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

LAST NAME:SOLUTIONS 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.

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}$$

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

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \begin{pmatrix} 7/10 & -3/10 \\ -3/10 & 1/5 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 1 \\ 3 \end{pmatrix} \\ &= \begin{pmatrix} 7/10 & -3/10 \\ -3/10 & 1/5 \end{pmatrix} \begin{pmatrix} 7 \\ 13 \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ .5 \end{pmatrix} \end{aligned}$$

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

$$s^2 = 1.5/2 = .75$$

$$\text{Cov}(b_0, b_1) = .75(-3/10) = -.225$$

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.

$$\begin{aligned} E(\mathbf{b}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta \\ &= \beta \end{aligned}$$

   (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 $\mathbf{X}'\mathbf{e} = \mathbf{0}$.

$$\begin{aligned} \mathbf{X}'\mathbf{e} &= \mathbf{X}'(\mathbf{I} - \mathbf{H})\mathbf{Y} \\ &= \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{Y} \\ &= 0 \end{aligned}$$

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

Variable	Sum	Mean	Uncorrected SS	Variance	Standard 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

Source	DF	Sum of Squares	Mean 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

Variable	DF	Parameter Estimate	Standard 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.

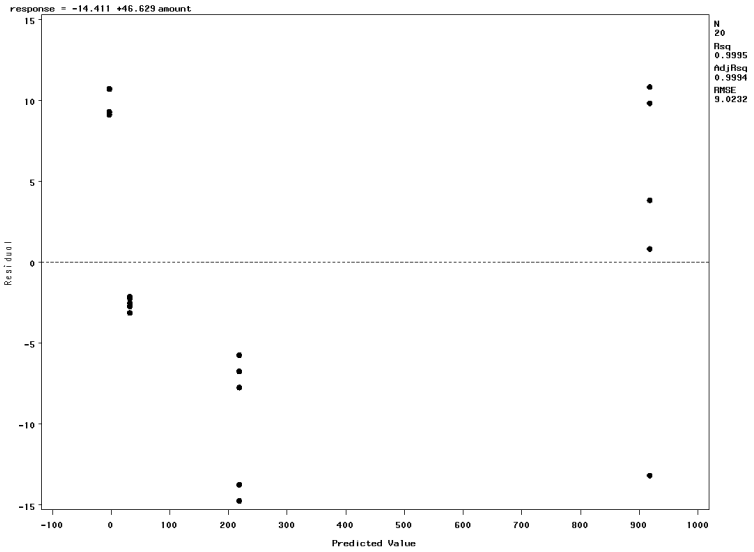
(a) (4 points) Predict the amount of the substance which will give a response of 500 units and construct an appropriate 95% interval for this prediction.

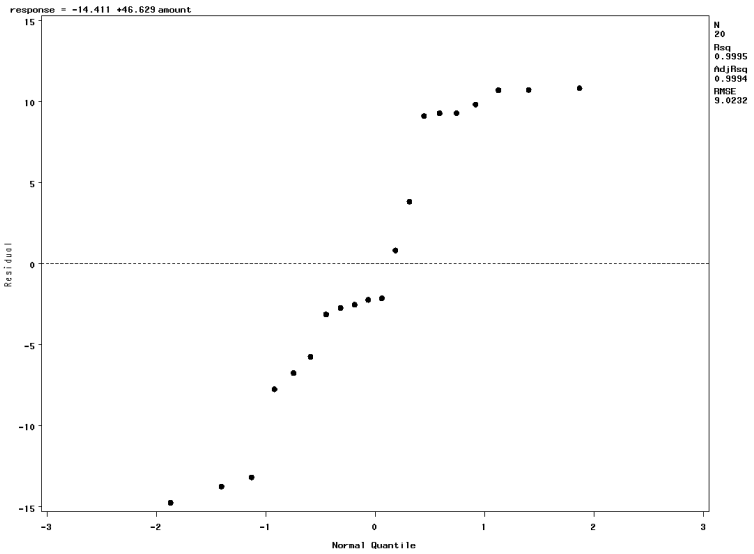
$\hat{X} = \frac{500-(-14.41)}{46.63} = 11.03$
$t_{18;.025} = 2.101$
Inverse PI:
$11.03 \pm \frac{2.101}{46.63}(9.0232)\sqrt{1 + \frac{1}{20} + \frac{(11.03-6.56)^2}{19(66.79)}} = (10.61, 11.45)$

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

Resdiduals versus time

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.

The same as the first plot on the previous page except for the scale on the horizontal axis.

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

The first plot shows curvature, so the linear model is not appropriate, and increasing variance. The second plot is not straight so data are not from a Normal distribution.

(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?

It is not correct. The prediction is not right since the linear model is not appropriate and the critical value is not right for a 95% interval since the data are not from a Normal distribution.

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

A transformation of Y (likely square root or log) will help with the curvature and the variance problems and hopefully will also fix the Normality problem as well.
(For a bonus mark: if the transformation of Y fixes the variance problem but there is still curvature, then transform X or fit a polynomial.)