STA 302 / 1001 F - Fall 2006
Test 2
November 22, 2006

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 7).
- A table of formulae is on the last page (page 8).
- For all questions you can assume that the formulae on page 8 are known.
- Total points: 40

1	2ab	2cde	2f,3	4

1. For the multiple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, assume $\mathbf{X}$ is non-random.

   (a) (3 marks) Assume the Gauss-Markov assumptions hold. Show that the variance-covariance matrix of $\mathbf{b}$, the least squares estimates of $\boldsymbol{\beta}$ is $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

   (b) (2 marks) Why is the matrix $\mathbf{H}$ given that letter as its name?

   (c) We will now generalize the multiple linear model to include the case where the variance-covariance matrix of $\boldsymbol{\epsilon}$ is the $n \times n$ matrix $\boldsymbol{\Sigma}$ (without restriction on $\boldsymbol{\Sigma}$ except that it is a valid variance-covariance matrix). We will still assume that $E(\boldsymbol{\epsilon}) = \mathbf{0}$. To obtain the gernalized least squares estimate, the quantity $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ is minimized with respect to $\boldsymbol{\beta}$.

      i. (4 marks) Show that the generalized least squares estimate is $\mathbf{b} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$.

      ii. (2 marks) In this case $\mathbf{H} = \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}$. Show that $\mathbf{H}$ is idempotent.

2. The data considered in this question are the salaries in dollars and years of experience for a sample of 50 social workers. Some SAS output from a regression of the natural logarithm of salary on experience follows:

```
                         The REG Procedure
                       Descriptive Statistics

                                          Uncorrected                    Standard
  Variable            Sum         Mean             SS      Variance     Deviation
  Intercept      50.00000      1.00000       50.00000             0             0
  experience    906.00000     18.12000          19304      58.92408       7.67620
  logsalary     537.34675     10.74694     5783.18290       0.17045       0.41286

                   Dependent Variable: logsalary
               Number of Observations Read          50
               Number of Observations Used          50

                        Analysis of Variance
                                 Sum of           Mean
Source                 DF       Squares         Square    F Value    Pr > F
Model                   1       7.21215        7.21215     303.65    <.0001
Error                  48       1.14006        0.02375
Corrected Total        49       8.35220

             Root MSE              0.15411    R-Square     0.8635
             Dependent Mean       10.74694    Adj R-Sq     0.8607
             Coeff Var             1.43403

                        Parameter Estimates
                        Parameter       Standard
    Variable      DF     Estimate          Error    t Value    Pr > |t|
    Intercept      1      9.84132        0.05636     174.63      <.0001
    experience     1      0.04998        0.00287      17.43      <.0001
```

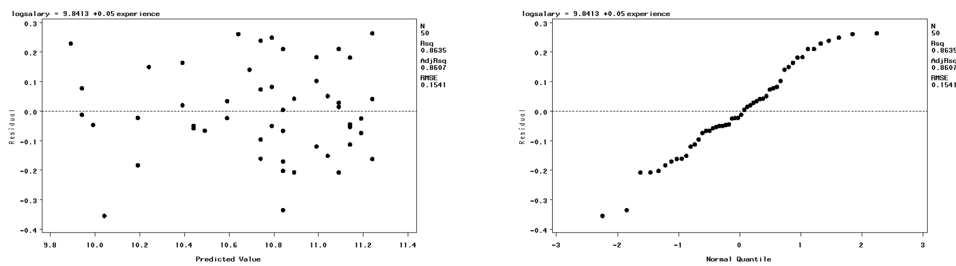
(a) (2 marks) Interpret the meaning of the slope in terms of the original variables.

(b) (5 marks) Estimate the number of years of experience for a social worker who makes \$35,000. Put a 95% interval around your estimate.

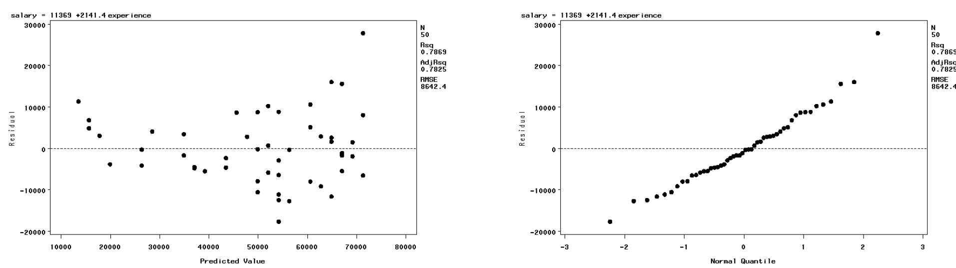
(c) (2 marks) The distribution of years of experience has a lighter left tail than a normal distribution. How does this affect your interval in part (b)? Explain.

(d) (4 marks) Below are 2 residual plots for the regression above, and 2 residual plots for the regression on the untransformed data. Based on these plots, which model do you prefer and why?

Regression of log of salary on experience



Regression of salary on experience



(e) (2 marks) Why do the values on the left and right ends of the normal quantile plot deserve more attention than the centre when looking at normal quantile plots?

(f) (4 marks) Also recorded for the social workers was the number of people they supervise. Below is some SAS output for the regression of log of salary on years of experience and number of people supervised.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.84548	0.06138	160.40	<.0001
experience	1	0.05095	0.00608	8.38	<.0001
supervise	1	-0.00189	0.01040	-0.18	0.8566

Interpret the meaning of the coefficient of `supervise` and the meaning of the corresponding p-value.

3. (3 marks) Find the variance stabilizing transformation of Y if $E(Y) = \mu$ and $\text{Var}(Y)$ is proportional to μ^4.

4. (a) (4 marks) Some data analysts prefer to use semi-studentized residuals instead of the raw residuals in plots. Explain how the interpretations of the plot of the residuals versus predicted values and of the normal quantile plot change when semi-studentized residuals are used.

(b) (3 marks) Consider a simple linear regression with non-random X. It could be argued that the assumption of normality could be assessed by evaluating a normal quantile plot of the Y's, rather than the residuals. Is this a reasonable approach? Why or why not?