**STA 302 / 1001 H - Fall 2003**
**Test 2**
November 10, 2003

LAST NAME: SOLUTIONS ______________ FIRST NAME: ______________

STUDENT NUMBER: ______________________________

ENROLLED IN: (circle one) STA 302 STA 1001

INSTRUCTIONS:
- Time: 50 minutes
- Aids allowed: calculator.
- A table of values from the $t$ distribution is on the last page.
- A table of formulae is on the second to last page.
- Total points: 30

______________________________

______________________________

| 1 (a) (b) | 2 (a) | 2 (b) | 3 (a) | 3 (b) (c) | 3(d) |
|---|---|---|---|---|---|
|  |  |  |  |  |  |

1. (a) (4 points) State the simple linear regression model in matrix terms, defining all matrices and vectors. Include the Gauss-Markov assumptions.

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \quad \text{(1 mark)}$$

*where*

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \qquad \text{(1 mark)}$$

*Gauss-Markov assumptions:* (2 marks)
$\mathrm{E}(\epsilon) = \mathbf{0}$
$\mathrm{Cov}(\epsilon) = \sigma^2\mathbf{I}$

(b) (4 points) Use matrix properties to prove $\mathbf{e}'\hat{\mathbf{Y}} = \mathbf{0}$ where $\mathbf{e}$ is the vector of residuals and $\hat{\mathbf{Y}}$ is the vector of predicted values. (Do not work with the individual elements of these matrices.)

$$\begin{aligned} \mathbf{e}'\hat{\mathbf{Y}} &= [(\mathbf{I} - \mathbf{H})\mathbf{Y}]'\mathbf{H}\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{H}\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{H}\mathbf{Y} - \mathbf{Y}'\mathbf{H}^2\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{H}\mathbf{Y} - \mathbf{Y}'\mathbf{H}\mathbf{Y} \\ &= \mathbf{0} \end{aligned}$$

2. The times to failure ($Y$) for 20 light bulbs were measured at 20 temperatures ($X$). A simple linear regression was carried out to explore how the failure time is related to the temperature.

   (a) (5 points) Here are some calculated values from the 20 data points.

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.105 & -0.0136 \\ -0.0136 & 0.00335 \end{pmatrix}$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 367.9 \\ 965.6 \end{pmatrix}$$

$$\mathbf{e}'\mathbf{e} = 789.7$$

What are the estimated slope and intercept of the regression line and their estimated standard errors?

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \begin{bmatrix} .105 & -.0136 \\ -.0136 & .00335 \end{bmatrix} \begin{bmatrix} 367.9 \\ 965.5 \end{bmatrix} \\ &= \begin{bmatrix} 25.5 \\ -1.77 \end{bmatrix} \end{aligned}$$

*So estimated intercept is 25.5 and estimated slope is -1.77.* (2 marks)

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

*Estimate of* $\sigma^2$*:* $s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-2} = \frac{789.7}{18} = 43.9$ (1 mark)
s.e. of $b_0 = \sqrt{43.9(.105)} = 2.15$ (1 mark)
s.e. of $b_1 = \sqrt{43.9(.00335)} = .38$ (1 mark)

(b) (2 points) A scatterplot and a plot of the residuals versus the predicted values for the light bulb problem are given below. What additional information does this provide?

*The point on the far right in the scatterplot and the far left in the residual plot is influential.* (1 mark) *The linear model may only be appropriate for temperatures less than* $10^o$*.* (1 mark)

3. An experiment was performed to determine which temperature in the manufacturing process results in the strongest product. In analysing the data, it was determined that a straight line model was an appropriate fit for the regression with square root of strength (`sqrtstrn`) as the dependent variable and temperature (`temp`) as the independent variable. Six temperatures ($80^o$ to $180^o$ in increments of $20^o$) were tested five times, resulting in 30 data points. Some SAS output is given below.

```
                               The REG Procedure
                            Descriptive Statistics

                                        Uncorrected                    Standard
Variable            Sum          Mean             SS      Variance     Deviation
Intercept      30.00000       1.00000      30.00000             0             0
temp         3900.00000     130.00000        542000    1206.89655      34.74042
sqrtstrn          20660     688.67158      15089253         29696     172.32658

                               The REG Procedure
                            Dependent Variable: sqrtstrn

                             Analysis of Variance
                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F
Model                     1         858611         858611    9295.97    <.0001
Error                    28     2586.18683       92.36382
Corrected Total          29         861197

                             Parameter Estimates
                          Parameter       Standard
     Variable     DF       Estimate          Error    t Value    Pr > |t|
     Intercept     1       44.77875        6.90488       6.49      <.0001
     x             1        4.95295        0.05137      96.42      <.0001
```

(a) (6 points) Construct simultaneous 90% confidence intervals for the expected strength for temperatures of $80^o$ and $180^o$. Use the Bonferroni method.
*When* `temp` *is 80, predicted value of* `sqrtstrn` *is* $44.79 + 4.953(80) = 441.03$
*When* `temp` *is 180, predicted value of* `sqrtstrn` *is* $44.79{+}4.953(180) = 936.33$
(1 mark each)
*Critical value:* $t_{28,.10/2/2} = t_{28,.025} = 2.048$ (1 mark)
*CI for E(*`sqrtstrn`*) when* `temp` *is 80:* (2 marks)
$441.03 \pm 2.048\sqrt{92.36}\sqrt{\frac{1}{30} + \frac{(80-130)^2}{29(1206.9)}} = 441.03 \pm 6.37 = (434.66, 447.4)$
*CI for E(*`sqrtstrn`*) when* `temp` *is 180:*
$936.33 \pm 6.37 = (929.96, 942.7)$ (1 mark)

(b) (2 points) The Bonferroni method is "conservative". Explain what this means in relation to your answer to (a).
*The probability that* **both** *CIs constructed in the manner in (a) capture the true expected strengths is* **at least 90%** *(probably more).* (1 mark for each concept indicated by bold words)

(c) (3 points) Before carrying out the regression using the square root of strength, a regression was first carried out with the raw data (strength regressed on temperature). What would be evident in the plot of residuals versus predicted values that would cause the people analysing the data to try the square root transform?
*(Wording of question has been corrected for answer below.)*
*curvature* (1 mark)
*non-constant variance* (1 mark)
*For the other mark, needed to note either that the curvature is monotone or the variance increases linearly in the mean value of* $Y$.

(d) (4 points) The normal probability plot of the residuals for the regression of square root of strength on temperature is given below. What additional information does this give you? Quote one number from the SAS output whose value is relevant to this additional information and explain how it is relevant.

*(Wording of question has been corrected for answer below.)*

*The errors do not follow a normal distribution.* (1 mark)

*The distribution of the residuals has heavier tails than a normal distribution.* (1 mark)

*You could pick any of the 3 p-values from the output.* (1 mark)

*The p-value may not be accurate as the relevant test statistic won't actually have an $F$ (or $t$) distribution.* (1 mark)