STA 302 / 1001 H - Summer 2004
Test 2
June 16, 2004

LAST NAME:____________________ FIRST NAME:____________________

STUDENT NUMBER:____

ENROLLED IN: (circle one) STA 302 STA 1001

INSTRUCTIONS:
- Time: 60 minutes
- Aids allowed: calculator.
- A table of values from the t distribution is on the second to last page (page 8).
- A table of formulae is on the last page (page 9).
- Total points: 30

1	2	3 ab	3 cdef

1. Suppose the following four pairs of observations have been made

Y_i	1	2	1	3
X_i	0	1	2	3

and a simple linear regression is to be fit to the data.

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 7/10 & -3/10 \\ -3/10 & 1/5 \end{pmatrix} \quad \text{and} \quad \mathbf{e}'\mathbf{e} = 1.5$$

(a) (1 point) State the $\mathbf{X}$ matrix.

(b) (3 marks) Find the least squares estimates of the slope and intercept of the regression line.

(c) (2 marks) Estimate the covariance between the estimators of the slope and intercept.

2. For the multiple linear regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, the least squares estimates are $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and the residuals are $\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b}$. Assume the Gauss-Markov conditions hold.

   (a) (3 marks) Show that $\mathbf{b}$ is an unbiased estimator of β.

   (b) (3 marks) You may recall that in simple linear regression, $\sum_{i=1}^{n} X_i e_i = 0$. For multiple linear regression, show that $\mathbf{X}'\mathbf{e} = \mathbf{0}$.

3. This question looks at the results of a study of gas chromatography, a technique which is used to detect very small amounts of a substance. Five measurements were taken for each of four specimens containing different amounts of the substance. The amount of the substance in each specimen was determined before the experiment. The response variable is the output reading from the gas chromatograph. Although the observations were made by the same machine over time, assume that they are independent. Some output from SAS is given below.

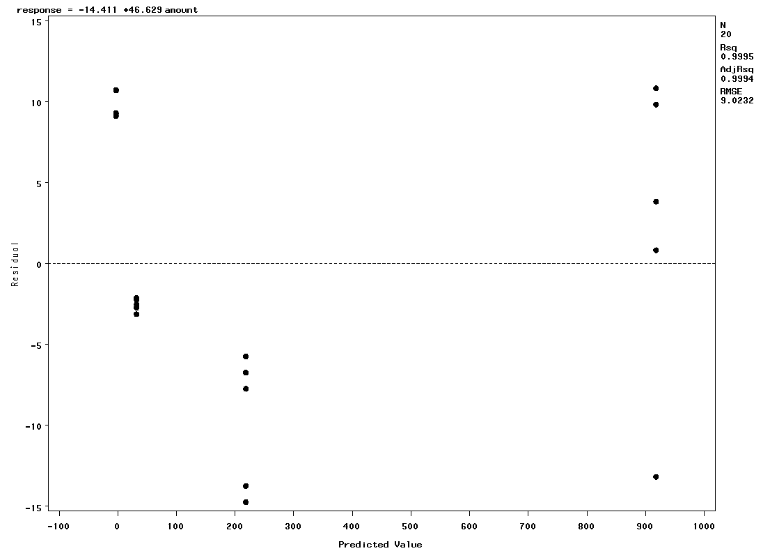
```
                         The REG Procedure
                       Descriptive Statistics

                                          Uncorrected                  Standard
Variable          Sum         Mean             SS       Variance      Deviation

Intercept    20.00000      1.00000       20.00000              0              0
amount      131.25000      6.56250     2130.31250       66.78865        8.17243
response   5831.80000    291.59000        4461028         145291      381.17090

                         The REG Procedure
                          Model: MODEL1
                    Dependent Variable: response

                        Analysis of Variance
                                 Sum of          Mean
Source               DF         Squares        Square    F Value    Pr > F
Model                 1         2759068       2759068    33887.5    <.0001
Error                18      1465.53180      81.41843
Corrected Total      19         2760534

             Root MSE            9.02322    R-Square     0.9995
             Dependent Mean    291.59000    Adj R-Sq     0.9994
             Coeff Var           3.09449

                        Parameter Estimates
                         Parameter      Standard
     Variable     DF      Estimate         Error    t Value    Pr > |t|
     Intercept     1     -14.41069       2.61421      -5.51      <.0001
     amount        1      46.62868       0.25330     184.09      <.0001
```

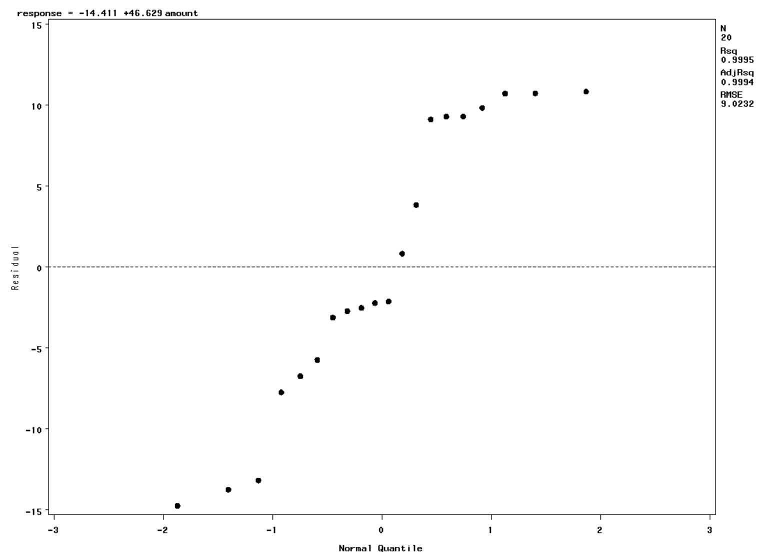
Questions based on this output start on the next page.

(a) (4 points) Predict the amount of the substance which will give a response of 500 units and construct an appropriate 95% interval for this prediction.

(b) (2 points) What residual plots would you like to see to check whether it is reasonable to "treat the observations as independent"?

Given below are a plot of the residuals versus the predicted values and a normal probability plot of the residuals for the gas chromatography data.





The questions that follow on the next page relate to these two residual plots.

(c) (2 marks) Describe (with a sketch or in words) what a plot of the residuals versus `amount` (the independent variable) would look like.

(d) (5 marks) Describe any problems you see with the given residual plots. Indicate what assumptions about the model are being violated.

(e) (2 marks) How do your comments about the residual plots made in the previous part affect the interpretation of your answer to part (a) on page 5?

(f) (3 marks) What would be appropriate action to remedy any problems identified in the residual plots? Justify your answer.

Some formulae:

$$b_1 = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2} \qquad b_0 = \overline{Y} - b_1\overline{X}$$

$$\mathrm{Var}(b_1) = \frac{\sigma^2}{\sum(X_i - \overline{X})^2} \qquad \mathrm{Var}(b_0) = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right)$$

$$\mathrm{Cov}(b_0, b_1) = -\frac{\sigma^2\overline{X}}{\sum(X_i - \overline{X})^2} \qquad \mathrm{SSTO} = \sum(Y_i - \overline{Y})^2$$

$$\mathrm{SSE} = \sum(Y_i - \hat{Y}_i)^2 \qquad \mathrm{SSR} = b_1^2\sum(X_i - \overline{X})^2 = \sum(\hat{Y}_i - \overline{Y})^2$$

$$\begin{aligned} \sigma^2\{\hat{Y}_h\} &= \mathrm{Var}(\hat{Y}_h) \\ &= \sigma^2\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right) \end{aligned} \qquad \begin{aligned} \sigma^2\{\mathrm{pred}\} &= \mathrm{Var}(Y_h - \hat{Y}_h) \\ &= \sigma^2\left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right) \end{aligned}$$

$\hat{X}_h \pm \frac{t_{n-2,1-\alpha/2}}{|b_1|} *$ appropriate s.e.
(valid approximation if $\frac{t^2s^2}{b_1^2\sum(X_i - \overline{X})^2}$ is small)

Working-Hotelling coefficient:
$W = \sqrt{2F_{2,n-2;1-\alpha}}$

$$\begin{aligned} \mathrm{Cov}(\mathbf{X}) &= E[(\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})'] \\ &= \mathrm{E}(\mathbf{X}\mathbf{X}') - (\mathrm{E}\mathbf{X})(\mathrm{E}\mathbf{X})' \end{aligned} \qquad \mathrm{Cov}(\mathbf{AX}) = \mathbf{A}\mathrm{Cov}(\mathbf{X})\mathbf{A}'$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad \mathrm{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$\hat{\mathbf{Y}} = \mathbf{Xb} = \mathbf{HY} \qquad \mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \qquad \mathrm{SSR} = \mathbf{Y}'(\mathbf{H} - \tfrac{1}{n}\mathbf{J})\mathbf{Y}$$

$$\mathrm{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} \qquad \mathrm{SSTO} = \mathbf{Y}'(\mathbf{I} - \tfrac{1}{n}\mathbf{J})\mathbf{Y}$$