Name Mr. Key
ID 7777777

ECO 5350
Intro. Econometrics

Prof. T. Fomby
Summer I, 2010

Mid-Term Exam

Instructions: Put your name and student ID in the upper right-hand-corner of this exam. This exam is worth a total of 6~~4~~ 66 points. The breakout of these points by question is as follows:

Q1 = 10, 2, 2, 2, 4 = 20 points
Q2 = 2, 2, 4, 4, 2 = ~~12 points~~ 14
Q3 = 3 points
Q4 = 4 points
Q5 = 4 points
Q6 = 4 points
Q7 = 2 points
Q8 = 4 points
Q9 = 2 points
Q10 = 5 points
Q11 = 4 points

You have one hour and thirty minutes to take this test. We will have lecture in the remaining 1 and one-half hours of the class remaining for the day. 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.

1. Consider the following STATA output concerning the Fair.dta program and its analysis of the vote on Presidents from 1880 – 2000.

```
. regress vote growth

      Source |       SS       df       MS              Number of obs =      31
-------------+------------------------------           F(  1,    29) =  16.369
       Model |  411.88         1   411.88              Prob > F      =  0.0004
    Residual |  729.669044    29   25.1610015          R-squared     =  0.3608
-------------+------------------------------           Adj R-squared =  0.3388
       Total |  1141.54952    30   38.0516506          Root MSE      =  5.0161

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      growth |   .6599232   .1631067    4.045   0.000      0.326        0.993
       _cons |   51.93868   .9054453    57.36   0.000    50.08683     53.79052
------------------------------------------------------------------------------
```

a) Fill in the above blanks. Your calculations don't have to be as accurate as the computations produced by STATA but at least close. Show me in detail how you calculate the SS(Model), the F-statistic, the t-ratio, and the 95% confidence intervals.

SS(Model): 1141.54952 = Total SS = Model SS + Resid SS

$\therefore$ 1141.54952 − 729.669 = 411.88

F-statistic:

$$F = \frac{MS(Model)}{MS(Residual)} = \frac{411.68}{25.161} = 16.3697$$

10

R-squared:

$$R^2 = \frac{Model\ SS}{Total\ SS} = \frac{411.88}{1141.549} = 0.36080$$

t-ratio:

coeff./se = 0.6599232 / 0.1631067 = 4.04596

95% confidence interval:

$0.6599232 \pm t_c\, se(b_2)$ $\quad t_c = t_{0.975,29} = 2.045$

$0.6599232 \pm 2.045(0.1631067) \Rightarrow [0.326, 0.993]$

b) Suppose you are interested in testing the significance of the growth variable in the above regression and that you suspect that the variable has a direct effect on the outcome of presidential elections. State the null hypothesis of your test and the alternative hypothesis of your test.

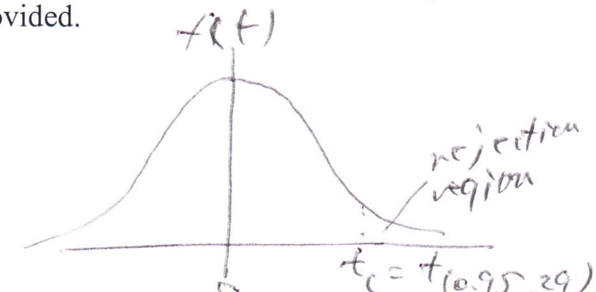
2

H_0: $\beta_2 = 0$ (i.e. growth has no effect on vote outcome)

H_1: $\beta_2 > 0$ (i.e. growth positively affects vote outcome)

c) For the above test, tell me the acceptance and rejection regions for the test and draw them below. Assume a 5% level of your test. See the t-table that you have been provided.

2



Rejection region $= [1.699, \infty)$

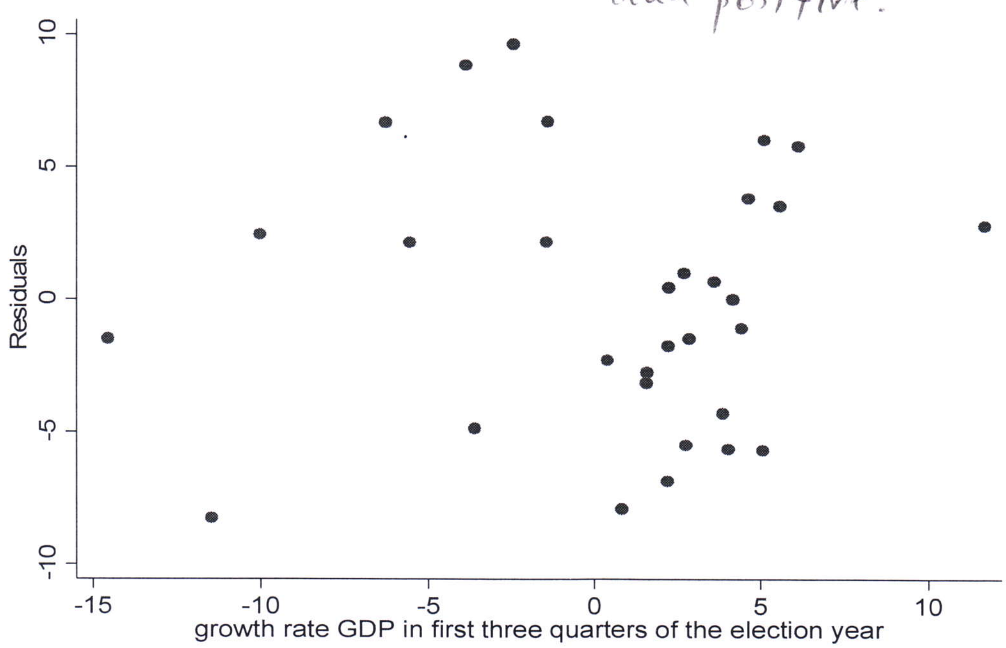
$t_c = t_{(0.95, 29)} = 1.699$

d) State the conclusion of the test you conducted in part c).

2

Since the observed t-statistic (4.04) is greater than 1.699, we reject H_0 and accept H_1 that the effect of growth is statistically significant and positive.

e) Consider the following residual plot.



What is the purpose of this plot? What does it imply with respect to the hypothesis testing that you conducted above?

4

The above residual plot helps us determine if we have heteroscedasticity in the errors of our regression. From the above plot it appears that we don't so we can proceed to do our hypothesis testing in the context of OLS estimation.

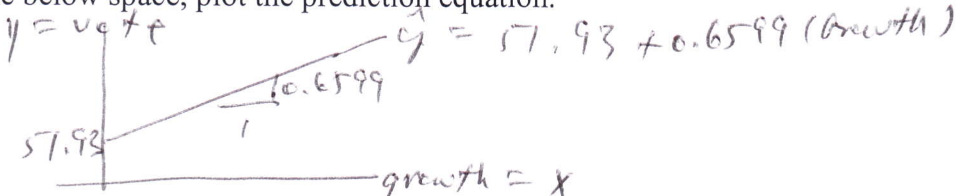
2. Consider the regression that is reported in Question 1 above.

a) Write out the prediction equation that one would use in predicting the 2004 election of candidate Bush.

(2) Prediction equation: $\hat{y}$ = $51.93868 + 0.6599232\,(\text{Growth})$.

b) In the below space, plot the prediction equation.

(2)



c) Suppose that in the first three quarters leading up to the 2004 election that the growth rate in the economy is shown to be -1.0%. What would be your predicted outcome of the race in percentage vote for the incumbent Bush? Show your work below.

(4) $\hat{y}_0 = 51.93868 + 0.6599232(-1.0) = 51.2787$

d) Given the information that you have in the STATA output in Question 1, compute a 95% confidence interval for your prediction in part c). Show your work below. You will need the following information to help you in the computation of your prediction confidence interval.

$Var(f) = \hat{\sigma}^2\left[1 + \frac{1}{N} + (x_0 - \bar{x})^2 Var(b_2)\right]$

$se(f) = \sqrt{Var(f)}$

$= 25.16100151\left[1 + \frac{1}{31} + (-1.0 - .5547097)^2 \cdot (0.1631067)^2\right]$

. summarize growth

Variable	Obs	Mean	Std. Dev.	Min	Max
growth	31	.5547097	5.614765	-14.557	11.677

$\Pr(\hat{y}_0 - se(f)\,t_c < y_0 < \hat{y}_0 + se(f)\cdot t_c) = 1 - \alpha$

$= 25.99$

(4) $\hat{y}_0 \pm se(f)\,t_c \Rightarrow 51.2787 \pm 5.098(2.045)$

$\therefore se(f) = \sqrt{25.99} = 5.098$

$[40.85, 61.70]$

e) Given the information that you have in the STATA output in Question 1, what level of growth would you have to have in order to predict a victory for the Democrats (Kerry) in 2004? Solve the following equation:

(2) $50.00 > 51.93868 + 0.6599232$ Growth

$\therefore$ Growth $< -2.95\%$ will provide Democratic victory

3. Match the following terms with the definitions:

time-series data = Definition A

(3) *cross-section data* = Definition C

panel data = Definition B

Definition A: data collected over discrete intervals of time—for example, the annual price of wheat in the US from 1880 to 2007, or the daily price of General Electric stock from 1980 to 2007.

Definition B: data that follow individual micro-units over time. For example, the U.S. Department of Education has several on-going surveys, in which the same students are tracked over time, from the time they are in the 8th grade until their mid-twenties.

Definition C: data collected over sample units in a particular time period—for example, income by counties in California during 2006, or high school graduation rates by state in 2006.

4. Consider the following regression equation: $\hat{Y} = 10 + 4X + 2(X^*D) + 3D$

Let D be 1 if the cross-sectional data is from a southern state and 0 if the data is from a northern state. Then the regression equation for northern states is

$\hat{Y}$ = $10 + 4X$.

4

The regression equation for the southern states is

$\hat{Y}$ = $10 + 3 + (4+2)X$. $= 13 + 6X$

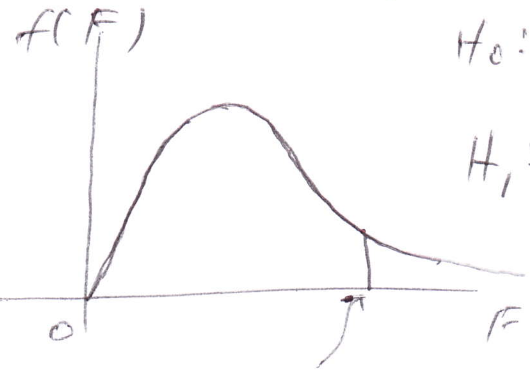
The Y-intercept for the northern states equation is 10 while the slope of the northern states equation is 4.

5. If, in the above Question 4, we wanted to test the significant difference between the north and south regressions we would apply the so-called ___________ test. Suppose that in the data we know that RSSU = 46, RSSR = 55, and N = 1000. Write out below the F-statistic that you would use to test the significant difference between the north and south regressions.

4

$$F = \frac{(RSSR - RSSU)/J}{RSSU/(N-K)} = \frac{(55-46)/2}{46/(1000-4)} = 97.43$$

6. Now given the F-statistic you have calculated in Question 5, use the F-table that you have been provided to form a critical region for your test of the north and south difference. Draw your critical region below. State the null and alternative hypotheses of your test and tell me the conclusion you draw from the F-statistic.



4

H_0: North and South regressions are same

H_1: North and South regressions are different

Rejection region = [3.00, ∞)

$F_{2,\infty,.05} = 3.00$

Since $F_c = 97.43 > 3.00$ we reject H_0 and accept H_1, North and South Regressions are different.

7. The **overall F-statistic** tests

(2)

a. The significance of the intercept term
b. The joint significance of the explanatory variables
c. The significance of the error term
d. The presence of heteroscedasticity

8. The estimators b_1 and b_2 of the intercept, β_1, and slope, β_2, respectively, in the conditional mean function $E(Y|X) = \beta_1 + \beta_2 X$ are both linear in the observations $Y_1, Y_2, \cdots, Y_N$ and are unbiased in that $E(b_1) = \beta_1$ and $E(b_2) = \beta_2$. Also these estimators are **BLU** estimators. This means that the least squares estimators b_1 and b_2 have smaller sampling variances than any other unbiased linear estimators.

9. The above theorem is called the Gauss-Markov theorem.

10. Consider the following multiple linear regression fit on 500 observations where y is the dependent variable and x1, x2, and x3 are explanatory variables. Using a backward selection algorithm which variable would you drop first. Under the coefficient estimates you will find in the parentheses the standard errors of the estimates, in the square brackets you should fill in the t-statistics. In the p= space below I have put the two-sided p-values associated with the t-statistics. So your job is to fill in the t-statistics and below indicate the first variable you would drop using the backward selection algorithm.

y = 12.0 +	3.0x1 −	2.0x2 +	1.50x3 +	e
(3.0)	(1.5)	(2.0)	(1.0)	
[4.0]	[2.0]	[1.0]	[1.5]	
p=0.00	p=0.04	p=0.32	p=0.14	

First variable to drop is (x1 / x2 / x3).

11. Suppose that you start out with 9 explanatory variables x1, x2, …, x9 and wind up with the following regression having three variables using backward selection with a chosen level of significance of 0.05. In the below regression the **conditional** p-values of the coefficient t-statistics are reported in the parentheses.

y = 8.0 +	2.0x4 −	4.0x5 +	1.20x7 +	e
	(0.04)	(0.01)	(0.12)	

After adjusting the above regression for the backward selection procedure the significant variable(s) at the **unconditional** level of significance of 0.05 is (are) X5.
We adjusted the above conditional p-values by the factor of 3.

9/3