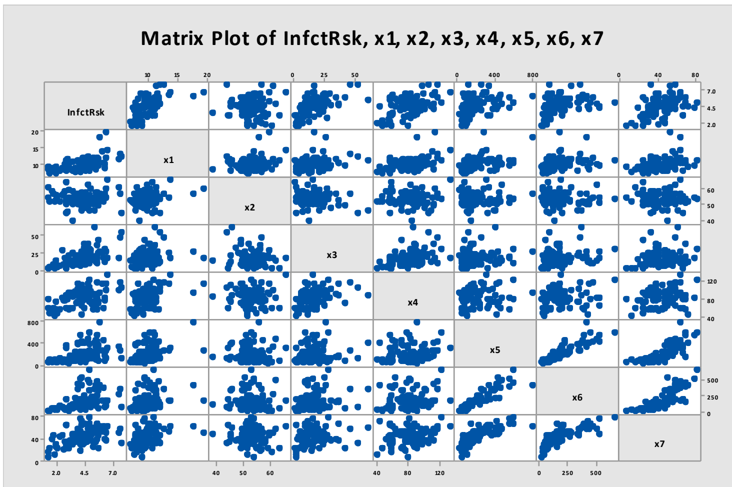
STAT 501 – Homework 9 Solutions – Fall 2015

1. **(8x6 = 48 points)** Use the "Infection Risk" dataset containing 97 observations, which relates Y=InfctRsk to 7 potential predictors: x1=Stay, x2=Age, x3 =Cult, x4= Xrays, x5=Census, x6=Nurses, and x7=Services.

   (a) Determine the four best linear predictors for Y in the context of simple linear regression. Use a Matrix Plot of Y, x1, x2, x3, x4, x5, x6, and x7, as well as SLR models based on each single predictor. The matrix plot can be obtained by the Minitab command sequence: Graph → Matrix Plot.



t-statistics and p-values of predictors from SLR models		
predictor	***t-statistic***	***p-value***
x1	***7.24***	***0.000***
x2	***-0.03***	***0.976***
x3	***7.10***	***0.000***
x4	***5.73***	***0.000***
x5	***4.53***	***0.000***
x6	***4.69***	***0.000***
x7	***4.92***	***0.000***

   Based on the plot and the t-statistic values, x1, x3, x4, x7 are the four best linear predictors.

   (b) Perform the stepwise procedure using all 7 predictors to determine the "best" model for predicting Y in the context of multiple linear regression. Use the Minitab command sequence: Stat → Regression → Regression → Fit Regression Model, click

"Stepwise," and select "Stepwise" for "Method" (leave everything at its default settings). For this dataset, does the "best" stepwise model match with your choice of best linear predictors in part (a)? [Note: In general it need not match.]

The "best" model contains the predictors x1, x3, x4, and x7. Its estimated regression equation is:

```
InfctRsk = -0.742 + 0.2416 x1 + 0.04893 x3 + 0.01204 x4
+ 0.02081 x7.
```

Yes, the "Best" model chosen here matches with the choice in part (a).

(c) Also perform the "Best Subsets" procedure using all 7 predictors. In Minitab, use Stat → Regression → Regression → Best Subsets and put all 7 predictors in the "Free predictors" box. Based on the adjusted R2 value and the Cp criterion, what is the "best" model consisting of:

i. 4 predictors: **x1, x3, x4, and x7**;
ii. 5 predictors: **x1, x3, x4, x6, and x7**.

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		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

(d) Use information from the Best Subsets procedure output in part (c) to test the hypothesis that x6 is not a significant predictor of Y upon controlling for x1, x3, x4, and x7.

H_0: $\beta_6 = 0$ vs H_a: $\beta_6 \neq 0$, where $E(Y) = \beta_0 + \beta_1x_1 + \beta_3x_3 + \beta_4x_4 + \beta_6x_6 + \beta_7x_7$ and $E(Y) = \beta_0 + \beta_1x_1 + \beta_3x_3 + \beta_4x_4 + \beta_7x_7$ are the full and reduced models respectively.
Then the test statistic is $F^* = [(SSE(R) – SSE(F))/1]/MSE(F)$
Now MSE(F)= s^2 = 0.91323^2 = 0.83399
SSE(F) = MSE(F)*(n-p) =0.83399*(97-6) = 75.8930
SSE(R) = (0.91381^2)*92=76.8245
F^* = (76.8245 - 75.8930) / 0.83399 =1.117 with a p-value = 0.293. Therefore, we fail to reject H_0 in favor of H_a and conclude that x6 is not a significant predictor of Y upon controlling for x1, x3, x4, and x7.

(e) Use output from either part (b) or part (c) to determine:

i. the most significant predictor out of x1, x3, x4, x7;

The output segment below is obtained from the stepwise procedure performed in part (b). I would say that x3 is the most significant predictor based on the t-values below.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.742	0.556	-1.33	0.185	
x1	0.2416	0.0571	4.23	0.000	1.37
x3	0.04893	0.00978	5.00	0.000	1.25
x4	0.01204	0.00578	2.08	0.040	1.37
x7	0.02081	0.00677	3.07	0.003	1.17

ii. the “best” model that includes predictor x6.

Based on the Best Subsets output in part (c), I would say that the model containing x1, x3, x4, x6, and x7 is the best model containing x6. Out of the models containing x6 such that $C_p < p$, this one has the highest adjusted R^2 value.

(f) Briefly describe any extra useful information that is provided by the Stepwise procedure that is not available in the Best Subsets procedure. Also, briefly describe any extra useful information that is provided by the Best Subsets procedure that is not available in the Stepwise procedure.

In addition to the fitted regression equation, the Stepwise procedure provides the t-values that can be used to rank the significance of predictors, as in part (e)(i). On the other hand, the Best Subsets procedure provides information about bias

for each model, identifies any underspecified models and also can be used to obtain information about a specific model consisting of certain predictor(s), as in part (e)(ii). In that sense, the Best Subsets procedure provides the user with a wider choice of models.

(g) What is the "best" model with interaction effects chosen by the Stepwise procedure? [Note: Choose your predictor candidate list based on your conclusion to part (d). Select only these predictors to be in the "Continuous predictors" box in the Regression dialog. Then click the "Model" button, highlight these predictors in the "Predictors" box, and click "Add" next to "Interactions through order 2" to add the interactions to the "Terms in model" box. Then click the "Stepwise" button to make sure all the main effect and interaction terms are included in "Potential terms" and click "Hierarchy" to make sure a hierarchical model is required at each step – see section 9.6 of the online notes.]

Based on the conclusion to part (d), the predictor candidate list should be x1, x3, x4, x7 and their interaction effects. The Best model from the Stepwise procedure contains the main effects x1, x3, x4, and x7 together with the interaction effect x3x7. The estimated regression equation is:

```
InfctRsk = -1.845 + 0.2528 x1 + 0.1310 x3 + 0.01012 x4
+ 0.04810 x7 - 0.001856 x3*x7
```

(h) Assume that a model that contains all the main effect and interaction terms considered as "Potential terms" in part (g) is unbiased. Based on this assumption, calculate Cp for the model in part (g) to determine if it is unbiased.

C_p = 67.157/0.7514 – (97–12) = 4.4 < p = 6, so this model is unbiased.

2. **(6 + 8 + 8 + 2 = 24 points)** Suppose we have a set of ten possible *x*-variables and that the model that with an intercept term and all ten *x*-variables has SSE = 1150. The sample size is n = 100 and SSTO = 5200. Consider two models that use subsets of the ten *x*-variables to predict *y*:
   - Model A with an intercept term and four *x*-variables with SSE = 1300.
   - Model B with an intercept term and five *x*-variables with SSE = 1210.

(a) Calculate R^2_{adj} for models A and B. (Remember that *p* is the number of regression parameters including the intercept!)

- **Model A:** $\boldsymbol{R^2_{adj} = \frac{\frac{SSTO}{n-1} - \frac{SSE}{n-p}}{\frac{SSTO}{n-1}} = \frac{\frac{5200}{100-1} - \frac{1300}{100-5}}{\frac{5200}{100-1}} = 0.739 \text{ or } 73.9\%.}$

- **Model B:** $R_{adj}^2 = \frac{\frac{SSTO}{n-1} - \frac{SSE}{n-p}}{\frac{SSTO}{n-1}} = \frac{\frac{5200}{100-1} - \frac{1210}{100-6}}{\frac{5200}{100-1}} = \mathbf{0.755}\ \boldsymbol{or}\ \mathbf{75.5}\%.$

(b) Calculate the AIC_p and BIC_p values for models A and B.

- **Model A: AICp = 100 × ln(1300) − 100 × ln(100) + 2 × 5 = 266.49 and BICp = 100 × ln(1300) − 100 × ln(100) + ln(100) × 5 = 279.52.**
- **Model B: AICp = 100 × ln(1210) − 100 × ln(100) + 2 × 6 = 261.32 and BICp = 100 × ln(1210) − 100 × ln(100) + ln(100) × 6 = 276.95.**

(c) Calculate the values of C_p for models A and B and indicate whether these values are desirable values for the C_p statistic.

- **Model A:** $\mathbf{C_p} = \frac{\mathbf{1300}}{\mathbf{1150/(100-11)}} - (\mathbf{100} - \mathbf{2 \times 5}) = \mathbf{10.6}$. **This is not a good value as it is greater than p = 5.**
- **Model B:** $\mathbf{C_p} = \frac{\mathbf{1210}}{\mathbf{1150/(100-11)}} - (\mathbf{100} - \mathbf{2 \times 6}) = \mathbf{5.6}$. **This is a good value as it is less than p = 6.**

(d) Which of models A and B do you prefer based on the results of parts (a), (b), and (c)? Explain.

Model B is preferable since it has a higher R^2_{adj}, lower AIC_p and BIC_p, and lower C_p.

3. **(25 + 3 = 28 points)** Suppose four x-variables are candidates to be in a model for predicting y. The sample size is n = 50 and SSTO = 884.8. SSE values for all possible models (including an intercept term) are given in the table below.

Model	SSE	Model	SSE	Model	SSE
X_1	452.2	X_1, X_2	297.9	X_1, X_2, X_3	261.6
X_2	316.9	X_1, X_3	277.7	X_2, X_3, X_4	262.2
X_3	279.5	X_1, X_4	370.7	X_1, X_2, X_4	297.6
X_4	389.3	X_2, X_3	262.5	X_1, X_3, X_4	273.2
		X_2, X_4	309.3	X_1, X_2, X_3, X_4	261.6
		X_3, X_4	273.3		

(a) Complete the following Best Subsets table.

Vars	R-Sq	R-Sq (adj)	Mallows Cp	S	X X X X 1 2 3 4
1	68.4	67.8	2.1	2.413	X (3)
1	**64.2**	**63.4**	**8.5**	**2.569**	**X** (2)
2	**70.3**	**69.1**	**1.2**	**2.363**	**X X** (2, 3)
2	**69.1**	**67.8**	**3.0**	**2.411**	**X X** (3, 4)
3	**70.4**	**68.5**	**3.0**	**2.385**	**X X X** (1, 2, 3)
3	**70.4**	**68.4**	**3.1**	**2.387**	**X X X** (2, 3, 4)
4	70.4	67.8	5.0	2.411	X X X X

(b) Describe in a few brief sentences the conclusions to be drawn from the results in part (a).

Candidate models based on Cp < p are (X_2, X_3), (X_1, X_2, X_3) and (X_2, X_3, X_4). Of these (X_2, X_3) has the lowest Cp and also the highest R-Sq (adj) and lowest S.