Stat/For/Hort 572 **March 9, 2004**

Midterm I

NAME: _____

Instructions:

1. For hypothesis testing, follow the standard procedures. That is, state the H_0 and the H_A, select and compute the test statistic, report the p-value, draw a conclusion at the $\alpha = 0.05$ level, and interpret the testing result.
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	30	
2	20	
3	50	
Total	100	

1. A researcher in forestry is interested in the relationship between diameter at breast height (DBH) and height of trees. The following data are recorded on the DBH in inches (x) and height in feet (y) of 8 trees.

x:	5.5	5.5	6.0	6.0	6.5	6.5	7.0	7.0
y:	50	54	56	60	58	68	66	72

   A simple linear regression (SLR) model $y = b_0 + b_1 x + e$ is fitted to the data set and some of the R code and output are as follows.

```
> x = c(5.5, 5.5, 6.0, 6.0, 6.5, 6.5, 7.0, 7.0);
> y = c(50, 54, 56, 60, 58, 68, 66, 72);
> tree.lm = lm(y~x);
> summary(tree.lm);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.500     14.885  -0.638  0.54689
x             11.200      2.372   4.722  0.00325 **
---

Residual standard error: 3.751 on 6 degrees of freedom
Multiple R-Squared: 0.7879,     Adjusted R-squared: 0.7526
F-statistic: 22.29 on 1 and 6 DF,  p-value: 0.003253

> anova(tree.lm);

Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value   Pr(>F)
x          1 313.600 313.600  22.294 0.003253 **
Residuals  6  84.400  14.067
---

> predict(tree.lm, data.frame(x=6.2),se.fit=TRUE);

$fit
[1] 59.94

$se.fit
[1] 1.331315

> tree.aov = lm(y~factor(x));
> anova(tree.aov);

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
factor(x)  3 314.00  104.67  4.9841 0.07739 .
Residuals  4  84.00   21.00
---
```

(a) Construct a 95% confidence interval of the expected height for trees of 6.2 inches in DBH.

(b) Perform a lack of fit test (LOF) for the SLR model $y = b_0 + b_1 x + e$.

2. A study is aimed at comparing the height of 20 seedlings of 4 species of pines (A, B, C, and D). Consider the model $y = b_0 + b_1 w_1 + b_2 w_2 + b_3 w_3 + e$. The first 5 rows of observations correspond to species A, the next 5 rows to species B, then the next 5 rows to species C, and the last 5 rows to species D. The following R output may be of use.

```
> seedling;
   w1 w2 w3    y
1   0  0  0 35.0
2   0  0  0 41.5
3   0  0  0 33.3
4   0  0  0 52.8
5   0  0  0 42.6
6   1  0  0 57.1
7   1  0  0 67.0
8   1  0  0 66.2
9   1  0  0 56.6
10  1  0  0 60.1
11  0  1  0 37.0
12  0  1  0 27.8
13  0  1  0 17.1
14  0  1  0  6.3
15  0  1  0 34.8
16  0  0  1 32.3
17  0  0  1 29.7
18  0  0  1 38.4
19  0  0  1 36.5
20  0  0  1 32.0

> seedling.lm = lm(y~w1+w2+w3, data=seedling);
> summary(seedling.lm);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   41.040      3.614  11.356 4.56e-09 ***
w1            20.360      5.111   3.984  0.00107 **
w2           -16.440      5.111  -3.217  0.00539 **
w3            -7.260      5.111  -1.420  0.17466
---

> anova(seedling.lm);
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
w1         1 2994.85 2994.85 45.8595 4.484e-06 ***
w2         1  546.99  546.99  8.3759   0.01057 *
w3         1  131.77  131.77  2.0177   0.17466
Residuals 16 1044.88   65.31
```

```
---
```

(a) Briefly interpret b_3 and formally test $H_0 : [b_3 = 0|b_0, b_1, b_2]$.

(b) Suppose that the expected tree heights are the same for species A and D, perform a test to determine whether the tree heights are the same for species A and C.

3. A research assistant lost an original data set that consisted of a dependent variable y and 3 independent variables x_1, x_2, x_3. However some of the R output has been recovered as follows.

```
> cor(lost);
            x1          x2          x3          y
x1  1.00000000 -0.07171971 -0.05350326  0.6507951
x2 -0.07171971  1.00000000  0.95493823 -0.5421249
x3 -0.05350326  0.95493823  1.00000000 -0.6726394
y   0.65079505 -0.54212492 -0.67263936  1.0000000

> lost.lm = lm(y~x3+x2+x1,data=lost);
> summary(lost.lm);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.1627    14.0706   1.433    0.167
x3           -3.9143     0.2341 -16.723 1.30e-13 ***
x2            2.8453     0.2471  11.517 1.55e-10 ***
x1            2.1125     0.1092  19.349 7.26e-15 ***
---

Residual standard error: 2.531 on 21 degrees of freedom
Multiple R-Squared: 0.977,Adjusted R-squared: 0.9737
F-statistic:   297 on 3 and 21 DF,  p-value: < 2.2e-16

> anova(lost.lm);
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x3         1 2643.38 2643.38  412.58 2.752e-15 ***
x2         1  665.93  665.93  103.94 1.380e-09 ***
x1         1 2398.60 2398.60  374.37 7.265e-15 ***
Residuals 21  134.55    6.41
---
```

(a) Consider the model $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e$. Test whether there is an overall effect of all the independent variables on y.

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

(c) Consider the model $y = b_0 + b_1x_1 + b_3x_3 + e$. Test for $H_0 : [b_1 = 0|b_0, b_3]$.

(d) Perform a forward selection to choose the best model. Give all the steps in the forward selection.