STAT 512   Midterm Exam July 7 2006

Name:

Show all your work (you can use this page).

All questions are for 5 points (100 total).

Good luck !

1. Formulate the simple linear regression model. Clearly state all the assumptions.

$Y_i = \beta_0 + \beta_1 X_i + \xi_i$, , $1 \le i \le n$, ξ_i 's independent $N(0,\sigma^2)$

2. You use your data to estimate parameters of the simple linear regression model. Your estimators based on n=120 observations are $b_0 = 2$, $b_1 = 0.04$, $s(b_1)=0.01$, MSE=0.4. You also computed the average of the response variable $\bar{Y}=3$.

a) Test if $\beta_1 = 0$ (Compute the test statistic, give the number of degrees of freedom, corresponding critical value and conclusion).

$t= b_1/ s(b_1) = 4$, df=118, $t^*=1.98$

$t> t^*$, reject H_0, at the significance level $\alpha=0.05$ we have enough evidence to conclude that Y is associated with X

b) Predict the value of Y for X=20.

$$\hat{Y} = 2 + 0.04 \cdot 20 = 2.8$$

c) Calculate $\bar{X}$ (Hint – see the formula for b_1).

$$\bar{X} = \frac{\bar{Y} - b_0}{b_1} = \frac{3-2}{0.04} = 25$$

d) Calculate $SSX = \sum (X_i - \bar{X})^2$ (Hint – see the formula for $s^2(b_1)$).

$$SSX = \frac{s^2}{s^2[b_1]} = \frac{0.4}{0.0001} = 4000$$

e) Estimate the variance of the error of the prediction you made in point b)

$$s^2[pred] = 0.04\left(1 + \frac{1}{120} + \frac{25}{4000}\right) = 0.4058$$

f) Construct the corresponding 95% prediction interval.

$$2.8 \pm 1.98\sqrt{0.4058} \approx [1.53, 4.07]$$

3. Here is the table of type I sums of squares for three explanatory variables used for the multiple regression model.

variable	Type I Sum of Squares
X1	250
X2	30
X3	20

SSE for the full model is equal to 550 and dfE=30.

a) How many cases do you have in your data file ?

34

b) Give the estimate of the standard deviation of the error term.

$$s = \sqrt{\frac{550}{30}} \approx 4.28$$

c) Test the hypothesis that the response variable is not associated with any of the explanatory variables.

$$F = \frac{300/3}{550/30} \approx 5.45$$

$$F^{*}(3,30) = 2.92$$

$$F > F^{*}$$

Reject H_0. At the significance level α=0.05 we have enough evidence to conclude that at least one explanatory variable is associated with Y.

d) Give the value of R^2 for the full model.

$R^2 = 300/850 \approx 0.35$

e) Compute the sample correlation coefficient between Y and X1.

$$r = \pm\sqrt{250/850} = \pm 0.54$$

3. We study the relation between selling price (Y) and assessed valuation of one-family residential dwellings (X). We also consider an additional explanatory variable – lot location (Z). Z is coded as 1 for dwellings located on corner lots and 0 in other case.

We build a regression model taking into account the possible interaction between X and Z (int=X · Z). Below you can find the SAS output from this analysis.

```
                    Dependent Variable: y

                     Analysis of Variance

                        Sum of         Mean
Source             DF      Squares       Square   F Value   Pr > F

Model               3    4237.05022    1412.35007     93.21   <.0001
Error              60     909.10463      15.15174
Corrected Total        63    5146.15484


     Root MSE           3.89252    R-Square     0.8233
     Dependent Mean      79.02344    Adj R-Sq     0.8145
     Coeff Var           4.92578


                     Parameter Estimates

                 Parameter      Standard
  Variable    DF      Estimate         Error    t Value    Pr > |t|

  Intercept    1    -126.90517      14.72247      -8.62      <.0001
  x           1       2.77590       0.19628      14.14      <.0001
  z           1      76.02153      30.13136       2.52      0.0143
  int          1      -1.10748       0.40554      -2.73      0.0083
```

a) Give the result of the overall ANOVA test for significance of any of the explanatory variables.

F=93.21, p-value<0.0001, reject H_0 , we have enough evidence to conclude that at least one explanatory variable is associated with Y.

b) Give the estimate of the standard deviation of the error term.

S=3.8925

c) Is there a significant difference between the slopes of dependence of Y on X for dwellings located on corner and non corner lots ?

test for the coefficient by int, p-value=0.0083, reject H_0 , we have enough evidence to conclude that the slopes of dependence of Y on X are different for dwellings located on corner and non corner lots.

d) Give the estimated equation describing the dependence of Y on X for dwellings located on corner lots.

$$\hat{Y} = -50.9 + 1.67X$$

4. You try to relate the job proficiency score (Y) to the results of four tests (X1, X2, X3, X4). Below you have the results of the ANOVA table for this analysis as well the table with type I and type II sums of squares.

```
                         Analysis of Variance

                              Sum of         Mean
Source              DF       Squares       Square   F Value    Pr > F

Model                4    8718.02248    2179.50562    129.74    <.0001
Error               20     335.97752      16.79888
Corrected Total     24    9054.00000
```

```
Variable        Type I SS    Type II SS

x2         1   2236.47088      12.21949
x1         1   1966.34919     759.83030
x3         1   4254.45924    1064.15000
x4         1    260.74317     260.74317
```

a) Test for significance of X2 in the full model .

$$F = \frac{12.2}{16.8} \approx 0.73$$
$$F^*(1,20) = 4.35$$
$$F < F^*$$

Do not reject H_0 , X2 is not a significant predictor in a full model.

b) Test for significance of X2 in the simple regression model.

$$F = \frac{2236.5}{(9054 - 2236.5)/23} \approx 7.54$$
$$F^*(1,23) = 4.26$$
$$F > F^*$$

reject H_0 , we have enough evidence to conclude that Y is associated with X2

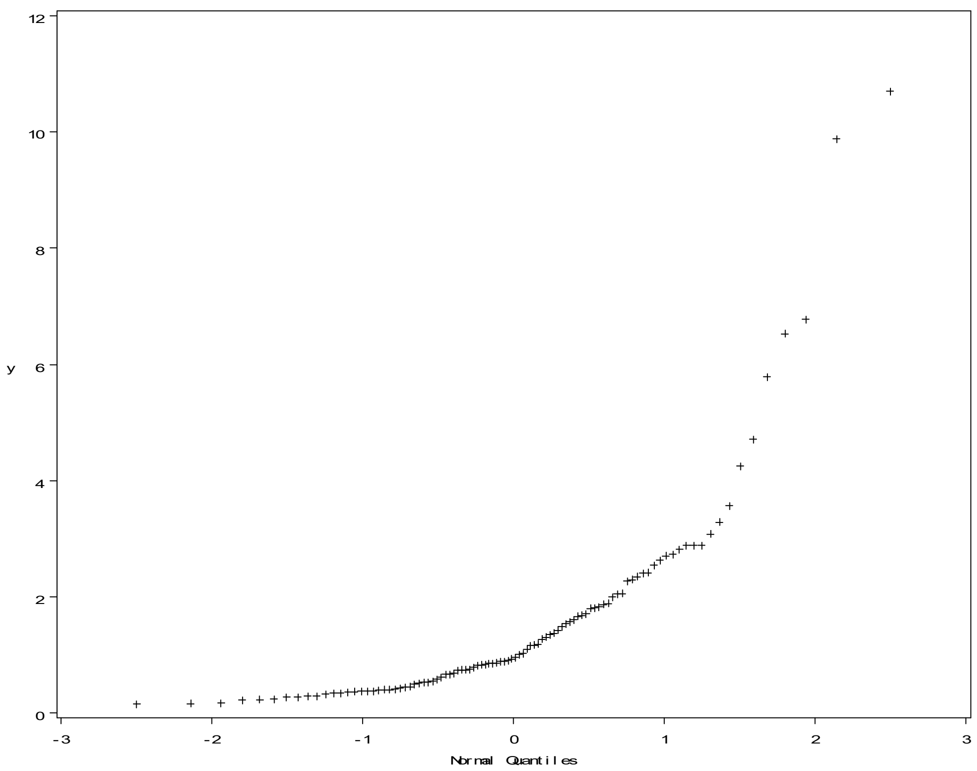
c) Which multiple regression model is ``the best'' according to the model selection criteria.

```
Dependent Variable: y

                    R-Square Selection Method

  Number in
```

```
Model     R-Square       C(p)          AIC           SBC    Variables in Model

     1      0.8047     84.2465      110.4685      112.90629    x3
     1      0.7558    110.5974       116.0546      118.49234    x4
     1      0.2646    375.3447       143.6180      146.05576    x1
     1      0.2470    384.8325       144.2094      146.64717    x2
------------------------------------------------------------------------------------
     2      0.9330     17.1130        85.7272      89.38384    x1 x3
     2      0.8773     47.1540      100.8605      104.51716    x3 x4
     2      0.8153     80.5653      111.0812      114.73788    x1 x4
     2      0.8061     85.5196      112.2953      115.95191    x2 x3
     2      0.7833     97.7978      115.0720      118.72864    x2 x4
     2      0.4642    269.7800       137.7025      141.35916    x1 x2
------------------------------------------------------------------------------------
     3      0.9615      3.7274       73.8473      78.72282    x1 x3 x4
     3      0.9341     18.5215        87.3143      92.18984    x1 x2 x3
     3      0.8790     48.2310      102.5093      107.38479    x2 x3 x4
     3      0.8454     66.3465      108.6361      113.51157    x1 x2 x4
------------------------------------------------------------------------------------
     4      0.9629      5.0000       74.9542      81.04859    x1 x2 x3 x4
```

Model with x1,x3 and x4 – the ``smallest'' (in terms of the number of regressors) model for which C(p) ≈ p, it also has a minimal AIC and SBC

5) What can you say about the distribution of the data demonstrated on the attached qqplot ?



The distribution is strongly skewed – long tail on the right, short tail on the left.