**STAT 501 – Homework 9 (covers Lesson 10) – Fall 2015 – Due Nov 1**

**Instructions**: Use Word to type your answers within this document. Then, submit your answers in the appropriate dropbox in ANGEL by the due date. The point distribution is located next to each question.

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.
   1. 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.



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model against SLR | X1 | X2 | X3 | X4 | X5 | X6 | X7 |
| S  R-sq  R-sq(adj) | **1.13508 35.53% 34.85%** | 1.41362 0.00% 0.00% | **1.14277 34.65% 33.96%** | **1.21890 25.65% 24.87%** | 1.28193 17.76% 16.90% | 1.27371 18.82% 17.96% | **1.26197 20.31% 19.47%** |

The four best linear predictors for Y in the context of simple linear regression are: X1, X3, X4 and X7

* 1. 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 minitab output is:

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

The “best” model for predicting Y in the context of multiple linear regressions is using X1, X3, X4 and X7. This is same result as in the first part.

* 1. 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:
     1. 4 predictors;
     2. 5 predictors.

The minitab output is:

Response is InfctRsk

R-Sq R-Sq Mallows x x x x x x x

Vars R-Sq (adj) (pred) Cp S 1 2 3 4 5 6 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

From the output we see:

|  |  |  |
| --- | --- | --- |
|  | Adjusted R2 value | Cp |
| 4 predictors | X1, X3, X4 and X7 | X1, X3, X4 and X7 |
| 5 predictors | X1, X3, X4, X6 and X7 | X1, X3, X4, X6 and X7 |

* 1. 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.

From the output in the above part, we see that when we add predictor x6 to a model that has predictors x1, x3, x4, and x7 the changes are as follows:

* R2 decreases only by 0.5 (60 – 59.5). This is a very small change.
* Even worse is that R2 (adj) remains the same at 57.8
* S only reduces by 0.0005

The above points clearly indicate that addition of X6 doesn’t improve the model significantly.

Alternatively, the minitab output we get on performing a regression is:

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant -0.566 0.580 -0.98 0.331

x1 0.2369 0.0573 4.14 0.000 1.38

x3 0.04788 0.00982 4.88 0.000 1.27

x4 0.01245 0.00579 2.15 0.034 1.38

x6 0.00115 0.00109 1.06 0.294 2.56

x7 0.0128 0.0101 1.27 0.209 2.62

The p-value is 0.294, which is greater than 0.05, so β6 in the above model is not significant at the 5% significance level.

* 1. Use output from either part (b) or part (c) to determine:

1. the most significant predictor out of x1, x3, x4, x7;
2. the “best” model that includes predictor x6.
3. ~~The most significant predictor is x1. This is clear from both parts b and c outputs. Part b shows that x1 is used as the step 1 and we know that stepwise regression begins with the best predictor. Part c shows that amongst the 2 models that use one predictor, x1 has higher values of R-Sq and R-Sq (adj) and lower value for Cp and S.~~

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. (x3 has higher t-value so more significant)



1. We look at output from part c, again:

Response is InfctRsk

R-Sq R-Sq Mallows x x x x x x x

Vars R-Sq (adj) (pred) Cp S 1 2 3 4 5 6 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

The best model from amongst the highlightes ones is with the predictors: x1, x3, x4, x6, x7 because amongst the highlightes models:

* + - * It has the smallest S
      * It has the highest R-Sq (adj)

The output from part b indicates that the “best” model has parameters: x1, x3, x4, and x7. Therefore any model that isn’t underspecified must have these parameters, again leading us to conclude that the “best” model that also contains X6 is the one using predictors: x1, x3, x4, x6, x7

* 1. 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.

Stepwise procedure does not assess all models but constructs a model by adding or removing one predictor at a time. On the other hand Best Subsets assesses all possible models and provides the best candidates.

Stepwise gives a single model at the end, which can be simpler to use. Best subsets provides more information by including more models, but can be more complex to choose the final.

Best Subsets assesses all possible models, and so large models may take a long time to process.

Stepwise may not select the model with the highest R2 value.

* 1. 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.]

The “best” model to be used from part d is the one with predictors: x1, x3, x4, x7

After the stepwise regression we get the “best” model as: x1, x3, x4, x7 and x3\*x7

The output is:

------Step 1------ -------Step 2------ -------Step 3------

Coef P Coef P Coef P

Constant -2.02 -3.63 -3.09

x1 0.546 0.000 0.335 0.008 0.2513 0.000

x3 0.1822 0.004 0.0935 0.142 0.04566 0.000

x1\*x3 -0.01194 0.047 -0.00454 0.444

x4 0.0387 0.013 0.0422 0.004

x7 0.0648 0.008 0.0703 0.003

x4\*x7 -0.000577 0.052 -0.000638 0.026

x3\*x7

S 0.958422 0.895927 0.893906

R-sq 55.00% 61.95% 61.70%

R-sq(adj) 53.55% 59.41% 59.59%

R-sq(pred) 48.57% 54.04% 55.52%

Mallows’ Cp 24.70 13.15 11.78

-------Step 4------ -------Step 5------

Coef P Coef P

Constant -2.47 -1.845

x1 0.2547 0.000 0.2528 0.000

x3 0.1217 0.000 0.1310 0.000

x1\*x3

x4 0.0197 0.224 0.01012 0.067

x7 0.0607 0.008 0.04810 0.000

x4\*x7 -0.000198 0.528

x3\*x7 -0.001668 0.006 -0.001856 0.000

* 1. 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.

The full model is:

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 28 147.377 5.26347 8.43 0.000

x1 1 0.337 0.33667 0.54 0.465

x2 1 1.851 1.85060 2.96 0.090

x3 1 0.001 0.00051 0.00 0.977

x4 1 0.485 0.48477 0.78 0.381

x5 1 0.211 0.21149 0.34 0.563

x6 1 0.022 0.02154 0.03 0.853

x7 1 1.141 1.14115 1.83 0.181

x1\*x2 1 1.120 1.12013 1.79 0.185

x1\*x3 1 1.102 1.10166 1.76 0.189

x1\*x4 1 2.443 2.44274 3.91 0.052

x1\*x5 1 1.108 1.10808 1.77 0.187

x1\*x6 1 0.971 0.97085 1.55 0.217

x1\*x7 1 0.243 0.24298 0.39 0.535

x2\*x3 1 0.024 0.02353 0.04 0.847

x2\*x4 1 0.004 0.00368 0.01 0.939

x2\*x5 1 0.123 0.12250 0.20 0.659

x2\*x6 1 0.131 0.13083 0.21 0.649

x2\*x7 1 0.622 0.62197 1.00 0.322

x3\*x4 1 0.021 0.02103 0.03 0.855

x3\*x5 1 0.969 0.96923 1.55 0.217

x3\*x6 1 1.333 1.33316 2.13 0.149

x3\*x7 1 1.755 1.75461 2.81 0.098

x4\*x5 1 1.132 1.13191 1.81 0.183

x4\*x6 1 0.200 0.20019 0.32 0.573

x4\*x7 1 0.476 0.47569 0.76 0.386

x5\*x6 1 0.382 0.38229 0.61 0.437

x5\*x7 1 0.465 0.46494 0.74 0.391

x6\*x7 1 0.337 0.33671 0.54 0.465

Error 68 42.465 0.62449

Total 96 189.842

The output for model from part (g) that includes x1, x3, x4, x7 and x3\*x7 is:

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 5 122.685 24.5371 33.25 0.000

x1 1 16.301 16.3008 22.09 0.000

x3 1 21.153 21.1526 28.66 0.000

x4 1 2.541 2.5409 3.44 0.067

x7 1 17.534 17.5344 23.76 0.000

x3\*x7 1 9.667 9.6666 13.10 0.000

Error 91 67.157 0.7380

Total 96 189.842



Cp = 6 + ((0.738 - 0.624) \* (97-6) / 0.624) = 22.625

Since Cp = 22.625 is much greater than p = 6, the bias is substantial.

1. **(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.

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

We have:



Model A: n = 100, p = 5: R-Sq (adj) = ((5200/99) - (1300/95)) / (5200/99) = 0.7395 or 73.95%

Model B: n = 100, p = 6: R-Sq (adj) = ((5200/99) - (1210/94)) / (5200/99) = 0.7549 or 75.49%

1. Calculate the AICp and BICp values for models A and B.



Model A: n = 100, p = 5: AICp = 100\*ln(1300)-100\*ln(100)+(2\*5) = 266.495

BICp = 100\*ln(1300)-100\*ln(100)+(5\*ln(100)) = 279.521

Model B: n = 100, p = 6: AICp = 100\*ln(1210)-100\*ln(100)+(2\*6) = 261.321

BICp = 100\*ln(1210)-100\*ln(100)+(6\*ln(100)) = 276.952

1. Calculate the values of Cp for models A and B and indicate whether these values are desirable values for the Cp statistic.

The Cp formulae is: 

MSEall = 1150/89= 12.92

Model A: n = 100, p = 5, MSE=13.68: Cp = 5+((13.68-12.92)\*(100-5)/12.92) = 10.6

Model B: n = 100, p = 6, MSE=12.87: Cp = 5+((12.87-12.92)\*(100-6)/12.92) = 5.6

The value for Model B is more desirable than Model B

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

On all the parameters we see that Model B performs better than the Model A. Model B has higher R-Sq(adj) and lower values for AICp, BICp and Cp

1. **(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 (which all include 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 |  | |

1. 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 |
| 1 | 64.18 | 63.44 | 8.5 | 2.569 | X |
| 2 | 70.33 | 69.07 | 1.2 | 2.363 | X X |
| 2 | 68.61 | 67.28 | 3.8 | 2.431 | X X |
| 3 | 70.43 | 68.51 | 3 | 2.385 | X X X |
| 3 | 70.37 | 68.43 | 3.1 | 2.387 | X X X |
| 4 | 70.4 | 67.8 | 5.0 | 2.411 | X X X X |

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

On all the parameters we see that the highlighted model performs better than the other non highlighted model with the same number of parameters. For instance Model with X3 has higher R-Sq(adj) and lower values for AICp, BICp and Cp when compared to model with predictor X2.

* Based on the R2-value criterion, the "best" model is the model with the two predictors x2 and x3. The addition of 3rd and 4th parameters have only increased R2 from 70.33 to 70.4x
* Based on the adjusted R2-value and MSE criteria, the "best" model is the same model as above with the predictors x2 and x3
* Based on the Cp criterion, there are two possible "best" models — the same model as above containing x2 and x3; the model containing x1, x2 and x3. Even the model with x2, x3 and x4 has a desirable Cp value.
  + It terms of relative comparison, the two predictor model with x2 and x3 have the most desirable Cp value