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

LAST NAME: SOLUTIONS 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: 44

1	2ab	2cd	3abc	3d	3efg
4	7	7	10	3	9

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 $\mathbf{A} = \begin{pmatrix} 1 & -2 \\ 0 & 3 \end{pmatrix}$.

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

$$
\begin{aligned}
\mathrm{E}(\mathbf{Y}) &= \mathbf{A}\mathrm{E}(\mathbf{X}) \\
&= \begin{pmatrix} 1 & -2 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \\
&= \begin{pmatrix} -3 \\ 6 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Cov}(\mathbf{Y}) &= \mathbf{A}\mathrm{Cov}(\mathbf{X})\mathbf{A}' \\
&= \begin{pmatrix} 1 & -2 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 4 & -1 \\ -1 & 9 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -2 & 3 \end{pmatrix} \\
&= \begin{pmatrix} 6 & -19 \\ -3 & 27 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -2 & 3 \end{pmatrix} \\
&= \begin{pmatrix} 44 & -57 \\ -57 & 81 \end{pmatrix}
\end{aligned}
$$

2. (14 points)

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

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$$

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

(b) Explain why $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$ follows from the assumptions of simple linear regression.
(4 marks)
Cov($\boldsymbol{\epsilon}$) *is the* $n \times n$ *matrix with* i*th diagonal entry equal to the variance of* ϵ_i *and* ij*th off-diagonal entry equal to Cov*(ϵ_i, ϵ_j)*. Since two regression assumptions are that Var*(ϵ_i) $=$ σ^2 *for all* i *and Cov*(ϵ_i, ϵ_j) $= 0$*, it follows that Cov*($\boldsymbol{\epsilon}$) $= \sigma^2\mathbf{I}$.

(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.
(5 marks)

$$\mathbf{b} = \begin{pmatrix} 3 & -2 \\ -2 & 7 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -5 \\ 9 \end{pmatrix}$$

$$\textit{estimate of } Cov(\mathbf{b}) = \frac{4}{16} \begin{pmatrix} 3 & -2 \\ -2 & 7 \end{pmatrix} = \begin{pmatrix} 3/4 & -1/2 \\ -1/2 & 7/4 \end{pmatrix}$$

$t_{16;.05} = 1.746$
90% confidence interval for b_0: $-5 \pm 1.746\sqrt{3/4} = (-6.51, -3.49)$

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

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\ &= \boldsymbol{\beta} + \mathbf{R}\boldsymbol{\epsilon} \end{aligned}$$

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
                                Uncorrected                   Standard
Variable         Sum       Mean          SS     Variance     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
                              Sum of           Mean
Source                DF     Squares         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
                      Parameter       Standard
    Variable     DF    Estimate          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.
(3 marks)
Using the Bonferroni method, need $t_{299,.005/2} \doteq 2.86$ *(approximating with 120 d.f.)*
C.I. for β_0*:* $-0.52612 \pm 2.86(0.96921) = (-3.298, 2.246)$
C.I. for β_1*:* $0.00358 \pm 2.86(0.00005446) = (0.00342, 0.00374)$

(b) What does it mean for the intervals in (a) to be "simultaneous"?
(2 marks)
The procedure that produces these C.I.s captures the true values of **both** β_0 *and* β_1 **at least** *99% of the time in repeated samples of size 301.*

(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.
(5 marks)
Estimate: $\hat{X} = \frac{100-(-0.52612)}{0.00358} = 28079$
Interval: $28079 \pm \frac{1.98}{0.00358}(12.99921)\sqrt{1 + \frac{1}{301} + \frac{(28079-11288)^2}{95320089347-301(11288)^2}} = (20859, 35298)$
Approximation is OK since the two-sided test for $H_0 : \beta_1 = 0$ *is highly significant guaranteeing that* $\frac{t^2 s^2}{b_1^2 S_{XX}}$ *is small.*

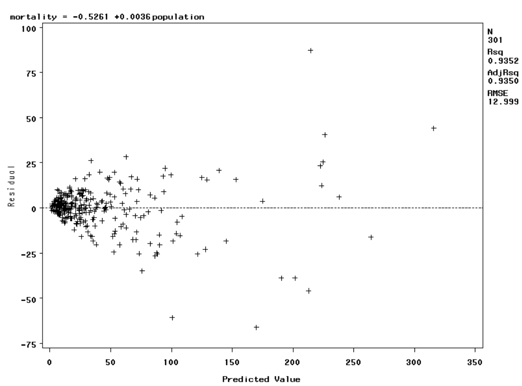
(d) Despite the best efforts of the U.S. census, it is well known that population is measured with error. How does this fact affect the estimate of the slope? Do you think this is a serious problem? Why or why not?

(3 marks)

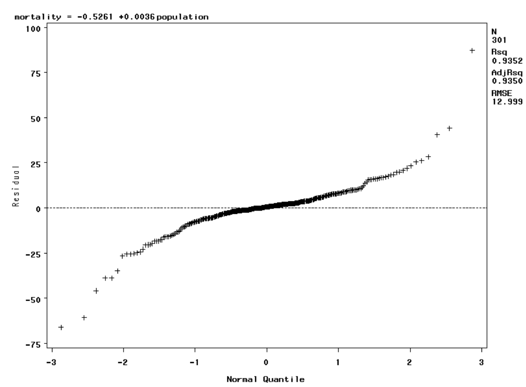
The estimated slope will be biased for the coefficient of population measured without error. But not a serious problem as the bias is small if the ratio of the variance of the measurement error in population is small compared to the variance of the populations, which is likely true.

(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.

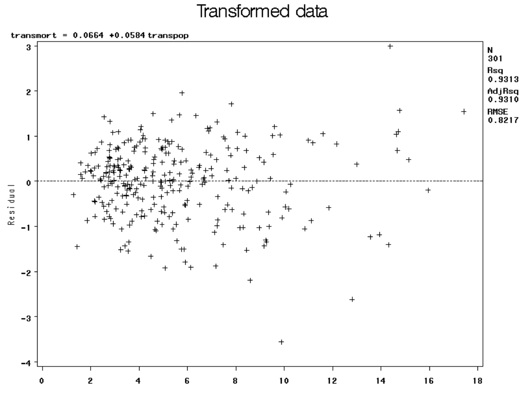
Raw data



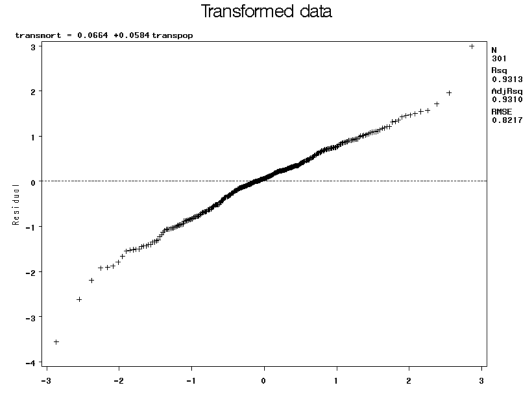
Raw data



Transformed data



Transformed data



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.

(5 marks)

Raw data first plot: shows variance increasing with predicted values

Raw data second plot: should not be considered at this point (the evident problems may be due to problems other than normality)

Transformed data first plot: looks OK; variance problems no longer exist; two values with large positive and large negative residuals may be worth further investigation

Transformed data second plot: looks OK as is close to straight but shows some indication of heavier tails than normal in the residuals but not serious

The transformed data model is appropriate as it more closely satisfies the linear regression assumptions.

(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?

(1 mark)

Square root of Y (mortality)

(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?

(3 marks)

$e_{(i)}$ *is the ith residual when they are ordered from smallest to largest.*

For a normal quantile plot, need the corresponding quantiles from the Normal distribution. $e_{(i)}$ *is plotted against the i/nth normal quantile (or a slight pertubation of i/n).*