STA 302 H1F / 1001 HF – Fall 2007
Test 1
October 24, 2007

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 8).
- Total points: 50

Some formulae:

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

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

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

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

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

$$r = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2}} \qquad \text{Working-Hotelling coefficient: } W = \sqrt{2\,F_{2,n-2;\,\alpha}}$$

1	2a	2bcdef	2ghi	2j	3

1. The following questions require derivations of results for the simple linear regression model.

   (a) (2 marks) In lecture we showed that $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} e_i X_i = 0$. Given these results, what is $\sum_{i=1}^{n} e_i \hat{Y}_i$? Justify your answer.

   (b) (5 marks) Show that the total Sum of Squares in a regression can be decomposed as

$$\sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

   You may use any results that were derived in lecture.

   (c) (5 marks) Assume that the X_i are non-random. Derive the formula for $\mathrm{Cov}(b_0, b_1)$ given on the first page. You may use any other formulae from the first page that you require except the formulae whose derivations require knowing $\mathrm{Cov}(b_0, b_1)$.
   (*Hint:* You may want to start with the formula for the estimated intercept.)

2. The SAS output that follows was produced to examine the relationship between full-scale IQ (`FSIQ`) and brain size as measured by MRI (`MRIcount`). Measurements were taken on 20 university students chosen because their full-scale IQ was at least 130.

The REG Procedure
Descriptive Statistics

Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	20.00000	1.00000	20.00000	0	0
MRIcount	18518961	925948	1.725648E13	5730703420	75701
FSIQ	2728.00000	136.40000	372396	15.62105	3.95235

Dependent Variable: FSIQ
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	89.22306	89.22306	7.74	0.0123
Error	18	207.57694	11.53205		
Corrected Total	19	296.80000			

Root MSE	3.39589	R-Square	0.3006
Dependent Mean	136.40000	Adj R-Sq	0.2618
Coeff Var	2.48965		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	109.89399	9.55947	11.50	<.0001
MRIcount	1	0.00002863	0.00001029	2.78	0.0123

(a) (5 marks) Give estimates of the following quantities:

the percent of total variability in `FSIQ` that is explained by its linear relationship with `MRIcount` __________

the `FSIQ` for a person whose `MRIcount` is 1,000,000 __________

the error variance __________

the variance of the slope __________

the difference in `FSIQ` between 2 people when the second person has `MRIcount` that is 10,000 units larger than the first person __________

(b) (2 marks) Explain the practical meaning of the estimated intercept.

(c) (2 marks) Give a 95% confidence interval for the slope.

(d) (1 mark) What are the null and alternative hypotheses for the test with p-value of 0.0123?

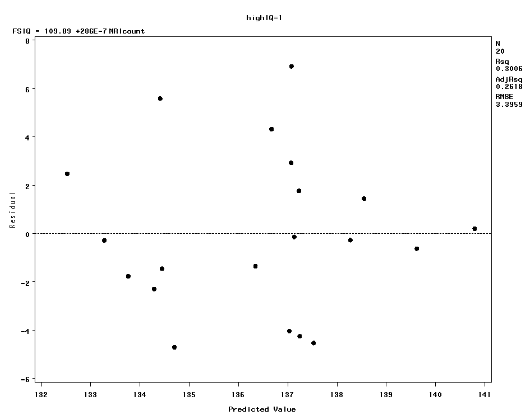
(e) (3 marks) The p-value of 0.0123 appears twice in the SAS output. Explain clearly how the test statistics are related for these data and show that this relationship holds for all simple linear regressions.

(f) (4 marks) Calculate a 90% interval estimate of `FSIQ` for an additional student whose `MRIcount` is 1,025,948.

(g) (2 marks) In addition to the student considered in part (f), suppose we also need an interval estimate of `FSIQ` for another student who has an `MRIcount` of 825,948. For these two intervals, a simultaneous confidence level of 90% is required. How should the interval for the student considered in part (f) be adjusted to take into account the fact that you are now interested in two additional students? (Note that you do not need to calculate anything for the second additional student and you do not need to re-calculate the interval from part (f), just explain how it would change.)

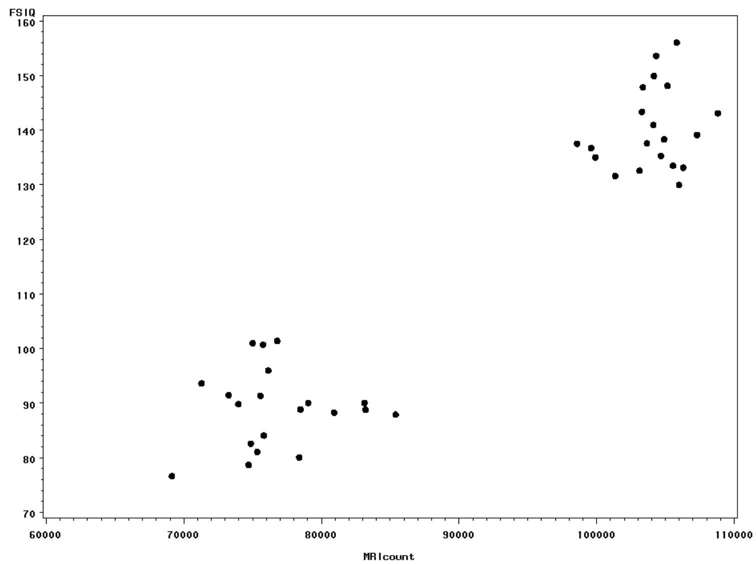
(h) (3 marks) Explain clearly what it means for the intervals discussed in part (g) to be "simultaneous".

(i) (5 marks) The following is a plot of the residuals versus the predicted values.



What assumptions can be evaluated from this plot? What do you conclude?

(j) Suppose the data were collected for 40 students, 20 with high IQs and 20 with low IQs and suppose the scatterplot of `FSIQ` versus `MRIcount` for all 40 observations looks like the following plot.



A regression line is fit to these 40 points.

i. (2 marks) Explain why you would expect a high R^2.

ii. (3 marks) In fact, R^2 for the 40 observations in this plot is 85%. Considering only this fact and the plot above, is this evidence of a strong linear relationship between `FSIQ` and `MRIcount`? Why or why not? Is there any additional information you would like?

3. (6 marks (2 each)) For each of the following statements regarding simple linear regression, state whether you agree or disagree. Briefly explain your choice.

   (a) 95% confidence limits for an intercept were constructed in a regression analysis for a study. The confidence interval may be interpreted as follows: If we were able to repeat the study and the corresponding analysis a large number of times with the same sample size, we would expect that 95% of the resulting estimated intercepts would fall in the original confidence interval.

   (b) The sample mean of the residuals always equals the true mean of the error term.

   (c) For the least squares method to be valid, the error terms ϵ_i must be normally distributed with mean zero and common variance σ^2.