Stat/For/Hort 572 **Zhu** **March 19, 2002**

Midterm I

NAME: ____

Instructions:

1. For hypothesis testing, the significant level is set at $\alpha = 0.05$.
2. This exam is open book. You may use textbooks, notebooks, and a calculator.
3. Do all your work in the spaces provided. If you need additional space, use the back of the preceding page, indicating *clearly* that you have done so.
4. To get full credit, you must show your work. Partial credit will be awarded.
5. Do not dwell too long on any one question. Answer as many questions as you can.
6. Note that some questions have multiple parts. For some questions, these parts are independent, and so, for example, you can work on parts (b) or (c) separately from part (a).

For grader's use:

Question	Possible Points	Score
1	40	
2	36	
3	24	
Total	100	

1. Agricultural researchers are interested in the growth of Golden Acre cabbage. The following data are recorded on the height (y) in cm of one cabbage after the number of weeks (x) since the first observation. In total, there are $n = 8$ observations.

x:	0	1	2	3	4	5	6	7
y:	4.2	5.6	6.3	8.2	10.5	10.8	15.3	17.6

   (a) The simple linear regression (SLR) model $y = b_0 + b_1 x + e$ is fitted to the complete data set. The following is the corresponding studentized residual plot. Perform model diagnosis (i.e., check the four model assumptions and any potential outliers). For violation against any of the model assumptions, suggest possible remedies. (You do not need to carry out the remedies you are to suggest.)

```
MTB > plot residual yhat

Plot


          -                                                        A
          -   A
       1.0+
          -                                                A
  residual-          A
          -
          -
       0.0+
          -                                 A
          -
          -                  A      A
          -
      -1.0+
          -
          -
          -
          -                                         A
           +---------+---------+---------+---------+---------+------yhat
          2.5       5.0       7.5       10.0      12.5      15.0
```

(b) The largest residual corresponds to the observation $x = 5, y = 10.8$. Suppose this observation is deleted and the SLR is fitted to the reduced data set. Given that $\hat{y}_{\text{est}} = 13.03$, $\text{se}(\hat{y}_{\text{est}}) = 0.39$, and SSError = 3.50, all of which are based on the reduced data set, test formally whether this observation is an outlier.

(c) Consider fitting the following two models to the complete data set:

I SLR: $y = b_0 + b_1 x + e$

II Polynomial regression: $y = b_0 + b_1 x + b_2 x^2 + e$

Given that SSTotal = 156.39, $R^2 = 95.1\%$ for Model I, and $R^2 = 98.3\%$ for Model II, test formally H_o : Model I (i.e., $y = b_0 + b_1 x + e$) versus H_a : Model II (i.e., $y = b_0 + b_1 x + b_2 x^2 + e$).

2. In another experiment, the growth of three different cabbages (A, B, and C) at bi-weekly intervals are recorded. The data are tabulated below with $y =$ the height in cm and $x =$ the number of weeks since the first observation.

A		B		C	
x	y	x	y	x	y
0	3.2	0	6.3	0	3.9
2	8.7	2	10.4	2	10.4
4	11.1	4	14.1	4	18.9
6	16.1	6	16.4	6	26.0

The following (edited) Minitab printouts may be of use:

```
MTB > print w1-w5 y

Data Display

 Row    w1    w2    w3    w4    w5      y

   1     0     1     0     0     0    3.2
   2     2     1     0     2     0    8.7
   3     4     1     0     4     0   11.1
   4     6     1     0     6     0   16.1
   5     0     0     1     0     0    6.3
   6     2     0     1     0     2   10.4
   7     4     0     1     0     4   14.1
   8     6     0     1     0     6   16.4
   9     0     0     0     0     0    3.9
  10     2     0     0     0     0   10.4
  11     4     0     0     0     0   18.9
  12     6     0     0     0     0   26.0

MTB > regress y 5 w1 w2 w3 w4 w5

Regression Analysis: y versus w1, w2, w3, w4, w5

The regression equation is
y = 3.58 + 3.74 w1 + 0.030 w2 + 3.12 w3 - 1.68 w4 - 2.04 w5

Predictor        Coef     SE Coef          T        P
Constant       3.5800      0.6124       5.85    0.001
w1             3.7400      0.1637      22.85    0.000
w2             0.0300      0.8661       0.03    0.973
w3             3.1200      0.8661       3.60    0.011
w4            -1.6850      0.2315      -7.28    0.000
w5            -2.0400      0.2315      -8.81    0.000

S = 0.7320      R-Sq = 99.3%     R-Sq(adj) = 98.8%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         5     473.147      94.629    176.60    0.000
Residual Error     6       3.215       0.536
Total             11     476.362

Source       DF      Seq SS
w1            1     374.500
w2            1      33.135
w3            1      18.000
w4            1       5.896
w5            1      41.616
```

```
MTB > regress y 3 w1 w4 w5

Regression Analysis: y versus w1, w4, w5

The regression equation is
y = 4.63 + 3.51 w1 - 1.68 w4 - 1.37 w5

Predictor        Coef     SE Coef          T        P
Constant       4.6300      0.6013       7.70    0.000
w1             3.5150      0.2104      16.70    0.000
w4            -1.6786      0.2353      -7.14    0.000
w5            -1.3714      0.2353      -5.83    0.000

S = 1.245       R-Sq = 97.4%     R-Sq(adj) = 96.4%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3      463.96      154.65     99.80    0.000
Residual Error     8       12.40        1.55
Total             11      476.36

Source       DF      Seq SS
w1            1      374.50
w4            1       36.80
w5            1       52.66
```

Consider the model $y = b_0 + b_1 w_1 + b_2 w_2 + b_3 w_3 + b_4 w_4 + b_5 w_5 + e$.

(a) Briefly interpret b_3 and test $H_0 : (b_3 = 0 | b_0, b_1, b_2, b_4, b_5)$.

(b) Test whether the intercepts of the regression lines relating height to time for the three cabbages are equal (without assuming that the slopes are equal).

(c) Test whether the slopes of the regression lines relating height to time for the three cabbages are equal (without assuming that the intercepts are equal).

3. It is of interest to study the possible relationships between a dependent variable y and three independent variables x_1, x_2, and x_3, based on $n = 30$ observations. The following (edited) Minitab printouts may be of use:

```
MTB > regress y 1 x1

The regression equation is
y = 0.94 + 0.898 x1

Predictor        Coef     SE Coef          T        P
Constant        0.938       1.440       0.65    0.520
x1            0.89840     0.08122      11.06    0.000

S = 3.801       R-Sq = 81.4%     R-Sq(adj) = 80.7%

MTB > regress y 1 x2

The regression equation is
y = 2.92 + 0.762 x2

Predictor        Coef     SE Coef          T        P
Constant        2.922       1.627       1.80    0.083
x2            0.76207     0.08863       8.60    0.000

S = 4.617       R-Sq = 72.5%     R-Sq(adj) = 71.5%

MTB > regress y 1 x3

The regression equation is
y = 10.1 + 0.274 x3

Predictor        Coef     SE Coef          T        P
Constant       10.136       1.936       5.23    0.000
x3            0.27449     0.08015       3.42    0.002

S = 7.394       R-Sq = 29.5%     R-Sq(adj) = 27.0%

MTB > regress y 2 x1 x2

The regression equation is
y = 0.95 + 0.751 x1 + 0.145 x2

Predictor        Coef     SE Coef          T        P
Constant        0.953       1.449       0.66    0.516
x1             0.7506      0.2022       3.71    0.001
x2             0.1452      0.1817       0.80    0.431

S = 3.826       R-Sq = 81.8%     R-Sq(adj) = 80.5%

MTB > regress y 2 x1 x3

The regression equation is
y = 0.94 + 0.895 x1 + 0.0030 x3

Predictor        Coef     SE Coef          T        P
Constant        0.940       1.467       0.64    0.527
x1             0.8949      0.1032       8.67    0.000
x3            0.00298     0.05235       0.06    0.955

S = 3.870       R-Sq = 81.4%     R-Sq(adj) = 80.0%

MTB > regress y 2 x2 x3

The regression equation is
```

```
y = 2.87 + 0.802 x2 - 0.0334 x3

Predictor        Coef     SE Coef          T        P
Constant        2.869       1.654       1.73    0.094
x2             0.8023      0.1225       6.55    0.000
x3           -0.03342     0.06918      -0.48    0.633

S = 4.681       R-Sq = 72.8%     R-Sq(adj) = 70.7%

MTB > regress y 3 x1 x2 x3

The regression equation is
y = 0.94 + 0.746 x1 + 0.169 x2 - 0.0168 x3

Predictor        Coef     SE Coef          T        P
Constant        0.939       1.475       0.64    0.530
x1             0.7458      0.2064       3.61    0.001
x2             0.1693      0.2027       0.84    0.411
x3           -0.01677     0.05771      -0.29    0.774

S = 3.892       R-Sq = 81.9%     R-Sq(adj) = 79.8%
```

(a) Perform a backward elimination to choose the best model. Give all the steps in the backward elimination.

(b) The following italicized statement is either True or False. Indicate whether the statement is true or false. Give a clear justification of your answer.
For this data set, the model chosen by a forward selection is the same as the model chosen by a stepwise procedure.