STA 302 / 1001 H1F – Fall 2006
Test 1

October 18, 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 last page (page 7).
- Total points: 50

Some formulae:

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

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

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

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

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

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

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 σ^2.

(a) (3 marks) What is the probability distribution of b_1? What is the probability distribution of β_1?

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(X_i - \overline{X})^2}\right) \quad \textit{(2 marks)}$$

β_1 is not random (1 mark)

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

Find the slope and intercept of the line that minimizes the sum of squares of the vertical deviations of the data points from the line. (2 marks)
The quantity that is being minimized as a proportion of variation in the Y's is $1 - R^2$. (2 marks)

(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.)

For a particular weight, there is a probability distribution for the possible values of blood pressure. The confidence interval for the mean of Y at that weight gives an interval that captures the mean of this probability distribution ($\beta_0 + \beta_1 X$) $100(1-\alpha)\%$ of the time. The prediction interval gives an interval that captures the actual value of the blood pressure for a person of that weight $100(1-\alpha)\%$ of the time.

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

Yes. Both variables are quantitative random variables.

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.

   $A = 22225.7$
   $B = 1077.58222$
   $C = 0.9997$
   $D = 0.18582$

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

   (2 marks for CI, 2 for interpretation)
   $t_{7,0.005} = 3.499$
   CI: $0.99373 \pm 3.499(0.00667) = (0.9704,\ 1.0171)$
   For repeated samples of size 9, a CI for the slope constructed in this manner will capture the value of β_1 99% of the time.

(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.

It would be smaller for 2 reasons: (1) $\sum(X_i - \overline{X})^2$ would be larger reducing the variance and increasing the value of the t test statistic; (2) the relevant t distribution would have larger degrees of freedom reducing the area in the tails.

(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 $X = 30$.

When $X = 30$, $\hat{Y} = -0.19487 + 0.99373(30) = 29.617$
$t_{7,.025} = 2.365$
Thus $s\sqrt{\frac{1}{n} + \frac{(30-\overline{X})^2}{\sum(X_i-\overline{X})^2}} = (29.617 - 29.43)/2.365 = 0.0791$
Since $s = 0.22016$, $\frac{1}{n} + \frac{(30-\overline{X})^2}{\sum(X_i-\overline{X})^2} = (0.0791/0.22016)^2 = 0.1291$
So the prediction interval is $29.617 \pm 2.365(0.22016)\sqrt{1 + 0.1291} = (29.064, 30.170)$

(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.

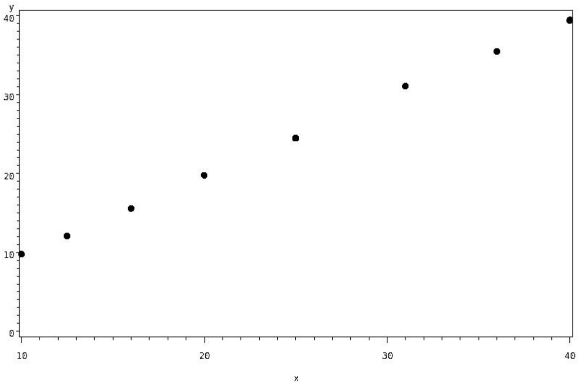
For the 2-sided test with null hypothesis that the intercept is 0, the p-value is 0.3292. Thus the data are consistent with $Y = 0$ when $X = 0$.

(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.

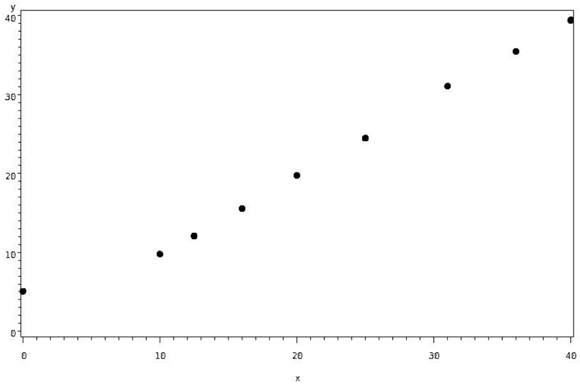
What to test $H_0 : \beta_1 = 1$ *versus* $H_a : \beta_1 \neq 1$
Test statistic: $t_{obs} = \frac{0.99373-1}{0.00667} = -0.94$
Estimating from a t distribution with 7 df, gives $2(.15) < p < 2(.2)$
So the data are consistent with a slope of 1.

(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.

Original data



New sample with very small amount added



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?

The slope will be smaller and R^2 will be smaller.

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?

Since the linear model seems to fit well on the original range of data but the new point outside the range does not follow the pattern, the model should be fit and used on the original range and something different should be done for samples with very small amounts of calcium.

(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.)

Since X is not random this should not be done; the researchers should continue to work with the model with Y as the dependent variable.

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 β is

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

Need to find the value of β that minimizes $S = \sum(Y_i - \beta X_i)^2$
$\frac{dS}{d\beta} = -2\sum X_i(Y_i - \beta X_i)$
Setting the derivative equal to zero gives the required result.

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

Assume $E(\epsilon_i) = 0$
Then $E(\hat{\beta}) = \frac{\sum X_i E(Y_i)}{\sum X_i^2} = \frac{\sum X_i(\beta X_i)}{\sum X_i^2} = \beta$ as required.
The only necessary assumption is that the expection of the ϵ's is 0.
(2 marks for derivation, 2 for assumptions)

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

$Var(\hat{\beta}) = \frac{\sum X_i^2 \, Var(Y_i)}{(\sum X_i^2)^2} = \frac{\sigma^2}{\sum X_i^2}$
In the calculation we needed to assume that the ϵ's are uncorrelated (so the Y's are also uncorrelated) and that the variance of the ϵ's is constant, labelled σ^2 (so the variance of the Y's is also σ^2).
(2 marks for derivation, 2 for assumptions)