**STA 302 / 1001 H1F – Fall 2006**
**Test 1**
October 18, 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 last page (page 7).
- Total points: 50

---

## Some formulae:

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

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

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

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

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

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}}$$

---

| 1 | 2ab | 2cdef | 2gh | 3 |
|---|---|---|---|---|
| | | | | |

1. A simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

is fit using least squares to $n$ data points. Assume that the Gauss-Markov conditions hold and that the error terms are normally distributed with mean 0 and variance $\sigma^2$.

(a) (3 marks) What is the probability distribution of $b_1$? What is the probability distribution of $\beta_1$?

(b) (4 marks) Describe the method of least squares. How is it related to $R^2$?

(c) (3 marks) Suppose the regression model is being used to predict blood pressure as a function of weight. Explain the difference between a confidence interval for the mean response at a new $X$ and a prediction interval at a new $X$ in this context. (Do not discuss the details of the formulae for calculating the intervals.)

(d) (2 marks) Is the correlation between blood pressure and weight meaningful in a practical manner? Why or why not?

2. To calibrate a measurement technique, researchers use a set of known $X$'s (determined in advance by the researchers) to obtain observed $Y$'s, then fit a model with $Y$ as the dependent variable and $X$ as the independent variable. This model can be used to convert future measured $Y$'s back into the corresponding $X$'s. The data in this exercise were collected for the calibration process of a technique designed to detect the quantity of calcium in a sample of material. $X$ is the known quantity of calcium in each sample of material, $Y$ is the amount of calcium measured by the technique being calibrated.

   Some output from SAS is given below. Note that some numbers have been replaced by letters.

```
                                The REG Procedure
                              Dependent Variable: y

                               Analysis of Variance
                                      Sum of           Mean
Source                   DF          Squares         Square    F Value    Pr > F
Model                     1       1077.24294     1077.24294      (A)      <.0001
Error                     7          0.33928        0.04847
Corrected Total           8          (B)

             Root MSE                0.22016    R-Square      (C)
             Dependent Mean         25.25556    Adj R-Sq     0.9996
             Coeff Var               0.87171

                               Parameter Estimates
                            Parameter        Standard
     Variable      DF        Estimate           Error     t Value    Pr > |t|
     Intercept      1        -0.19487             (D)       -1.05      0.3292
     x              1         0.99373         0.00667      149.08      <.0001
```

   (a) (4 marks) Find the 4 missing values (A through D) in the SAS output.

   (b) (4 marks) Find a 99% confidence interval for the slope. Explain clearly how to interpret the confidence interval.

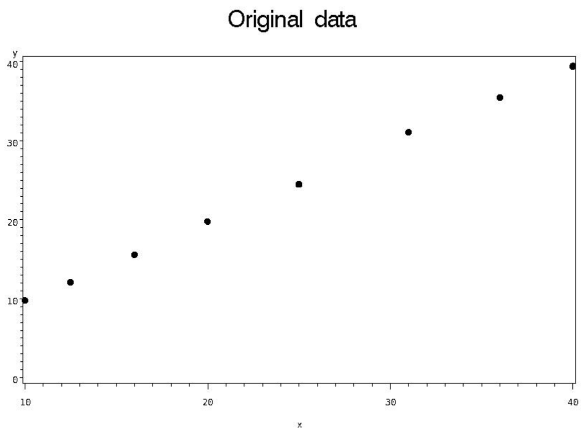
(c) (3 marks) How would the $p$-value for the $t$-test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ change if the sample size were doubled? Justify your answer. You may assume that the new data values are similar to the original data.

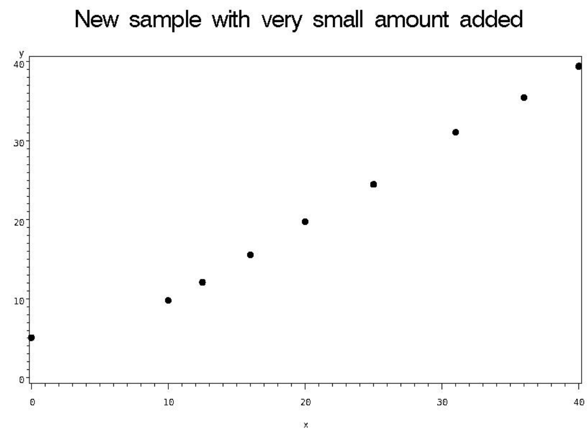
(d) (5 marks) A 95% confidence interval for the mean of $Y$ when $X = 30$ is (29.43, 29.80). Find a 95% prediction interval for the value of $Y$ for a new sample with $X = 30$.

(e) (2 marks) If there were no calcium present the technique should not detect any. Thus if $X = 0$, $Y$ should also be 0. Do the data give evidence to support this? Justify your answer.

(f) (3 marks) If the technique is any good at all, then the slope in the simple linear regression should be 1. Do the data give evidence to support this? Justify your answer using an appropriate hypothesis test.

(g) A scatterplot of the original data is given below. (Not all points are visible because some are close together.) Later, a new measurement is taken for a sample with a very small quantity of calcium. The second scatterplot below includes the original data and this new measurement.




i. (2 marks) A simple linear regression model is fit to the data in the second scatterplot. How will the values of the slope and $R^2$ compare to the corresponding values from the regression fit to the original data?

ii. (2 marks) Since the fitted regression equation changes when this new point is added to the data, what would you recommend the researchers do to model these data?

(h) (2 marks) Since the goal of the researchers is to be able to predict the quantity of calcium that is actually in the sample (what we've labelled $X$) given what the technique measures (what we've labelled $Y$), it is proposed that the regression be carried out with $X$ as the dependent variable and $Y$ as the independent variable. Comment briefly on how this proposal should be carried out and how the resulting regression equation would compare to the original. (Consider the original data values only for this question.)

3. In the previous question, it could be argued that the model should be forced through the origin because when $X = 0$, $Y$ must necessarily be 0. Then the model would reduce to

$$Y_i = \beta X_i + \epsilon_i, \quad i = 1, \ldots, n$$

As in the previous question, assume the $X_i$'s are known values set by the researchers, and that the usual assumptions for the normal errors regression model apply.

(a) (3 marks) For this model, show that the least squares estimate of $\beta$ is

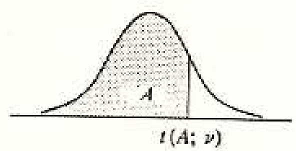
$$\hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2}$$

(b) (4 marks) Show that the estimate of $\beta$ in part (a) is unbiased. What assumptions do you need to impose on the model to do this?

(c) (4 marks) Find the variance of the estimate of $\beta$ from part (a). What assumptions do you need for this derivation?

**TABLE B.2** Percentiles of the $t$ Distribution.

Entry is $t(A; \nu)$ where $P\{t(\nu) \leq t(A; \nu)\} = A$




| $\nu$ | A: .60 | .70 | .80 | .85 | .90 | .95 | .975 |
|---|---|---|---|---|---|---|---|
| 1 | 0.325 | 0.727 | 1.376 | 1.963 | 3.078 | 6.314 | 12.706 |
| 2 | 0.289 | 0.617 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 |
| 3 | 0.277 | 0.584 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 |
| 4 | 0.271 | 0.569 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 |
| 5 | 0.267 | 0.559 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 |
| 6 | 0.265 | 0.553 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 |
| 7 | 0.263 | 0.549 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 |
| 8 | 0.262 | 0.546 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 |
| 9 | 0.261 | 0.543 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 |
| 10 | 0.260 | 0.542 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 |
| 11 | 0.260 | 0.540 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 |
| 12 | 0.259 | 0.539 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 |
| 13 | 0.259 | 0.537 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 |
| 14 | 0.258 | 0.537 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 |
| 15 | 0.258 | 0.536 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 |
| 16 | 0.258 | 0.535 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 |
| 17 | 0.257 | 0.534 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 |
| 18 | 0.257 | 0.534 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 |
| 19 | 0.257 | 0.533 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 |
| 20 | 0.257 | 0.533 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 |
| 21 | 0.257 | 0.532 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 |
| 22 | 0.256 | 0.532 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 |
| 23 | 0.256 | 0.532 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 |
| 24 | 0.256 | 0.531 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 |
| 25 | 0.256 | 0.531 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 |
| 26 | 0.256 | 0.531 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 |
| 27 | 0.256 | 0.531 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 |
| 28 | 0.256 | 0.530 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 |
| 29 | 0.256 | 0.530 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 |
| 30 | 0.256 | 0.530 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 |
| 40 | 0.255 | 0.529 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 |
| 60 | 0.254 | 0.527 | 0.848 | 1.045 | 1.296 | 1.671 | 2.000 |
| 120 | 0.254 | 0.526 | 0.845 | 1.041 | 1.289 | 1.658 | 1.980 |
| $\infty$ | 0.253 | 0.524 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 |

**TABLE B.2** *(concluded)* Percentiles of the $t$ Distribution.

| $\nu$ | A: .98 | .985 | .99 | .9925 | .995 | .9975 | .9995 |
|---|---|---|---|---|---|---|---|
| 1 | 15.895 | 21.205 | 31.821 | 42.434 | 63.657 | 127.322 | 636.590 |
| 2 | 4.849 | 5.643 | 6.965 | 8.073 | 9.925 | 14.089 | 31.598 |
| 3 | 3.482 | 3.896 | 4.541 | 5.047 | 5.841 | 7.453 | 12.924 |
| 4 | 2.999 | 3.298 | 3.747 | 4.088 | 4.604 | 5.598 | 8.610 |
| 5 | 2.757 | 3.003 | 3.365 | 3.634 | 4.032 | 4.773 | 6.869 |
| 6 | 2.612 | 2.829 | 3.143 | 3.372 | 3.707 | 4.317 | 5.959 |
| 7 | 2.517 | 2.715 | 2.998 | 3.203 | 3.499 | 4.029 | 5.408 |
| 8 | 2.449 | 2.634 | 2.896 | 3.085 | 3.355 | 3.833 | 5.041 |
| 9 | 2.398 | 2.574 | 2.821 | 2.998 | 3.250 | 3.690 | 4.781 |
| 10 | 2.359 | 2.527 | 2.764 | 2.932 | 3.169 | 3.581 | 4.587 |
| 11 | 2.328 | 2.491 | 2.718 | 2.879 | 3.106 | 3.497 | 4.437 |
| 12 | 2.303 | 2.461 | 2.681 | 2.836 | 3.055 | 3.428 | 4.318 |
| 13 | 2.282 | 2.436 | 2.650 | 2.801 | 3.012 | 3.372 | 4.221 |
| 14 | 2.264 | 2.415 | 2.624 | 2.771 | 2.977 | 3.326 | 4.140 |
| 15 | 2.249 | 2.397 | 2.602 | 2.746 | 2.947 | 3.286 | 4.073 |
| 16 | 2.235 | 2.382 | 2.583 | 2.724 | 2.921 | 3.252 | 4.015 |
| 17 | 2.224 | 2.368 | 2.567 | 2.706 | 2.898 | 3.222 | 3.965 |
| 18 | 2.214 | 2.356 | 2.552 | 2.689 | 2.878 | 3.197 | 3.922 |
| 19 | 2.205 | 2.346 | 2.539 | 2.674 | 2.861 | 3.174 | 3.883 |
| 20 | 2.197 | 2.336 | 2.528 | 2.661 | 2.845 | 3.153 | 3.849 |
| 21 | 2.189 | 2.328 | 2.518 | 2.649 | 2.831 | 3.135 | 3.819 |
| 22 | 2.183 | 2.320 | 2.508 | 2.639 | 2.819 | 3.119 | 3.792 |
| 23 | 2.177 | 2.313 | 2.500 | 2.629 | 2.807 | 3.104 | 3.768 |
| 24 | 2.172 | 2.307 | 2.492 | 2.620 | 2.797 | 3.091 | 3.745 |
| 25 | 2.167 | 2.301 | 2.485 | 2.612 | 2.787 | 3.078 | 3.725 |
| 26 | 2.162 | 2.296 | 2.479 | 2.605 | 2.779 | 3.067 | 3.707 |
| 27 | 2.158 | 2.291 | 2.473 | 2.598 | 2.771 | 3.057 | 3.690 |
| 28 | 2.154 | 2.286 | 2.467 | 2.592 | 2.763 | 3.047 | 3.674 |
| 29 | 2.150 | 2.282 | 2.462 | 2.586 | 2.756 | 3.038 | 3.659 |
| 30 | 2.147 | 2.278 | 2.457 | 2.581 | 2.750 | 3.030 | 3.646 |
| 40 | 2.123 | 2.250 | 2.423 | 2.542 | 2.704 | 2.971 | 3.551 |
| 60 | 2.099 | 2.223 | 2.390 | 2.504 | 2.660 | 2.915 | 3.460 |
| 120 | 2.076 | 2.196 | 2.358 | 2.468 | 2.617 | 2.860 | 3.373 |
| $\infty$ | 2.054 | 2.170 | 2.326 | 2.432 | 2.576 | 2.807 | 3.291 |