STA 302 / 1001 H - Fall 2003
Test 1
October 6, 2003

LAST NAME:_____ 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:

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

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

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

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

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

---

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.

(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.

(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?

(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.

(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?

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

Variable	Sum	Mean	Uncorrected SS	Variance	Standard 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

Source	DF	Sum of Squares	Mean 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

Variable	DF	Parameter Estimate	Standard 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 $H_0 : \beta_1 = 0$?

(b) (2 points) What is the value of R^2? What does this value mean?

(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.

(d) (2 points) Do you trust your prediction in (c)? Explain.

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.