STA 302 / 1001 F - Fall 2005
Test 2
November 16, 2005

LAST NAME:______________________ FIRST NAME:______________________

STUDENT NUMBER:____

ENROLLED IN: (circle one) STA 302 STA 1001

INSTRUCTIONS:
- Time: 90 minutes
- Aids allowed: calculator.
- A table of values from the t distribution is on the second to last page (page 9).
- A table of formulae is on the last page (page 10).
- For all questions you can assume that the formulae on page 10 are known.
- Total points: 40

1	2ab	2cd	3abc	3d	3efg

1. (4 points)

   Suppose that $\mathbf{X}$ is a 2×1 random vector with $\mathrm{E}(\mathbf{X}) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and

   $\mathrm{Cov}(\mathbf{X}) = \begin{pmatrix} 4 & -1 \\ -1 & 9 \end{pmatrix}$. $\mathbf{Y}$ is another random vector with $\mathbf{Y} = \mathbf{AX}$ where $\mathbf{A}$ is the constant matrix $\mathbf{A} = \begin{pmatrix} 1 & \text{-2} \\ 0 & 3 \end{pmatrix}$.

   Find the expectation of $\mathbf{Y}$ and the variance-covariance matrix for $\mathbf{Y}$.

2. (14 points)

(a) Write the simple linear regression model in matrix terms, defining all terms.

(b) Explain why $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$ follows from the assumptions of simple linear regression.

(c) A simple linear regression model is fit to data with 18 observations and the following are calculated:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 3 & -2 \\ -2 & 7 \end{pmatrix}$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

$$\mathbf{e}'\mathbf{e} = 4$$

Find a 90% confidence interval for the intercept.

(d) Show $\mathbf{b} = \boldsymbol{\beta} + \mathbf{R}\,\boldsymbol{\epsilon}$ where $\mathbf{R} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

3. (22 points)

The data considered in this question are values for breast cancer mortality (counts of number of women dying from breast cancer) from 1950 to 1960 and the adult white female population in 1960 for 301 counties in North Carolina, South Carolina, and Georgia. Interest is in considering how population can be used to predict the number of breast cancer cases.

Some output from SAS is given below.

The REG Procedure

Number of Observations Read	301
Number of Observations Used	301

Descriptive Statistics

Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	301.00000	1.00000	301.00000	0	0
population	3397705	11288	95320089347	189888678	13780
mortality	11997	39.85714	1257787	2598.73619	50.97780

The REG Procedure
Model: MODEL1
Dependent Variable: mortality

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	729096	729096	4314.70	<.0001
Error	299	50525	168.97946		
Corrected Total	300	779621			

Root MSE	12.99921	R-Square	0.9352
Dependent Mean	39.85714	Adj R-Sq	0.9350
Coeff Var	32.61451		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.52612	0.96921	-0.54	0.5876
population	1	0.00358	0.00005446	65.69	<.0001

Questions related to this SAS output are on the next 3 pages.

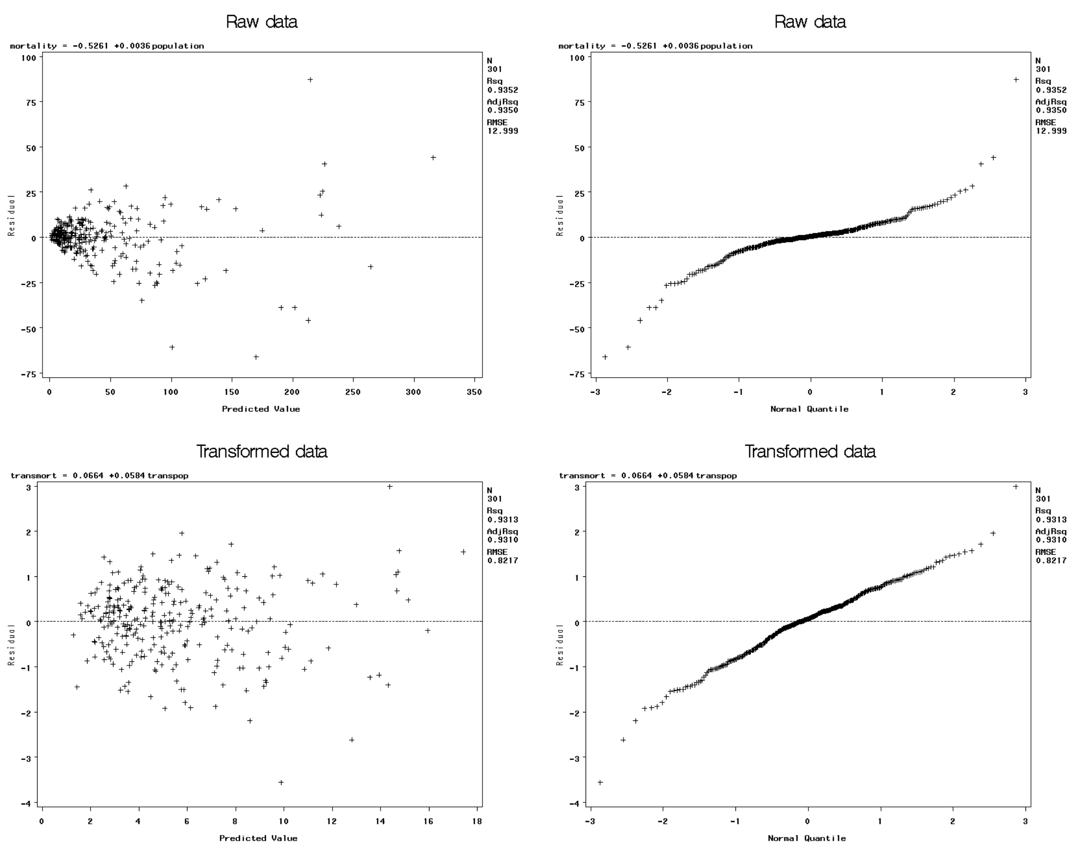
(a) Find simultaneous 99% confidence intervals for the slope and intercept.

(b) What does it mean for the intervals in (a) to be "simultaneous"?

(c) Estimate the population in 1960 for a county with 100 breast cancer deaths in the years from 1950 to 1960. Include an appropriate 95% interval for your estimate. Verify that the approximation used in the derivation of the interval formula holds.

(d) Despite the best efforts of the U.S. census, it is well known that population is measured with error. How does this fact affect the estimate of the slope? Do you think this is a serious problem? Why or why not?

(e) Given below are two residual plots for the regression of mortality on population and two residual plots for the same data after undergoing appropriate transformations.



Part (e) continues on the next page.

(e) continued . . .
Describe the relevant features of each of the 4 residual plots on the previous page and, as a consequence, whether you think it is appropriate to use the raw or transformed data.

(f) For count data such as mortality, the variance typically increases proportionally to the mean. Based on this information, which variance stabilizing transformation is appropriate?

(g) The normal quantile plots use $e_{(i)}$ in their construction. What is $e_{(i)}$ and how is it used in constructing a normal quantile plot?

Some formulae:

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

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

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

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

$$\sigma^2\{\hat{Y}_h\} = \mathrm{Var}(\hat{Y}_h) = \sigma^2\left(\frac{1}{n} + \frac{(X_h-\overline{X})^2}{\sum(X_i-\overline{X})^2}\right)$$

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

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

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

$$r = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2}}$$

$$\mathrm{Cov}(\mathbf{X}) = \mathrm{E}[(\mathbf{X} - \mathrm{E}\mathbf{X})(\mathbf{X} - \mathrm{E}\mathbf{X})'] = \mathrm{E}(\mathbf{X}\mathbf{X}') - (\mathrm{E}\mathbf{X})(\mathrm{E}\mathbf{X})' \qquad \mathrm{Cov}(\mathbf{A}\mathbf{X}) = \mathbf{A}\mathrm{Cov}(\mathbf{X})\mathbf{A}'$$

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

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{Y} \qquad \mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

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

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

$$\sigma^2\{\hat{Y}_h\} = \mathrm{Var}(\hat{Y}_h) = \sigma^2\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h$$

$$\sigma^2\{\text{pred}\} = \mathrm{Var}(Y_h - \hat{Y}_h) = \sigma^2(1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$$