STA 302 H1F / 1001 HF – Fall 2009
Test
October 22, 2009

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 10).
- Total points: 50

Some formulae:

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

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

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

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

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

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

1abc	1de	2a	2bcd	2efgh	3ab	3c	3de

1. The method of least squares is used to fit a simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$ to n observations (X_i, Y_i). The values of the X_i's are not realizations of random variables but are fixed in advance by the researcher. Assume the following: the form of the model is appropriate, the Gauss-Markov conditions hold, and the distribution of the error terms is Normal. In your answers, you may use any of the formulae on the front page.

   (a) (4 marks) Which of the assumptions are needed to fit the model using least squares? How would you assess the necessary assumptions?

   For least squares, the only assumption needed is the form of the model.
   To check this look at plots of Y versus X and the residuals versus either X or the predicted values. Look for outliers / influential points and curvature.

   (b) (3 marks) What is $\mathrm{E}(Y)$? Which of the assumptions did you use to determine your answer?

   $E(Y) = \beta_0 + \beta_1 X$
   Assumptions used: form of model and $E(\epsilon) = 0$

   (c) (2 marks) Suppose the researcher is interested in the relationship between X and Y on a certain range of X's. She uses the smallest value in the range of X for half of the observations and the largest value in the range of X for the other half of the observations and fits the simple linear regression model to the resulting data. What is the advantage of fixing the X's to be these values? What is the disadvantage?

   Advantage: This choice of X's will make S_{XX} as large as possible, giving more precise estimates (smaller variance) of the model parameters.
   Disadvantage: Won't be able to determine if the form of the relationship really is linear without observations on more values of X.

(Question 1 continued)

(d) (4 marks) Show that the least squares estimator b_1 is an unbiased estimator of β_1.

$$\begin{aligned}
E(b_1) &= E\left(\frac{\sum X_i Y_i - n\overline{XY}}{S_{XX}}\right) \\
&= \frac{1}{S_{XX}}\left(\sum X_i E(Y_i) - n\overline{X}\frac{1}{n}\sum E(Y_i)\right) \\
&= \frac{1}{S_{XX}}\left(\sum X_i(\beta_0 + \beta_1 X_i) - \overline{X}\sum(\beta_0 + \beta_1 X_i)\right) \\
&= \frac{1}{S_{XX}}\left(\beta_0 n\overline{X} + \beta_1 \sum X_i^2 - \overline{X}n\beta_0 - \beta_1\overline{X}n\overline{X}\right) \\
&= \frac{\beta_1}{S_{XX}}\left(\sum X_i^2 - n\overline{X}^2\right) \\
&= \beta_1
\end{aligned}$$

(e) (2 marks) Are b_0 (the estimator of the intercept) and b_1 (the estimator of the slope) independent? Explain.

No. Their covariance is not 0 (unless $\overline{X} = 0$*).*

2. CFC-11 atmospheric concentrations in parts per trillion were measured monthly. The following SAS output shows the results of the regression of atmospheric concentration on time (in years) for the period 1977 to 1989. On the next page is SAS output for the regression for the period 1995 to 2004. Some of the output has been removed and, in the first regression, some of the numerical values have been replaced by letters. Answer the questions assuming that the usual regression model assumptions hold.

```
                 Before Montreal Protocol (before January 1990)

                               Descriptive Statistics

                                            Uncorrected                       Standard
Variable           Sum          Mean            SS          Variance       Deviation
Intercept    153.00000       1.00000     153.00000              0               0
time            303466    1983.43464     601906139       14.16876         3.76414
cfc11            30286     197.94771       6199037     1342.05751        36.63410


                               The REG Procedure
                            Dependent Variable: cfc11

                                  Analysis of Variance
                                   Sum of           Mean
Source                   DF       Squares         Square    F Value    Pr > F
Model                    (A)       203119         203119        (B)    <.0001
Error                    151    874.00094        5.78809
Corrected Total          (C)       203993

             Root MSE                  (D)     R-Square    0.9957
             Dependent Mean      197.94771

                              Parameter Estimates
                            Parameter       Standard
     Variable      DF        Estimate          Error     t Value     Pr > |t|
     Intercept      1          -19064            (E)     -185.40       <.0001
     time           1         9.71152        0.05184      187.33       <.0001
```

(a) (5 marks) Find the values of the numbers that have been replaced by letters:

(A) = 1

(B) = $203119/5.78809 = 35092$

(C) = 152

(D) = $\sqrt{5.78809} = 2.406$

(E) = $-19064/-185.40 = 102.8$

(Question 2 continued)

After Montreal Protocol (after December 1994)

Descriptive Statistics

Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	116.00000	1.00000	116.00000	0	0
time	231985	1999.86710	463939247	7.92451	2.81505
cfc11	30641	264.14741	8096893	27.17849	5.21330

The REG Procedure
Dependent Variable: cfc11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3061.55314	3061.55314	5455.70	<.0001
Error	114	63.97289	0.56117		
Corrected Total	115	3125.52602			

Root MSE	0.74911
Dependent Mean	264.14741

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3929.67750	49.62630	79.19	<.0001
time	1	-1.83289	0.02481	-73.86	<.0001

(b) (2 marks) Calculate a 90% confidence interval for the intercept for the regression for the period 1995 to 2004 (after the Montreal Protocol).

$t_{114,.05} = 1.671$ *(approximating with 60 d.f.)*
Confidence interval: $3929.6775 \pm 1.671(49.62630) = (3846.8, 4012.6)$

(c) (2 marks) For the regression for the time period 1995 to 2004, find R^2 and explain what it measures.

$R^2 = 3061.55314/3125.52602 = 0.9795$
Almost 98% of the variation in CFC-11 concentration for this time period is explained by its linear relationship with time.

(d) (1 mark) Interpret the estimated slope in practical terms.

On average, in this time period CFC-11 is going down 1.83 *parts per trillion per year.*

(Question 2 continued)

(e) (4 marks) Carry out an hypothesis test to determine whether the slopes of the lines for the regressions for the two time periods differ. If you do not have all the information you need to completely answer the question, indicate what is missing and give the most complete answer you can.

s.d. of the difference in the slopes is $\sqrt{0.05184^2 + 0.02481^2} = 0.05747$
Testing $H_0 : \beta_{1,before\ MP} = \beta_{1,after\ MP}$ *versus* $H_a : \beta_{1,before\ MP} \neq \beta_{1,after\ MP}$
Test statistic: $(9.71152 - (-1.83289))/0.05747 = 200.9$
In order to estimate the p-value, we need to know the appropriate degrees of freedom for the t-distribution of the test statistic. However, 200.9 is far in the tails of all t distributions so the p-value will be very small.
So we have strong evidence that the 2 slopes differ.

(f) (4 marks) Use one of the fitted models to predict what the atmospheric concentration of CFC-11 on October 1, 2009 was (when `time` = 2009.75) and give a 99% interval for your prediction.

(Use the second model since it is closer in time to 2009.)
On October 1, 2009, $\widehat{CFC\text{-}11} = 3929.6675 - 1.83289(2009.75) = 246.0$
$t_{114,.005} = 2.660$ *(estimating with 60 d.f.)*
Prediction interval: $246.0 \pm 2.660(0.74911)\sqrt{1 + \frac{1}{116} + \frac{(2009.75-1999.8671)^2}{115(7.9245)}} = (243.9, 248.1)$

(g) (2 marks) Do you feel confident that the actual concentration of CFC-11 measured on October 1, 2009 is in the interval you calculated in part (f)? Why or why not?

No because October 1, 2009 is outside the range of the data and we can't be sure that the linear model is still appropriate after 2004.

(h) (2 marks) Using only what you know about how the data were collected, does it seem possible that there are any violations in the Gauss-Markov conditions for these regressions? Explain.

Yes since the data are collected over time there are likely non-zero correlations in the ϵ_i*'s for observations close together.*

3. Golf tournaments take place over a few days. On each day of the tournament one round of golf is played. In this question, we are looking at the relationship between golfers' scores on the first round and their scores on the second round in the 2000 British Open. In golf, low scores are good. Some output from SAS is given below.

```
                         The REG Procedure
                     Dependent Variable: round2

                        Analysis of Variance
                                Sum of           Mean
Source                DF       Squares         Square    F Value    Pr > F
Model                  1      29.51923       29.51923       3.10    0.0804
Error                154    1467.22436        9.52743
Corrected Total      155    1496.74359

          Root MSE              3.08665    R-Square     0.0197
          Dependent Mean       72.24359    Adj R-Sq     0.0134
          Coeff Var             4.27256

                        Parameter Estimates
                      Parameter       Standard
     Variable     DF   Estimate          Error    t Value    Pr > |t|
     Intercept     1   62.42741        5.58218      11.18      <.0001
     round1        1    0.13449        0.07641       1.76      0.0804
```

(a) (2 marks) Is there evidence of a linear relationship between golf scores on the first and second round of the tournament? Explain.

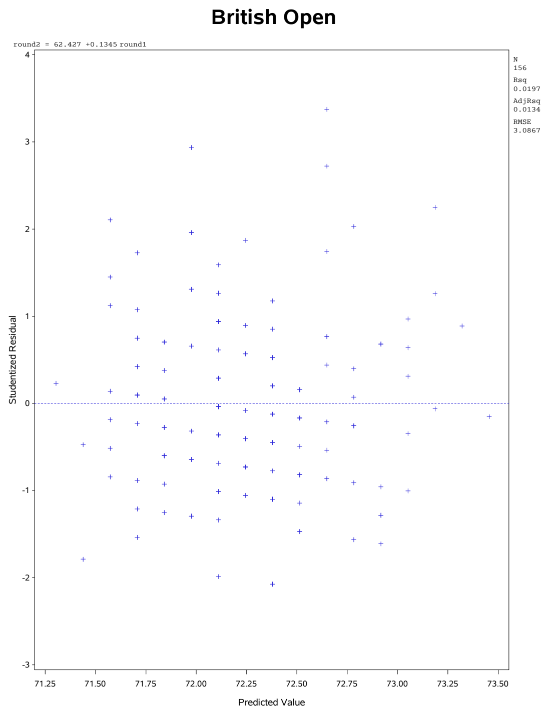
The p-value for the test with null hypothesis that the slope is 0 is 0.0804 so we have only weak evidence of a linear relationship.

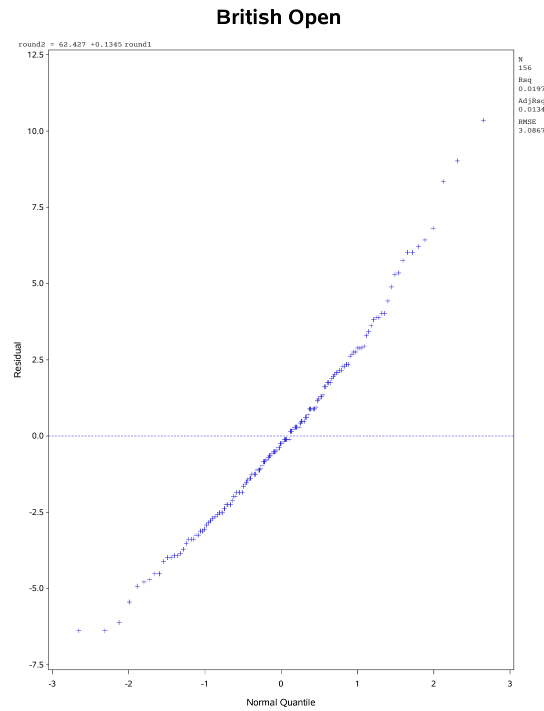
(b) (3 marks) The lowest score obtained on the first round was 66. Predict the second round score of the golfer who achieved this. Is this surprising? Explain your answer in terms of known facts about simple linear regression.

$\widehat{round2} = 62.42741 + 0.13449(66) = 71$

So we expect the second round score to jump to up 71. It is not surprising to see it go up as this is an example of regression to the mean. In any test/re-test situation we expect people with low scores the first time to have higher scores, on average, the second time.

(c) (4 marks) Below are plots of the studentized residuals versus the predicted values, and a normal quantile plot of the residuals. What additional information do you learn from the plots? Be specific.





The first plot looks like random scatter about 0. So there are no problems with: outliers / influential points, curvature, constant variance.

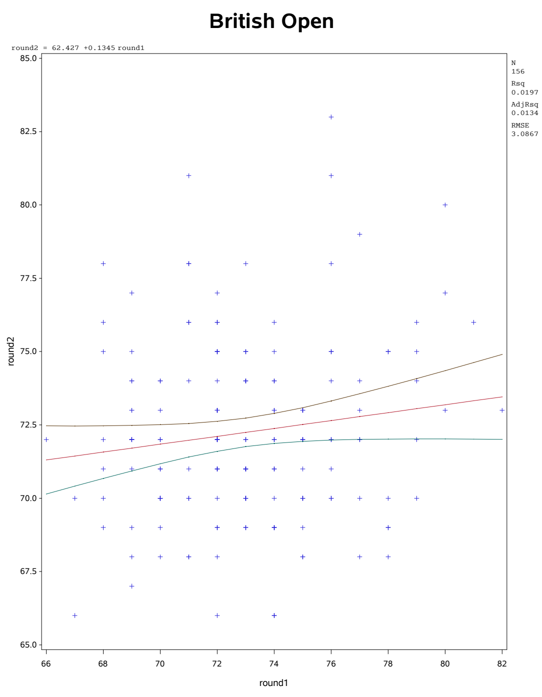
The second plot looks like a straight line so there is no problem with the assumption that the errors are normally distributed.

(Question 3 continued)

(d) (2 marks) In assignment 1 we considered the relationship between football kickers' field goal percentages one year with the percentage of field goals scored the previous year and found problems with violations of the Gauss-Markov conditions in the initial analysis. In the regression here we are examining the relationship between golf scores on one round with golf scores on the previous round. Do we have a similar problem with violations of the Gauss-Markov assumptions? Why or why not?

No because in this example we don't have multiple observations on the same golfers.

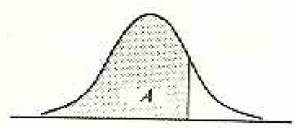
(e) (2 marks) The plot of the data below includes 95% confidence intervals for the mean score in round 2 given the score in round 1. About 90% of the data points fall outside the confidence limits. Explain how it can occur that so many observations are missed.



These are CIs for $E(Y)$ and only take into account error in the estimation of the regression line and not the fact that individual points vary about the line (the ϵ component of the model). We would expect prediction intervals that also include this other variability to capture more of the points.

TABLE B.2 Percentiles of the t Distribution.

Entry is $t(A; \nu)$ where $P\{t(\nu) \leq t(A; \nu)\} = A$



$t(A; \nu)$

ν	A: .60	.70	.80	.85	.90	.95	.975
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

TABLE B.2 *(concluded)* Percentiles of the t Distribution.

ν	A: .98	.985	.99	.9925	.995	.9975	.9995
1	15.895	21.205	31.821	42.434	63.657	127.322	636.590
2	4.849	5.643	6.965	8.073	9.925	14.089	31.598
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.849
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.690
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.659
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	2.054	2.170	2.326	2.432	2.576	2.807	3.291