STA 302 / 1001 H - Fall 2005
Test 1
October 19, 2005

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:

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

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

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

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

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

$$r = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2}} \qquad \text{Working-Hotelling coefficient: } W = \sqrt{2\,F_{2,n-2;1-\alpha}}$$

1	2	3	4abc	4def	4ghi
10	8	7	11	9	5

1. (10 points) A simple linear regression model is fit on n observed data points.

   (a) What is the difference between β_1 and b_1?
   (3 marks)
   β_1: slope of model, unobserved parameter
   b_1: estimate of β_1, calculated from data

   (b) What does it mean if $R^2 = 1$?
   (1 mark)
   Data points fit exactly on a line.

   (c) In lecture we showed $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} e_i X_i = 0$. Show that $\sum_{i=1}^{n} e_i \hat{Y}_i = 0$. (You may use the results shown in class if they are helpful.)
   (2 marks)

$$\begin{aligned}\sum e_i \hat{Y}_i &= \sum e_i (b_0 + b_1 X_i) \\ &= b_0 \sum e_i + b_1 \sum e_i X_i \\ &= 0\end{aligned}$$

   (d) Explain why the result in (c) implies that the residuals and predicted values are uncorrelated and why this is useful.
   (4 marks)

$$r = \frac{\sum e_i \hat{Y}_i - n\overline{e}\overline{\hat{Y}}}{\sqrt{\sum (e_i - \overline{e})^2 \sum (\hat{Y}_i - \overline{\hat{Y}})^2}} = 0$$

   using $\overline{e} = 0$ since $\sum e_i = 0$

   This is useful for residual plots since we then don't expect a pattern in the plot of the e_i's versus the $\hat{Y}_i$'s.

2. (8 points) In order to carry out linear regression analyses, in addition to the assumption that a linear model is appropriate for the data, we have made the following assumptions:

- the expectation of the random errors is zero
- the variance of the errors is constant
- the errors are uncorrelated
- the errors are normally distributed

Assume that the independent variable is not random.

(a) Which of these additional assumptions are necessary to show that b_1 is unbiased for β_1?
(2 marks – 1 for assumption, 1 for not stating unnecessary assumptions)
$E(\epsilon) = 0$

(b) Derive the formula for the variance of b_1 and state which of the additional assumptions are necessary for the derivation.
(6 marks – 3 for derivation, 2 for necessary assumptions, 1 for not stating unncessary assumptions)

$$
\begin{aligned}
Var(b_1) &= Var\left(\frac{\sum X_i Y_i - n\overline{XY}}{S_{XX}}\right) \\
&= Var\left(\frac{\sum (X_i Y_i - \overline{X} Y_i)}{S_{XX}}\right) \\
&= \frac{1}{S_{XX}^2} \sum Var[(X_i - \overline{X})Y_i] \\
&= \frac{\sum (X_i - \overline{X})^2}{S_{XX}^2} Var(Y_i) \\
&= \frac{\sigma^2}{S_{XX}}
\end{aligned}
$$

where $S_{XX} = \sum (X_i - \overline{X})^2$.

Assumptions: errors uncorrelated and variance constant.

3. (7 points) In lecture we have considered the Snow Gauge example. In this experiment, scientists measured the number of gamma rays (the `gain`) that make it through 10 samples of each of 9 densities of polystyrene. We fit a simple regression model with the logarithm of gain (`loggain`) as the dependent variable and `density` as the independent variable to these 90 points. A scientist argues that, since 10 samples were measured at each density, taking the mean of `loggain` at that density will result in a better estimate and the regression should then be run using the 9 resulting points. Will the least squares estimates of the slope and intercept change? Will the estimate of the error variance change? If there is a change, say whether it is larger or smaller. Justify your answers.

$\overline{Y}$ will not change, neither will $\overline{X}$ so $b_0 = \overline{Y} - b_1\overline{X}$ won't change unless b_1 changes.
S_{XX} will be 10 times larger than for value based on means.
For one of the X_i's:

$$\begin{aligned}
\sum_{\textit{these 10 points}} (Y_i - \overline{Y})(\textit{this } X - \overline{X}) &= (\textit{this } X - \overline{X}) \sum_{\textit{these 10 points}} (Y_i - \overline{Y}) \\
&= (\textit{this } X - \overline{X})(10 \times (\textit{mean of } Y \textit{ for this } X) - 10\overline{Y} \\
&= \textit{10 times value using means}
\end{aligned}$$

So $b_1 = \frac{\sum (Y_i - \overline{Y})(X_i - \overline{X})}{S_{XX}}$ *does not change.*

The estimated error variance, s^2, will be larger for the regression not based on means. s^2 is an estimate of the variability in Y after the variation due to X has been controlled for. A mean of 10 Y's will be less variable than individual observations.

4. (25 points) The data analysed in this question are from a random sample of records of esales of homes in 1993 in the U.S. city of Albuquerque. The data collected include many variables about the homes sold, but we will only consider how well the size of the home (in square feet of usable floor space, variable name: `sqft`) can be used to predict the selling price (in hundreds of dollars, variable name: `price`) of the home.

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

Descriptive Statistics

Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	116.00000	1.00000	116.00000	0	0
sqft	189751	1635.78448	337777165	238134	487.98988
price	123045	1060.73276	147252397	145518	381.46781

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	(A)	13229494	(B)	430.28	(C)
Error	114	(D)	30746		
Corrected Total	(E)	16734535			

Root MSE	(F)	R-Square	0.7906
Dependent Mean	1060.73276	Adj R-Sq	0.7887
Coeff Var	16.53058		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-76.20835	57.17689	-1.33	0.1852
sqft	1	0.69504	0.03351	(G)	<.0001

(a) Find the 7 missing values (A through G) in the SAS output.
(7 marks)
A=1
B=13229494
C$< .0001$
D=3505041
E=115
F=$\sqrt{30746} = 175.3$
G=20.74

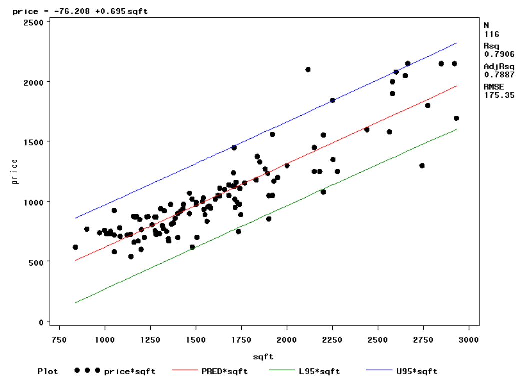
(b) How many houses are in the sample?
(1 mark)
116

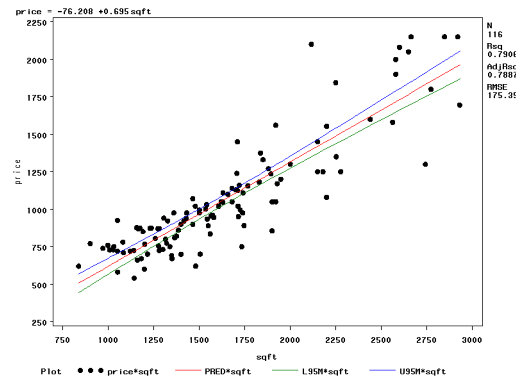
(c) Is the intercept statistically significantly different from 0? Justify your answer. Explain the meaning of the intercept for a real estate agent.
(3 marks)
Not statistically significantly different from 0: p-value for test=.1852.
No meaning: there are no houses with 0 square feet.

(d) A house with 2000 square feet of usable space came on the market (under the same market conditions as the houses used in this analysis). Predict its selling price.
(1 mark)
$\hat{Y} = -76.20835 + .69504(2000) = 1313.9$

(e) What is the standard error of the prediction in part (d)?
(3 marks)
$\sqrt{30746}\sqrt{1 + \frac{1}{116} + \frac{(2000-1635.78)^2}{337777165-116(1635.78)^2}} = 176.52$

(f) Plots of the data including the regression line and 95% confidence intervals for the mean of Y and 95% prediction intervals for Y are given below.

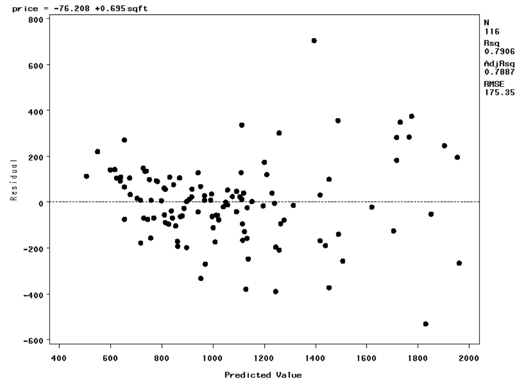




i. Which plot is which? How do you know?
(2 marks)
Plot on the right is CI for mean of Y.
The mean of Y has smaller variance than a prediction which leads to a narrower interval.

ii. For the plot on the right, show how to calculate the value on the lowest curve corresponding to $X = 1500$. In your answer include the numeric value.
(3 marks)
$\hat{Y} = -76.20835 + .69504(1500) = 966.35$
$t_{114,.025} \doteq 2.000$
Value on plot is $966.35 - 2.0\sqrt{30746}\sqrt{\frac{1}{116} + \frac{(1500-1635.78)^2}{337777165-116(1635.78)^2}}$
$= 966.35 - 2.0(16.90)$
$= 932.55$

(g) A plot of the residuals versus predicted values is below.



Describe any problems you see in the residual plot. If the plot shows that any assumptions are being violated indicate which.
(2 marks)
Increasing variance violating constant variance of error.

(h) A student hired by the real estate board to analyse these data argues that we should consider correlation rather than regression since the predictor variable is random. Respond to this comment.
(2 marks)
Fitting regression model is OK since there is a clear choice of dependent/independent variable.

(i) Several of the homes in the random sample used in this analysis were from a new housing development. Why should this be considered in carrying out the analyses?
(1 mark)
Measurements will be correlated.