STA 302 / 1001 H - Fall 2003
Test 1
October 6, 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.
- Total points: 30

Some formulae:

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

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

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

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

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

---

1 (a) (b)	1 (c) (d) (e)	2 (a) (b) (c)	2 (d)	3

1. (a) (4 points) State the simple linear regression model for dependent variable Y and independent variable X and the Gauss-Markov conditions.
Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Gauss-Markov conditions:
$E(\epsilon_i) = 0$
$Var(\epsilon_i) = \sigma^2$
ϵ_i*'s are uncorrelated*
(The GM conditions can also be stated in terms of the Y_i*'s.)*

(b) (5 points) The least squares estimate of the Y intercept for the model in (a) is b_0 as given on the first page. Under the model you specified in (a), derive the formula for the variance of b_0. You may **not** asssume that any of the formulae for variance or covariance are known.
Sketch of solution:
First show that b_0 *is a linear combination of the* Y_i*'s, i.e.*

$$b_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\overline{X}(X_i - \overline{X})}{\sum (X_i - \overline{X})^2} \right) Y_i$$

Since the Y_i*'s are uncorrelated,* $Var(\sum k_i Y_i) = \sum k_i^2 \, Var(Y_i) = \sum k_i^2 \sigma^2$
Squaring the k_i*'s and summing gives the result.*

(c) (3 points) In order to do inference about the slope (such as testing whether or not the slope is 0) we need to make one more assumption about the model in (a). What is the usual assumption and why is it necessary?

The errors have a normal distribution. We need a distributional assumption for the errors in order to have a sampling distribution for the estimators which is used for critical values and p-values.

(d) (2 points) An estimate is more precise than another if it has smaller variance. The estimate of the expected value of Y varies with the value of X. At what value of X will there be the most precise estimate of the expected value of Y? Justify your answer.

The variance of the expected value of Y at X^* is

$$\sigma^2 \left(\frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right)$$

This is smallest when X^* is the average value of the X's.

(e) (2 points) Suppose the dependent variable Y could be measured with less error. Why would this lead to more precise estimation of the intercept of the regression line?

The variance of the intercept is $\sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right)$. If Y has less error, σ^2 is smaller, so the variance of the intercept is smaller.

2. Some engineers are interested in examining the relationship between load, in pounds, and the corresponding deformation, in inches, on a mild steel bar. Based on the physical properties of the steel, the engineers believe there will be a linear relationship between the natural logarithm of load (`logL`) and the natural logarithm of deformation (`logD`). Data were collected for 24 loads ranging from 1000 to 9900 pounds and a regression of `logD` on `logL` was run. Some output from SAS is given below. (Some numbers have been purposely removed from the output.)

```
                            Descriptive Statistics
                                        Uncorrected                    Standard
Variable            Sum          Mean            SS      Variance     Deviation
Intercept      24.00000       1.00000      24.00000             0             0
logL          212.18446       8.84102    1883.83553       0.34386       0.58639
logD          174.88129       7.28672    1304.29718       1.30375       1.14182

                           Dependent Variable: logD
                             Analysis of Variance
                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F
Model                     1       20.57353       20.57353      48.09    <.0001
Error                    22        9.41263        0.42785
Corrected Total          23       29.98616

                             Parameter Estimates
                          Parameter       Standard
     Variable     DF       Estimate          Error    t Value    Pr > |t|
     Intercept     1       -6.97275        2.06066      -3.38      0.0027
     logL          1        1.61288        0.23259       6.93      <.0001
```

The following questions (labelled (a) through (d)) relate to the SAS output on the previous page.

(a) (3 points) Construct a 95% confidence interval for the slope. What can you conclude from this confidence interval about a test of $H_0 : \beta_1 = 0$?
$t_{22,0.025} = 2.074$
95% CI for β_1*:* $1.61288 \pm 2.074(0.233259) = (1.1305, 2.0953)$
The p-value for a two-sided test of $H_0 : \ \beta_1 = 0$ *is less than 0.05.*

(b) (2 points) What is the value of R^2? What does this value mean?
$R^2 = 20.57353/29.98616 = 0.686$
68.6% of the variation in the log of deformation is explained by its linear relationship with the log of load.

(c) (4 points) What is the predicted value of `logD` when the load is 9600 pounds? Construct an appropriate 90% interval for the expected value of `logD` at this load.
The predicted value of `logD` *when the load is 9600 pounds is*
$-6.97275 + 1.61288[\log(9600)] = 7.8166$
$t_{22,.05} = 1.717$
90% interval for mean of `logD` *when load is 9600:*

$$7.8166 \pm 1.717\sqrt{.42785}\sqrt{\left(\frac{1}{24} + \frac{[log(9600) - 8.84102]^2}{(23)(.34386)}\right)} = (7.5525, 8.0807)$$

(d) (2 points) Do you trust your prediction in (c)? Explain.
No. As can be seen from the scatterplot, a linear model is not appropriate.

3. (3 points) A simple linear regression is performed to study the relationship between a child's intelligence at age 5 (the dependent variable) and the length of time the child was breastfed as an infant (the independent variable). The value of r^2 was 0.56 and for the two-sided test of whether the slope was 0, the p-value was 0.013. A newspaper headline reporting on the study said "Study shows that to make a smarter child, mothers should breastfeed longer". Comment on the validity of the headline.
While there is an indication of a linear relationship between intelligence at age 5 and length of time breastfed, the headline claims breastfeeding causes intelligence. This is an observational study so causality can't be claimed.