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

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 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 $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

$$\begin{aligned}
\mathrm{Cov}(\mathbf{b}) &= \mathrm{Cov}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\sigma^2\mathbf{I}\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

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

   Fitted values are $\mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$ *and are usually given the symbol* $\hat{\mathbf{Y}}$ *which is read "$\mathbf{Y}$ hat".*

   (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 $\mathbf{b} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$.

      Need to minimise $Q = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ *with respect to* $\boldsymbol{\beta}$.

$$\begin{aligned}
Q &= \mathbf{Y}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} - \boldsymbol{\beta}'\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} - \mathbf{Y}\boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{Y}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{\beta}
\end{aligned}$$

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} + 2\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{\beta}$$

      Setting $\frac{\partial Q}{\partial \boldsymbol{\beta}} = 0$ *gives* $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\mathbf{b}$
      so $\mathbf{b} = \left(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$

      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.

$$\begin{aligned}
\mathbf{H}^2 &= \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1} \\
&= \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1} \\
&= \mathbf{H}
\end{aligned}$$

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.

Fitted line is: `logsalary` $= 9.84 + 0.04998$ `experience`
so `salary` $= e^{9.84} e^{0.04998\,\texttt{experience}}$
If `experience` *increases by 1 year (additive),* `salary` *increases by a factor of* $e^{0.04998} = 1.05$ *(multiplicative), i.e. increases by 5%.*

(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.

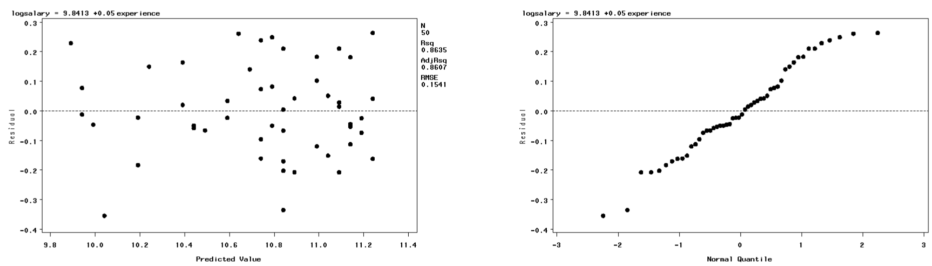
Estimate of `experience` *for a salary of \$35,000 is* $(\log(35000) - 9.84132)/0.04998 = 12.4$
$t_{48,0.025} \doteq 2.021$ *(estimating with 40 d.f.)*
95% inverse prediction interval: $12.4 \pm \frac{2.021}{0.04998}(0.15411)\sqrt{1 + \frac{1}{50} + \frac{(12.4-18.12)^2}{19304-50(18.12)^2}} = (6.07,\ 18.73)$
(There are at least 2 other ways of getting S_{XX} *from what was given.)*

(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.

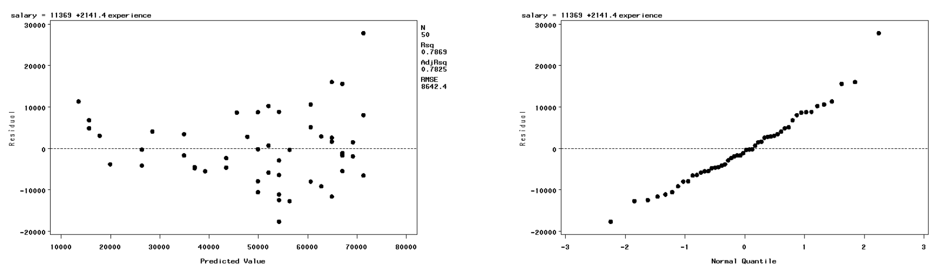
It doesn't affect it. There are no distributional assumptions on the predictor variable.

(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



In the regression of salary on experience, the first plot shows increasing variance, and the normal quantile plot shows a normal distribution except for one large point.
In the regression of the log of salary on experience, the first plot shows no problems (random scatter) and the second shows some skew.
Since non-constant variance is a more serious problem than lack of normality, I prefer the regression using log of salary.

(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?

The assumption of normality is used for calculating p-values and quantiles used in confidence intervals, both of which use values in the tails of the distribution.

(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.

Coefficient: For social workers with the same experience, an increase of 1 in the number of people supervised results in a change in salary by a factor of $e^{-0.00189} = 0.998$.
p-value: For a model with `experience` *in it, the data are consistent with the coefficient of* `supervise` *being 0.*

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

Need to find the function f such that

$$\left(f'(\mu)\right)^2 \propto \frac{1}{\mu^4}$$

so

$$f'(\mu) \propto \frac{1}{\mu^2}$$

and

$$f(\mu) \propto \frac{1}{\mu}$$

So the required transformation is $1/Y$.

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.

   Plot of residuals versus predicted values:
   Points that are outlying in that they don't fit the regression model are easy to judge numerically as they are greater than 3 (or 2) in absolute value. (No difference for constant variance or curvature evaluation.)
   Normal quantile plot:
   No difference in interpretation. Semi-studentization will not affect whether it is a straight line. (It will just change the slope of the line.)

   (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?

   No. Although the Y's are normally distributed they don't have the same mean. The normal quantile plot evaluates whether a set of data from the same distribution appear to be from a normal distribution.

Some formulae:

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

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

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

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

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

$$\hat{X}_h \pm \frac{t_{n-2,1-\alpha/2}}{|b_1|} * \text{appropriate s.e.} \qquad r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2\sum(Y_i - \overline{Y})^2}}$$

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

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

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

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

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

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