STA 302 H1F / 1001 HF – Fall 2007

Test 1

October 24, 2007

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 8).
- Total points: 50

Some formulae:

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

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

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

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

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

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

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.

$$\begin{aligned}\sum_{i=1}^{n} e_i \hat{Y}_i &= \sum_{i=1}^{n} e_i \left(b_0 + b_1 X_i\right) \\ &= b_0 \sum_{i=1}^{n} e_i + b_1 \sum_{i=1}^{n} e_i X_i \\ &= 0\end{aligned}$$

(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.

$$\begin{aligned}SSTO &= \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2 \\ &= \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}_i)\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}_i) &= \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y}) e_i \\ &= \sum_{i=1}^{n} e_i \hat{Y}_i - \overline{Y} \sum_{i=1}^{n} e_i = 0\end{aligned}$$

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

(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.)
From the formula for b_0*:* $\overline{Y} = b_0 + b_1 \overline{X}$
So

$$\begin{aligned}Var(\overline{Y}) &= Var(b_0) + \overline{X}^2\, Var(b_1) + 2\overline{X}\, Cov(b_0, b_1) \\ \frac{\sigma^2}{n} &= \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right) + \overline{X}^2 \frac{\sigma^2}{S_{XX}} + 2\overline{X}\, Cov(b_0, b_1)\end{aligned}$$

Rearranging gives $Cov(b_0, b_1) = -\frac{\sigma^2 \overline{X}}{S_{XX}}$ *where* $S_{XX} = \sum_{i=1}^{n} (X_i - \overline{X})^2$

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
                                    Uncorrected                     Standard
Variable            Sum        Mean          SS      Variance      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
                                   Sum of          Mean
Source                  DF        Squares        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
                          Parameter       Standard
     Variable     DF       Estimate          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 <u>30.06%</u>

the FSIQ for a person whose MRIcount is 1,000,000 <u>138.5</u>

the error variance <u>11.532</u>

the variance of the slope <u>1.06×10^{-10}</u>

the difference in FSIQ between 2 people when the second person has MRIcount that is 10,000 units larger than the first person <u>0.286</u>

(Question 2 continued.)

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

It has no practical meaning in this context because an `MRIcount` *of 0 is impossible.*

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

$t_{18,.025} = 2.101$
CI: $0.00002863 \pm 2.101(0.00001029) = (0.00000701, 0.0000502)$

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

$H_0 : \ \beta_1 = 0$ *versus* $H_a : \ \beta_1 \neq 0$

(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.

The test statistics are $t_{obs} = 2.78$ *and* $F_{obs} = 7.74$ *and* $2.78^2 = 7.74$ *(within round-off error)*
In general,

$$t_{obs}^2 = \left(\frac{b_1}{s.e. \ \ of \ b_1}\right)^2 = \left(\frac{b_1}{\sqrt{MSE/S_{XX}}}\right)^2 = \frac{b_1^2 S_{XX}}{MSE} = \frac{SSR}{MSE} = F_{obs}$$

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

$\hat{FSIQ} = 109.89 + 0.0000286(1025948) = 139.27$
$S_{XX} = (n-1)s_X^2 = 19(5730703420) = 108883364980$
$t_{18,.05} = 1.734$
PI: $139.37 \pm 1.734(3.396)\sqrt{1 + \frac{1}{20} + \frac{(1025948-925948)^2}{108883364980}} = (132.97, \ 145.56)$

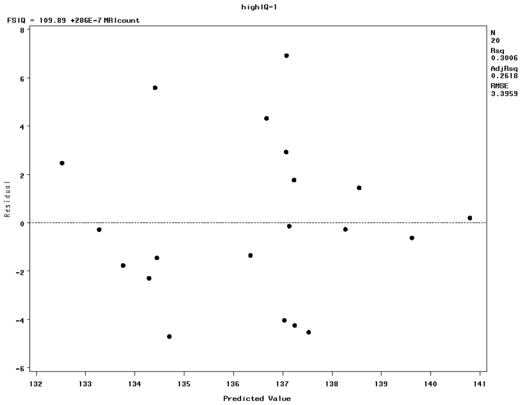
(g) (2 marks) In addition to the student considered in part (f), suppose we also need an interval estimate of `FSIQ` for another student who has an `MRIcount` of 825,948. For these two intervals, a simultaneous confidence level of 90% is required. How should the interval for the student considered in 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.)

Use the Bonferroni method; that is, make each interval at confidence level 95%. So the only adjustment necessary is to substitute $t_{18,.05}$ *with* $t_{18,.025} = 2.101$.

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

In repeated samples of 20 students with the prediction intervals re-calculated for each sample, at least 90% of the pairs of prediction intervals will both capture the actual FSIQ of the 2 additional students.

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

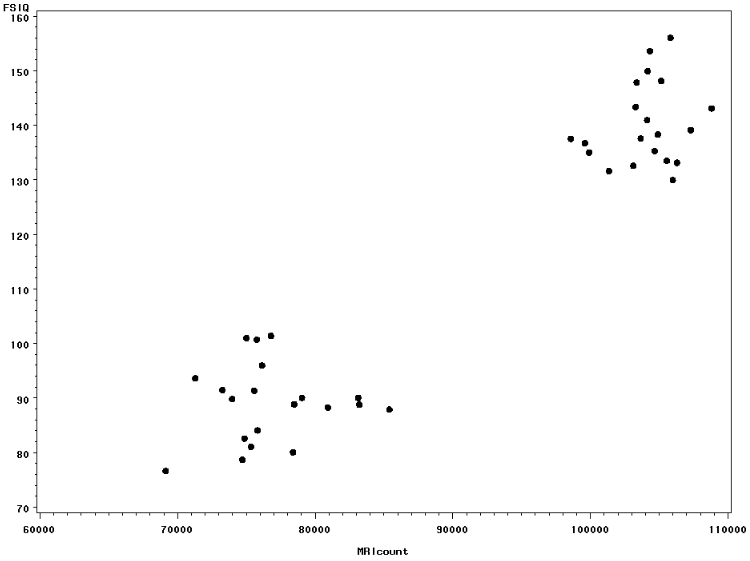


What assumptions can be evaluated from this plot? What do you conclude?

Assumptions:
- linear model is appropriate (no curvature, outliers, influential points)
- variance is constant
Both of these assumptions appear to be satisfied.

(Question 2 continued.)

(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.

There is lots of variation in the Y's, particularly between the two groups. But a regression line (which will roughly connect the centres of the 2 groups) accounts for much of it.

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?

It is not evidence of a strong linear relationship as the line just connects the 2 clumps of points and may not fit either clump well.
I'd like to see a separate regression line fit to each group and evaluate each of them for fit.

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.

   DISAGREE
   We expect that 95% of the confidence intervals will include the true (unknown) intercept.

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

   AGREE
   The true mean of the errors is $E(\epsilon) = 0$ *and we have shown that* $\sum_{i=1}^{n} e_i = 0$ *so* $\overline{e} = 0$.

   (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.

   DISAGREE
   For least squares the only assumption is that the relationship is linear.