STA 302 H1F / 1001 HF – Fall 2010
Test
October 21, 2010

LAST NAME: SOLUTIONS FIRST NAME:

STUDENT NUMBER:

INSTRUCTIONS:
- Time: 90 minutes
- Aids allowed: calculator.
- All of the formulae below can be taken as known unless a question indicates otherwise.
- Total points: 50

Some formulae:

$$b_1 = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2} = \frac{\sum x_i y_i - n\overline{x}\overline{y}}{\sum x_i^2 - n\overline{x}^2} \qquad b_0 = \overline{y} - b_1\overline{x}$$

$$\mathrm{Var}(\hat{\beta}_1 \mid X) = \frac{\sigma^2}{\sum(x_i-\overline{x})^2} \qquad \mathrm{Var}(\hat{\beta}_0 \mid X) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i-\overline{x})^2}\right)$$

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1 \mid X) = -\frac{\sigma^2\overline{x}}{\sum(x_i-\overline{x})^2} \qquad \mathrm{SST} = \sum(y_i-\overline{y})^2$$

$$\mathrm{RSS} = \sum(y_i-\hat{y}_i)^2 \qquad \mathrm{SSReg} = b_1^2\sum(x_i-\overline{x})^2 = \sum(\hat{y}_i-\overline{y})^2$$

$$\mathrm{Var}(\hat{y}|X=x^*) = \sigma^2\left(\frac{1}{n} + \frac{(x^*-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right) \qquad \mathrm{Var}(Y-\hat{y}|X=x^*) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x^*-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)$$

$$r = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2}} \qquad SXX = \sum(x_i-\overline{x})^2 = \sum x_i^2 - n\overline{x}^2$$

$$h_{ij} = \frac{1}{n} + \frac{(x_i-\overline{x})(x_j-\overline{x})}{SXX} \quad \left(h_{ii} > \frac{4}{n}\right) \qquad \mathrm{DFBETAS}_{ik} = \frac{b_k - b_{k(i)}}{s.e.(b_k)} \quad \left(> 1 \text{ or } \frac{2}{\sqrt{n}}\right)$$

$$\mathrm{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{s.e.(\hat{y}_i)} \quad \left(> 1 \text{ or } 2\sqrt{\frac{2}{n}}\right) \qquad D_i = \frac{\sum(\hat{y}_{j(i)}-\hat{y}_j)^2}{2S^2} \quad \left(> \frac{4}{n-2}\right)$$

1abc	1d	2	3abc	3def	3gh, 4

1. Consider the example we have been examining in lecture in which we have been using simple linear regression to model how the atmospheric concentrations of CFC-11 (in parts per trillion) were changing as a function of time (in years) before the implementation of the Montreal Protocol. The data consist of 153 measurements of CFC-11 taken monthly from 1977 to the end of 1989 and the date on which the measurements were taken.

   (a) (2 marks) State the simple linear regression model being used. Which terms in the model are random variables?

   Model: $Y = \beta_0 + \beta_1 x + e$
   Random: e *and* Y
   x *is not random here since it is planned times*

   (b) (3 marks) State the Gauss-Markov conditions for the model in part (a).

   $E(e) = 0$
   $Var(e) = \sigma^2$ *(same variance for all observations)*
   e*'s uncorrelated for different observations*

   (c) (4 marks) Assume that the Gauss-Markov conditions and the usual distributional assumptions hold. State fully the distributions of the random variables in the model. How do the distributions change with time?

   $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$
   The mean differs with time (x) *but the variance doesn't.*

   $e \sim N(0, \sigma^2)$
   Distribution does not differ with time.

(d) Suppose that the regression model has been fit to the data and the usual statistics have been calculated.

i. (3 marks) Show that $\sum_{i=1}^{n} \hat{e}_i x_i = 0$.

$$\begin{aligned}
\sum (y_i - \hat{y}_i) x_i &= \sum (y_i - b_0 - b_1 x_i) x_i \\
&= \sum x_i y_i - n\overline{xy} + b_1 n\overline{x}^2 - b_1 \sum x_i^2 \\
&= (\sum x_i y_i - n\overline{xy}) - \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\overline{x}^2} (\sum x_i^2 - n\overline{x}^2) \\
&= 0
\end{aligned}$$

ii. (3 marks) Show that estimator of the slope of the regression line is unbiased.

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_1) &= E\left(\frac{\sum x_i Y_i - \overline{x} \sum Y_i}{SXX}\right) \\
&= \frac{1}{SXX}\left(\sum x_i \mathrm{E}(Y_i) - \overline{x} \sum \mathrm{E}(Y_i)\right) \\
&= \frac{1}{SXX}\left(\sum x_i (\beta_0 + \beta_1 x_i) - \overline{x} \sum (\beta_0 + \beta_1 x_i)\right) \\
&= \frac{1}{SXX}\left(n\beta_0 \overline{x} + \beta_1 \sum x_i^2 - n\beta_0 \overline{x} - n\beta_1 \overline{x}^2\right) \\
&= \frac{1}{SXX}\left(\beta_1 \, SXX\right) \\
&= \beta_1
\end{aligned}$$

2. Suppose a simple linear regression is carried out to investigate the relationship between a dependent variable Y and an independent variable X. The data consist of n pairs of observed values of X and Y, (x_i, y_i), $i = 1, \ldots, n$.

   (a) (2 marks) What is the first step you should carry out in the regression analysis? What do you hope to accomplish in this step?

   Plot the data to see if a linear regression model appears to be appropriate for the data.

   (b) (1 mark) As part of the output for the simple linear regression analysis, SAS gives the results of the statistical test with null hypothesis $H_0 : \beta_1 = 0$ and alternative hypothesis $H_a : \beta_1 \neq 0$. Why is this test of particular interest?

   If $\beta_1 = 0$ then there is no linear relationship between X and Y.

   (c) You suspect that there is a strong linear relationship between Y and X.

      i. (2 marks) For the test with null hypothesis $H_0 : \beta_1 = 0$, do you expect the test statistic to be large or small? Explain.

      Large in absolute value. The test statistics is $\frac{b_1}{s.e.(b_1)}$ and if there is a strong linear relationship you expect the estimate of the slope to be large relative to its standard error.

      ii. (2 marks) For the test with null hypothesis $H_0 : \beta_1 = 0$, do you expect the p-value to be large or small? Explain.

      Small. A large test statistic will be out in the tails of the distribution resulting in a small p-value. OR A small p-value gives evidence that the slope is not 0 consistent with the suspicion that there is a strong linear relationship.

3. In a paper published in the *British Medical Journal* in 1965, Lea looked at data from counties in regions of Great Britain, Norway, and Sweden. He was interested in how the mean annual temperature (in degrees Fahrenheit) affected the mortality index for breast cancer. (The mortality index is a measure of the death rate for women diagnosed with breast cancer. The index Lea used measures death rate relative to the average death rate for England and Wales. On his scale, England and Wales was given the value of 100. Mortality indices greater than 100 indicate a higher death rate than that of England and Wales.)

   Here are some quantiles from t-distributions which may be useful for some of the questions that follow.

Degrees of	Upper-tail probability				
freedom	0.005	0.010	0.025	0.05	0.10
14	2.977	2.624	2.145	1.761	1.345
15	2.947	2.602	2.131	1.753	1.341
16	2.921	2.583	2.120	1.746	1.337

   Some SAS output is given below and on the next page for the analysis Lea carried out.

The REG Procedure

Descriptive Statistics

Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	16.00000	1.00000	16.00000	0	0
temperature	713.50000	44.59375	32285	31.17663	5.58360
mortality	1333.50000	83.34375	114535	226.42929	15.04757

Dependent Variable: mortality

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2599.53358	2599.53358	(A)	<.0001
Error	14	796.90580	56.92184		
Corrected Total	15	3396.43938			

Root MSE	7.54466	R-Square	0.7654
Dependent Mean	83.34375	Adj R-Sq	0.7486
Coeff Var	9.05246		

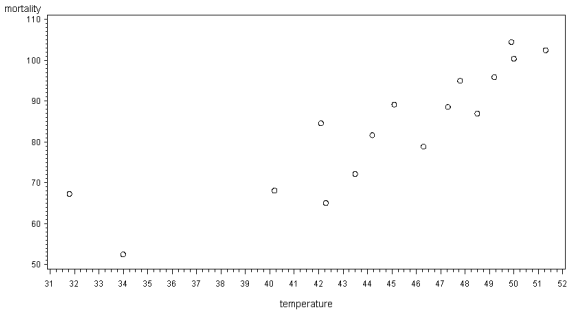
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-21.79469	15.67190	-1.39	0.1860
temperature	1	2.35769	0.34888	(B)	<.0001

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual
1	102.5000	99.1550	3.0053	3.3450	6.920	0.483
2	104.5000	95.8543	2.6429	8.6457	7.067	1.223
3	100.4000	96.0900	2.6674	4.3100	7.057	0.611
4	95.9000	94.2039	2.4779	1.6961	7.126	0.238
5	87.0000	92.5535	2.3270	-5.5535	7.177	-0.774
6	95.0000	90.9031	2.1929	4.0969	7.219	0.568
7	88.6000	89.7243	2.1093	-1.1243	7.244	-0.155
8	89.2000	84.5373	1.8944	4.6627	7.303	0.638
9	78.9000	87.3666	1.9779	-8.4666	7.281	-1.163
10	84.6000	77.4642	2.0772	7.1358	7.253	0.984
11	81.7000	82.4154	1.8912	-0.7154	7.304	-0.0980
12	72.2000	80.7650	1.9244	-8.5650	7.295	-1.174
13	65.1000	77.9358	2.0489	-12.8358	7.261	-1.768
14	68.1000	72.9846	2.4305	-4.8846	7.142	-0.684
15	67.3000	53.1800	4.8457	14.1200	5.783	2.442
16	52.5000	58.3669	4.1494	-5.8669	6.301	-0.931

Obs	-2-1 0 1 2	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS
1	| |	0.022	0.4697	0.1587	1.3329	0.2040
2	| |**	0.105	1.2475	0.1227	1.0544	0.4666
3	| |*	0.027	0.5965	0.1250	1.2558	0.2254
4	| |	0.003	0.2298	0.1079	1.2895	0.0799
5	| *|	0.031	-0.7621	0.0951	1.1744	-0.2471
6	| |*	0.015	0.5533	0.0845	1.2092	0.1681
7	| |	0.001	-0.1497	0.0782	1.2538	-0.0436
8	| |*	0.014	0.6244	0.0630	1.1668	0.1620
9	| **|	0.050	-1.1789	0.0687	1.0164	-0.3203
10	| |*	0.040	0.9826	0.0758	1.0874	0.2814
11	| |	0.000	-0.0944	0.0628	1.2358	-0.0244
12	| **|	0.048	-1.1915	0.0651	1.0082	-0.3143
13	| ***|	0.124	-1.9327	0.0738	0.7555	-0.5454
14	| *|	0.027	-0.6703	0.1038	1.2090	-0.2281
15	| |****	2.093	3.1052	0.4125	0.6507	2.6020
16	| *|	0.188	-0.9264	0.3025	1.4632	-0.6100



Questions related to this output begin on the next page.

(Question 3 continued.)

(a) (5 marks) What are the values of each of the following:

- the number of observations 16

- the number replaced by (A) in the SAS output 45.67

- the number replaced by (B) in the SAS output 6.76

- the estimate of the correlation between
mortality index and mean annual temperature 0.875

- the estimate of the variance of the error 56.92184

(b) (4 marks) Give two different numbers from the SAS output that give some indication of the strength of the linear relationship between mortality index and mean annual temperature. For each number, state what it measures. Do not choose numbers that are missing from the output and do not choose two numbers that are equal.

1. the p-value(< 0.0001) for the two-sided test with null hypothesis that the slope is 0
What this measures: assuming that the slope is 0, this is the probability of getting the value we got or a value of the test statistic (estimated slope divided by its s.e.) further from 0.

2. R^2 (0.7654)
What this measures: the proportion of variation in mortality index that is explained by its linear relationship with mean annual temperature.

(c) For the test with p-value 0.1860:

i. (1 mark) What are the null and alternative hypotheses?

$H_0 : \beta_0 = 0$ *versus* $H_a : \beta_0 \neq 0$

ii. (2 marks) What do you conclude? State your conclusion in the practical context of the data being analysed.

The data are consistent with possibly having a 0 intercept. There is no practical interpretation since a mean annual temperature of 0 is well outside the range of the data.

(d) (3 marks) Calculate a 90% confidence interval for the slope. How is it related to your answers to parts (b) and/or (c)?

$t_{14,.05} = 1.761$
90% CI for the slope: $2.35769 \pm 1.761 * 0.34888 = (1.743, 2.972)$
The confidence interval does not contain 0, which is consistent with the p-value in (b) which showed strong evidence that the slope is not 0.

(e) (2 marks) DFFITS is given in the output statistics and its formula is given on the first page. Explain what it measures.

$DFFITS_i$ measures how the fitted value of the ith observation changes when the ith observation is part of the data and when the regression line has been calculated without it in the data (scaled by it's s.e.).

(f) (4 marks) Based on the given output statistics, what concerns do you have about the fit of the regression line to the data? Give at least two numbers in the output that indicate that this concern exists. Draw a sketch that illustrates the implications of your concern.

Obervation 15 is influential. $DFFITS_{15} = 2.6020 > 1$ *(using the small dataset cut-off) and Cook's distance for observation 15 is* $2.093 > 4/14 = 0.286$.
For the plot, you should give an indication of how the regression line changes with and without observation 15 which is the point that has the smallest temperature. Observation 15 pulls the line to it, resulting in a fitted line with smaller slope.

(Question 3 continued.)

(g) (4 marks) Calculate a 95% prediction interval for the predicted mortality rate for a county with a mean annual temperature of 37 degrees Fahrenheit.

$\hat{y} = -21.79469 + 2.35769(37) = 65.44$
$t_{14,.025} = 2.145$
95% PI at 37 degrees: $65.44 \pm 2.145 * 7.54466\sqrt{1 + \frac{1}{16} + \frac{(37-44.59375)^2}{15*31.17663}} = (47.82, 83.06)$

(h) (1 mark) How does your answer to part (f) affect your interpretation of the prediction interval in part (g)?

I do not trust the prediction because it would be quite different without the influential point in the model.

4. (2 marks) A study was carried out to examine the effect of taking Vitamin D tablets on levels of LDL in the blood. (LDL is "bad" cholesterol and is measured in mg/dL.) Thirty subjects who were taking no medication for their cholesterol were recruited into the study and their LDL was measured. They then took Vitamin D tablets for 30 days while otherwise maintaining their usual diet and their LDL was measured again. A two-sided t-test with null hypothesis that the mean is 0 for the before-after change in LDL had p-value 0.9207. The researchers also fit a simple linear regression model with response variable LDL after 30 days and explanatory variable LDL at the start of the study. The fitted regression line had intercept 13.0 and slope 0.887. The p-value was < 0.001 for the two-sided test with null hypothesis that the slope is 0. The researchers concluded the following: "On average, LDL did not differ with Vitamin D intake, but, importantly, for subjects with higher LDL values than the population average of 115 mg/dL, LDL levels were lower on average after taking Vitamin D with greater reductions for patients with higher initial LDL levels."

   Is the researchers' conclusion supported by the analysis? Why or why not?

   No. This is an example of regression to the mean. We expect high initial values to have lower values after taking vitamin D.