Name ____Mr. Key____________
ID __7777777________________

ECO 5350
Intro. Econometrics

Prof. T. Fomby
Fall 2007

Mid-Term Exam I

Instructions: Put your name and student ID in the upper right-hand-corner of this exam. This exam is worth a total of 88 points. The breakout of these points by questions is as follows:

Q1 = (a) 4 (b) 2 (c) 3 (d) 1,1,2 (e) 2,2,2 (f) 2,2,2 = 25 points
Q2 = (a) 2 (b) 2 (c) 2 (d) 2 (e) 2 (f) 2 (g) 2 (h) 2 = 16 points
Q3 = (a) 5 (b) 3 (c) 2 (d) 4 = 14 points
Q4 = (a) 5 (b) 4 (c) 2 (d) 2 (e) 4 (f) 4 (g) 2 (h) 4 (i) 2 (j) 4 = 33 points

You have one hour and twenty minutes to take this test. A word from the wise: Don't get hung up on any one question. Answer the easy questions first and then go back and pick up the hard ones. Good luck.

Oh, by the way, here is a bonus questions worth **2** points. How many Aggies does it take to screw in a light bulb? __**5**_____. This time you have to get it exactly right to get any bonus points!

+2

1. Let's start off with some short answer questions.

a) Match the below terms by circling the correct alternative.

1 The $N(\mu,\sigma^2)$ distribution is a (discrete / **continuous**) distribution.

1 The Bernoulli distribution is a (**discrete** / continous) distribution.

The single toss of a coin would best be described by the

1 ($N(\mu,\sigma^2)$ or **Bernoulli**) distribution.

The family incomes of Dallas Families would best be described by the

1 ($N(\mu,\sigma^2)$ or Bernoulli) distribution. **Answer:** $N(\mu,\sigma^2)$

b) In the first part of this course we are going to be focusing on the regression analysis of
1 ___**Cross-Section**__________ data. Later we will focus on the regression analysis of
1 ___**Time Series**__________ data. In a sequel to this course, Eco 6352, we will consider the analysis of panel data.

c) Match up the following data types with an example of the data type.

EXAMPLES:

1 Real GDP observed quarterly From 1900 Q1 to 2000 QIV ____**A**____

1 Employment in each of the 50 States of the Union in January, 1999 ____**C**____

1 Real Per Capita annual growth rates in 10 Countries observed from 1990 – 2000. ____**B**____

POSSIBLE DATA TYPES:

A. Time Series Data, **B.** Panel Data, **C.** Cross-Section Data

d) In a given population of two-earner male/female couples, male earnings have a mean of \$40,000 per year and a standard deviation of \$12,000. Female earnings have a mean of \$45,000 per year and a standard deviation of \$18,000. The correlation between male and female earnings for a couple is 0.80. Let C denote the combined earnings for a randomly selected couple.

1 i) The mean of C is ____**85,000**__________. Show your work below.

Answer:

Let X = Male earnings and Y = Female earnings

E(X+Y) = E(X) + E(Y) = 40K + 45K = 85,000.

ii) The covariance of male and female earnings is __**172,800,000**________.
Show your work below.

1 **Answer: Corr(X,Y) = 0.8 = Cov(X,Y)/(SD(X)SD(Y))**

Therefore, Cov(X,Y) = 0.8(12x10^3)(18x10^3) = 172.8x10^6

iii) The Standard Deviation of C is ____**28,523**______. Show your work below.

Answer: Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)

2 $= (12\cdot 10^3)^2 + (18\cdot 10^3)^2 + 2\cdot 172.8\cdot 10^6 = 813.6\cdot 10^6$

Therefore, SD(C) =

$$\sqrt{Var(X+Y)} = \sqrt{813.6\cdot 10^6} = \sqrt{813.6}\cdot 10^3 = 28.523\cdot 10^3 = 28{,}523$$

e) Define the following terms:

2 i) **Population Regression Function**:
The conditional mean function: E(Y|X) = $\beta_0 + \beta_1 X$

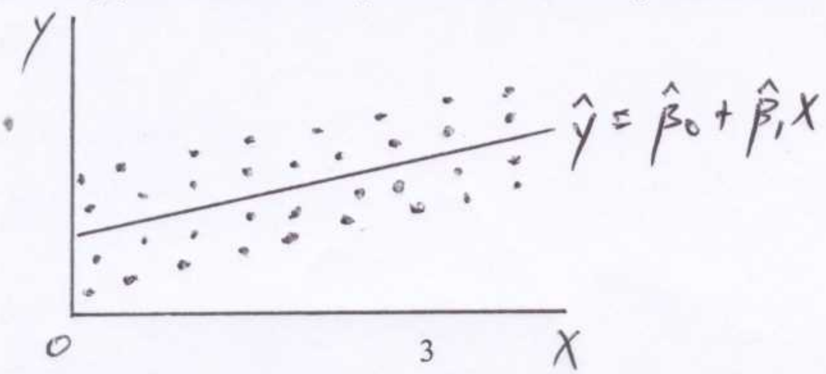
2 ii) **Sample Regression Function**:
The Fitted Regression Line obtained by OLS:
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

2 iii) **Sampling Distribution** of $\hat{\beta}_1$:
The probability distribution of the $\hat{\beta}_1$'s that would be obtained in an infinite number of repeated samples.

f) In the spaces below I want you to

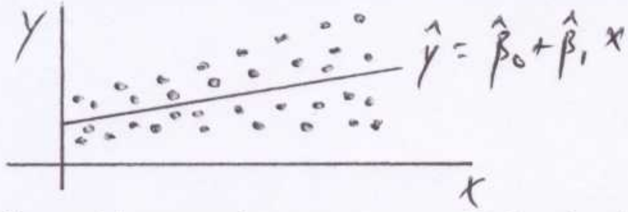
i) Draw a scatter plot of (x,y) values with a regression line through the points that would imply **homoskedasticity** in the errors of a regression model.

2



ii) Draw a scatter plot of (x,y) values with a regression line through the points that would imply **heteroskedasticy** in the errors of a regression model.

2



iii) Briefly explain to me the consequences OLS estimation of coefficients and statistical inference about them if the errors of one's regression model are heteroskedastic.

ANSWER:

2

The OLS estimates are still unbiased but they are no longer efficient. Moreover, the OLS standard errors of the coefficient estimates are no longer appropriate for constructing t-statistics for hypothesis testing purposes. To conduct statistical testing in our regression model we must either use Weighted Least Squares (WLS) or use heteroskedasticity-robust standard errors in calculating our t-statistics.

2. Let's review some of your QQ questions.

2

a. **True** or False: In **experimental data** you usually have a **control group** and a **treatment group** and you want to use the data to determine **the effectiveness of a treatment.** In contrast, **observational data** is collected **without experimental control** through government or telephone surveys, administrative records, or data obtained from government documents. These latter data pose challenges to the econometrician in estimating **causal effects**.

2

b. One of the mainstays in econometrics is
   a. randomized controlled experiments
   b. multiple regression
   c. panel data
   d. simultaneous causality

The following equation represents the so-called **Taylor Rule**:

$$i_t = r_t + \pi_t + \delta(\pi_t - \pi_t^*) + \omega(y_t - y_t^*).$$

2

c. This rule is used to characterize the r _**e**_ _**a**_ _**c**_ _**t**_ _**i**_ _**o**_ _**n**_ function of the Federal Open Market Committee for determining its target for the Fed Funds rate.

2

d. The term $(\pi_t - \pi_t^*)$ is called the __**inflation**_______ gap. The term $(y_t - y_t^*)$ is called the __**output**_________________ gap.

e. One reason Professor Fomby introduced this model in class is because he wanted to discuss two different purposes for which statistical models (like the Taylor Rule) can be used. Two of these purposes are hy _**p**_ _**o**_ _**t**_ _**h**_ _**e**_ _**s**_ _**i**_ _**s**_ testing and p _**r**_ _**e**_ _**d**_ _**i**_ _**c**_ _**t**_ _**i**_ _**o**_ _**n**_. A third purpose would be for simulating future scenarios, as in "favorable," "ordinary," and "bad" times for the economy over the near term.

2

f. Suppose that we observed an exact t-statistic of t = 1.65 when testing two samples to have equal population means and the t-statistic has 20 degrees of freedom. Suppose we know from this information that $\Pr(-\infty < t < 1.65) = 0.95$. Given this information and knowing that we are interested in testing $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ at the 5% level of confidence, we conclude that the probability value of the test statistic for this test is p = ___**0.10**___. Therefore, we (**accept** / reject) the null hypothesis.

2

g. One equivalent way of computing an exact t-statistic for comparing two population means is to run a regression of Y on a _**dummy**___ variable which takes the value of 1 when the observation is from group 1 and 0 when the observation is from group 2. Here Y represents the observations taken on the two populations.

2

h. **True** or False. Among all unbiased linear estimators of population mean μ, the sample mean $\overline{Y}$ is the most efficient estimator of μ. That is, the variance of the sampling distribution of the sample mean is less than the variance of the sampling distributions of all other unbiased linear estimators of μ. This is called the Gauss-Markov theorem for estimating the mean of a population.

2

3. Consider the following **ANOVA table**.

a) Fill in the blanks

5

Source	**SS**	**DF**	**MS**	**F**	**P-Value**
Explained	12	2	_ **6**___	__**3**__	0.03
Residual	_**40**__	_**20**___	___**2**___		
Total	52	22			

b) The number of observations used to generate the above ANOVA table is __**23**_____. The number of explanatory variables (apart from the intercept) in the above regression model is ___**2**_____. The explanatory variables in the regression (**are** / are not) jointly significant. Circle the correct alternative.

3

(2) c) The R^2 in this model is _**12/52 = 0.23**__. This means that _______ **23% of the variation in the dependent variable is explained by the two proposed explanatory (independent) variables.**_________

d) In the below space draw two sampling distributions one for the OLS estimator $\hat{\beta}_1$ and one for Extreme Values Estimator, say, $b_1^{(E)}$ so as to illustrate the Gauss-Markov Theorem.

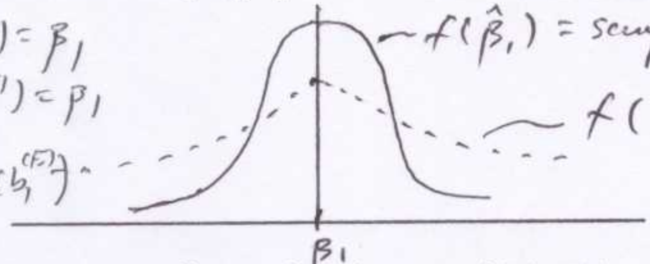
(4)

$E(\hat{\beta}_1) = \beta_1$

$E(b_1^{(E)}) = \beta_1$

but

$Var(\hat{\beta}_1) < Var(b_1^{(E)})$



4. Consider the computer output that you have been provided. This output analyses the relationship between the exam scores (variable name "midterm") that my Fall 2006 Eco 5350 students had on their first mid-term exam and the number of times they missed (variable name "miss") either a quick quiz or handing in an assigned exercise up to the time of the mid-term exam. Use the SAS program file and listing file to answer the following questions.

(5) a) This program has ____**3**_____ **data steps** in it. The **"end-of-file marker"** for reading in the data is denoted by the character ___**; (i.e. semicolon)**____. This program has ____**2**_____ **proc steps** in it. The **first delimiter** of a comment section is __/*____ while the **ending delimiter** of a comment section is __*/____.

b) Use you computer output to write out the estimated regression equation in "conventional" form in the below space.

ANSWER:

(4)

$$\text{Midterm} = \underset{(2.74)}{83.948} - \underset{(0.97)}{3.402}\text{miss} + \hat{u}$$

Root MSE = 10.455 $\quad R^2 = 0.378$

c) Suppose that we are interested in testing the null hypothesis $H_0 : \beta_1 = 0$. In a sentence or two explain to me the meaning of this hypothesis in the below space.

(2) **ANSWER: The conditional mean of midterm scores given the number of misses in the class is independent of the number of misses in the class. In other words, the slope of the Population Regression Function is not affected by miss.**

d) If you were to specify an alternative hypothesis to the above null hypothesis what would it be? Explain your reasoning below. H_1 : _$\beta_1 < 0$______.

2

ANSWER: One would expect that the expected mid-term score is going to be negatively related to the number of misses the student has.

e) Assuming a 5% level of significance, test your null and alternative hypotheses. What is your conclusion? Explain your reasoning.

4

ANSWER: The t-statistic for the explanatory variable miss is t = -3.40/0.974 = -3.49. This t-statistic has a two-sided p-value of 0.0023. Its left-tail p-value is p = 0.0023/2 = 0.00115. Since this p-value is less than 0.05 we reject the null hypothesis of independence between midterm scores and misses and accept the alternative hypothesis that midterm scores are, on average, negatively affected by the number of misses the student has in class.

f) In looking at a t-table we know that $t_{20,0.025} = 2.086$. Use this information to form a 95% confidence interval for β_1 in this regression problem. Show your work.

ANSWER:

4

$$\mathbf{-3.40205 \pm 0.97462*2.086 = -3.40205 \pm 2.03505}$$

[-5.435,-1.369]

g) In a sentence or two briefly explain to me the meaning of the confidence interval that you constructed in part f) above.

2

ANSWER: In many repeated samples 95% of so-constructed confidence intervals will encompass the true unknown β_1 coefficient. The β_1 coefficient in this problem is the slope of the PRF (conditional mean) in the direction of the variable "miss."

4

h) Given the sample regression function you wrote out in part b) above, how many points would you expect a student to (**lose** / gain) in going from 1 miss to two misses? __**3.4**___. In going from 2 misses to three misses? _**3.4**____. Suppose that a person misses 3 QQs and/or exercises during the semester. What is his/her expected score? __**73.74**____. Show your work below.

$$\widehat{midterm} = 83.94882 - 3.40205(3) = 83.94882 - 10.20615 = 73.74$$

i) In your output, I have provided two residual plots, one plotting the OLS residuals versus misses and the other plotting the squared residuals versus misses. What are the purposes of residual plots?

2

ANSWER: To determine if there is any heteroskedasticity in the errors of our regression model. If so, we need to seek a remedy for it before we can conduct any hypothesis tests.

j) One test of heteroskedasticity is called **White's test for heteroskedasticity**. (Even though we haven't discussed this test formally in class, you should be able to answer this question by reading my description below.) White's test is conducted by regressing the squared OLS residuals (denoted r2 in your output) on misses and squared misses (denoted miss2). The null hypothesis of homoskedasticity (i.e. no heteroskedasticity) is supported **if the population regression function of the squared residuals is flat, i.e. is equal to a constant function.** On the other hand, if "misses" and "squared misses" are important explanators of the variation in the squared residuals, then we have heteroskedasticity. Given the second regression output, does there appear to be a significant amount of heteroskedasticity in the errors of the midterm/miss regression? Explain your answer.

4

ANSWER: No. The overall F-statistic reported in the ANOVA table of this test regression is 2.32 with a probability value of 0.1259. At both the 5% and 10% levels of significance, we accept the null hypothesis that miss and miss2 do not significantly affect the level of the squared residuals. We accept the null hypothesis of homoskedasticity in the errors of our model and our statistical hypothesis testing can proceed vis-à-vis OLS.

```
/*  This data relates to the mid-term scores ("midterm") recorded for
    students in Tom Fomby's Introduction to Econometrics class (ECO 5350)
    offered during the Fall 2006 term at SMU.  The students mid-term
    scores were regressed on the variable "miss" that represents the
    total number of "quick quizzes" missed and/or assigned exercises not turned
    in for the class from the beginning of the semester to the time of the mid-term
    exam.  */
Options nodate;

data score;
   input midterm miss;
datalines;
92.5 0
86.8 0
81.1 0
75.5 0
86.8 0
92.5 0
83.0 2
96.2 0
94.3 0
83.0 0
75.1 1
75.5 2
62.3 3
81.3 3
77.4 0
52.8 4
47.2 3
73.6 8
67.9 3
66.0 7
84.9 0
88.7 0
;

proc reg data = score;
  model midterm = miss;
  output out=result r = residual;
run;

data score;
  merge score result;
  run;

data score;
  set score;
   r2 = residual**2;
   miss2 = miss**2;
   run;

proc reg data = score;
   model r2 = miss miss2;
   run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: midterm

Number of Observations Read	22
Number of Observations Used	22

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1332.05867	1332.05867	12.18	0.0023
Error	20	2186.45406	109.32270		
Corrected Total	21	3518.51273			

Root MSE	10.45575	R-Square	0.3786
Dependent Mean	78.38182	Adj R-Sq	0.3475
Coeff Var	13.33951		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	83.94882	2.74093	30.63	<.0001
miss	1	-3.40205	0.97462	-3.49	0.0023

The REG Procedure
Model: MODEL1
Dependent Variable: r2

Number of Observations Read	22
Number of Observations Used	22

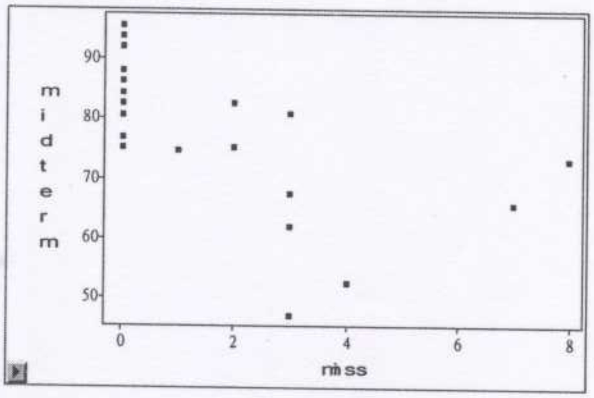
Analysis of Variance

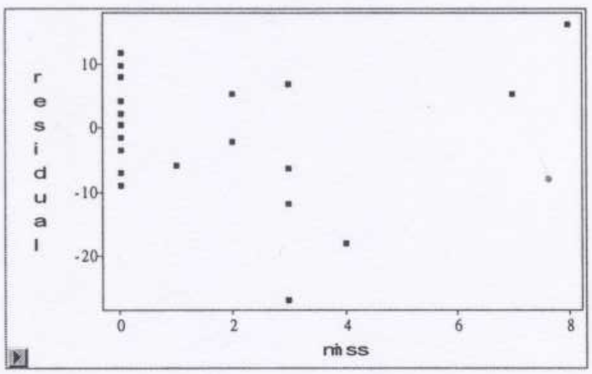
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	104233	52117	2.32	0.1259
Error	19	427581	22504		
Corrected Total	21	531814			

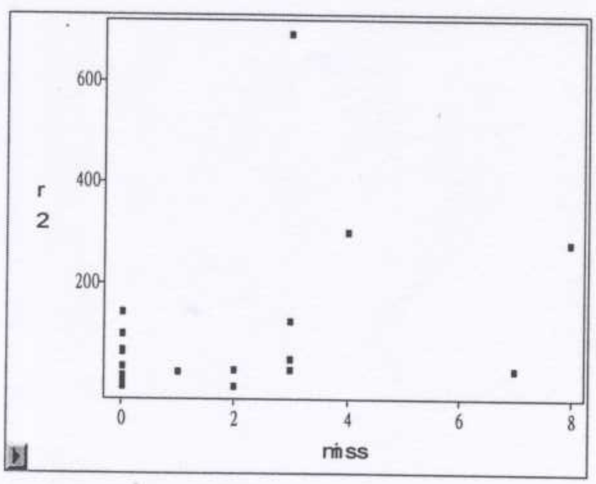
Root MSE	150.01416	R-Square	0.1960
Dependent Mean	99.38428	Adj R-Sq	0.1114
Coeff Var	150.94356		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	39.63301	42.48718	0.93	0.3626
miss	1	63.87055	39.33068	1.62	0.1209
miss2	1	-5.65984	5.57224	-1.02	0.3225


midterm
90
80
70
60
50
0
2
4
6
8
miss


residual
10
0
-10
-20
0
2
4
6
8
miss


r2
600
400
200
0
2
4
6
8
miss

FORMULA SHEET FOR
MID-TERM

BASIC STATISTICS:

1. Var(X) = $E(X-\mu_x)^2$

2. Cov(X,Y) = $E(X-\mu_x)(Y-\mu_y)$; $Corr(X,Y) = Cov(X,Y)/(Var(X)\cdot Var(Y))^{1/2}$

3. E(aX + bY) = aE(X) + bE(Y)

4. Var(aX + bY) = $a^2Var(X)+b^2Var(Y)+2Cov(X,Y)$

5. Sample Mean: $\overline{Y} = \sum_{1}^{N} Y_i$

6. Sample Variance: $s^2 = \sum_{1}^{N}(Y_i - \overline{Y})^2/(N-1)$

7. t-statistic for testing population mean:

$$t_{N-1} = \frac{\overline{Y}-\mu_{Y,0}}{se(\overline{Y})};\ \text{where } se(\overline{Y}) = s/\sqrt{N}$$

8. $(1-\alpha)$% confidence interval for μ

$$\Pr(\overline{Y} - t_{N-1,\alpha/2}\cdot se(\overline{Y}) < \mu < \overline{Y} + t_{N-1,\alpha/2}\cdot se(\overline{Y})) = 1-\alpha$$

9. Approximate t-statistic for testing difference in means (variances assumed unequal):

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \to Z = N(0,1)$$

10. Exact t-statistic for testing difference in means (variances assumed equal):

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \to t_{n_1+n_2-2}$$

where $s_p^2 = \dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

11. F-Test for equal variances across two populations

$$F_{v_1,v_2} = \frac{s_1^2}{s_2^2}$$

where $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$. Also, $F_{1-\alpha/2}(v_1, v_2) = \dfrac{1}{F_{\alpha/2}(v_1, v_2)}$

SOME OLS REGRESSION FORMULAS:

12. $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}$; $\quad Var(\hat{\beta}_0) = \dfrac{\sigma^2 \sum_1^N X_i}{N\sum_1^N (X_i - \overline{X})^2}$

13. $\hat{\beta}_1 = \dfrac{\sum_1^N (X_i - \overline{X})Y_i}{\sum_1^2 (X_i - \overline{X})^2} = \sum_1^N w_i Y_i$; $\quad Var(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_1^N (X_i - \overline{X})^2}$

14. $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

15. TSS = ESS + SSR ; $\sum_1^N (Y_i - \overline{Y})^2 = \sum_1^N (\hat{Y}_i - \overline{Y})^2 + \sum_1^N (Y_i - \hat{Y}_i)^2$; $R^2 = \dfrac{ESS}{TSS}$

16. $t = \dfrac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)}$

17. one-tailed p-value: $\Pr(t_0 < t)$ or $\Pr(t < t_0)$

18. two-tailed p-value: $\Pr(|t_0| < t)$

19. $\Pr(\hat{\beta}_i - t_{N-K,\alpha/2} \cdot se(\hat{\beta}_i) < \beta_i < \hat{\beta} + t_{N-K,\alpha/2} \cdot se(\hat{\beta}_i)) = 1 - \alpha$

20. $F_{overall} = \dfrac{R^2/(K-1)}{(1-R^2)/(N-K)}$

21. $F = \dfrac{(RSS_R - RSS_U)/J}{RSS_U/(N-K)} = \dfrac{(R_U^2 - R_R^2)/J}{(1-R_U^2)/(N-K)}$