ECON 581 Startz/ Tsang
Econometrics II Winter 2008

Homework 3
Due Feb 26^{th}, 2007

QUESTION 1

1) Finish exercises 1, 5, 6, 8, and 9 from Chapter 8 of Greene (6^{th} edition).

2) Finish exercises 1, 2, and 3 from Chapter 19 of Greene (6^{th} edition).

QUESTION 2

a) Given a standard normal PDF $\phi(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2\right)$, find for the derivatives $\phi'(z)$, $\phi''(z)$, $\phi'''(z)$ and $\phi''''(z)$ all as functions of $\phi(z)$ and z.

b) Use the results in a) to find the first, second, third and fourth moments for z, where the k -th moment is defined by $\int_{-\infty}^{\infty} z^k f(z)dz$.

c) Let $z_1, z_2, ..., z_m$ be mutually independent standard normal random variables, then $y \equiv \sum_{i=1}^{m} z_i^2$ follows the chi-squared distribution with degrees of freedom m, or $y \sim \chi^2(m)$.

Show that $E(y)=m$ and $Var(y)=2m$ (You need the results in b)).

QUESTION 3

Matlab Exercise

Consider the following DGP:

$y_i = 1 + 2x_i + u_i$, with $u_i \sim N(0,1)$ and $x_i \sim U(0,1)$ (standard uniform)

Now do an experiment with 10,000 runs, with a sample size of 1000. In each iteration, run the regression $y_i = \alpha + \beta x_i + u_i$, and another regression $y_i = \alpha + \beta x_i^* + u_i$, where $x_i^* = x_i + v_i$, with $v_i \sim N(0, 0.0025)$ being a rather tiny random noise. For each regression, perform the *t*-test for the null hypothesis that $\beta = 2$ at 5% level (i.e. the critical value is 1.96).

For each regression, out of what percentage of the 10,000 runs is the null hypothesis wrongly rejected? Is it close to 5%? Why and why not?

QUESTION 4

EViews Exercises

Back to our lovely extramarital affairs sample (See **Fair, Ray C**, 1978. "A Theory of Extramarital Affairs," *Journal of Political Economy*, University of Chicago Press, vol. 86(1), pages 45-61, February. I get the data from Wooldridge's website). Download the workfile **affairs.wf1** from my website.

1) Run an OLS regression for naffairs (i.e. number of affairs) on a constant, educ, age, kids (i.e. whether the respondent has kids), male, yrsmarr (number of years married), vryhap and vryunhap.

2) Perform a White test for heteroskedasticity (with cross terms). Do you reject the null hypothesis of homoskedasticity at 5% level?

3) Re-run the regression in 1), but use the White heteroskedasticity consistent standard error this time. Have the results in 1) changed much?

QUESTION 5

EViews Exercises

I introduce a dataset that we will play around quite a bit: **Monthly Phillips Curve Data 1958:01 – 2006:12** (phillips.xls): It includes the month-to-month annual % change in the consumer price index (less food and energy) and the unemployment rate.

What is the Natural Rate of Unemployment? The expectation-augmented Phillips curve is specified as: $\pi_t - \pi_{t-1} = \beta_1 \left(un_t - un^* \right) + e_t$, where un^* is the natural rate of unemployment, the rate at which inflation stops changing. The lagged inflation is a proxy for expected inflation. The Phillips curve can be estimated by OLS as

[1] $\pi_t - \pi_{t-1} = \beta_0 + \beta_1 un_t + e_t$ with $\beta_0 = -\beta_1 un^*$

i) Estimate [1] with OLS. What is the estimated natural rate of unemployment? Use the delta method to find the 95% confidence interval for the natural rate.

ii) Perform the Breusch-Godfrey test for AR(1) serial correlation in the error term (EViews can do it for you). Do you reject the null of no serial correlation at 5%?

iii) Re-estimate [1] with AR(1) correction (you can do it in EViews by adding "AR(1)" at the end of your regression command). What is the estimated natural rate? Use the delta method to find the 95% confidence interval for the natural rate.

QUESTION 6

Matlab Exercises

How Serial Correlation Can Screw Up Your Regression

a) We want to know what will happen when the error term in a regression is serially correlated. You are given the DGP:

$u_t = e_t + 1.5e_{t-1}$, $e_t \sim N(0,1)$

$Y_t = 1 + 2X_t + u_t$

$X_t = 0.6X_{t-1} + v_t$, $v_t \sim N(0,1)$ and $X_1 = 0$

Using a sample size of 500, obtain the OLS estimates from regressing Y_t on a constant and X_t, and perform a *t*-test for the coefficient on X_t under the null that it equals the true value, at 5% level. Repeat the experiment 5000 times. Is the rejection rate close to 5%? Why and why not?

b) Repeat the above experiment with $u_t = 0.7u_{t-1} + e_t$, $e_t \sim N(0,1)$ and $u_1 = 0$.

QUESTION 7

Matlab Exercises

The following two exercises ask you to plot the power functions of various tests.

A. Plotting the Power Function of a *t*-test under Heteroskedasticity

Consider the following DGP:

$y_t = 1 + kx_t + u_t$, $x_t \sim N(0,1)$, $u_t \sim N(0, x_t^4)$

Use different true values k : -1, -0.9, …, 0, …, 0.9, 1. **For each true value**, do the following: Generate 100 observations from the DGP, run the regression $y_t = \alpha + \beta x_t + u_t$, and do a *t*-test for the null hypothesis that $\beta = 0$ at 5% level, separately for two different standard errors: i) the usual $s^2 (\mathbf{X}'\mathbf{X})^{-1}$ and ii) the White estimator

$\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{t=1}^{T}\hat{u}_t^2\mathbf{x}_t\mathbf{x}_t'\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ which corrects for heteroskedasticity. Repeat it 5000 times.

Record the % that the null hypothesis is rejected.

When the above experiment is done for all true values, plot the rejection % on the y-axis, and the true values on the x-axis, in the same graph for i) and ii).

Notice that unlike our previous exercises, the null is wrong except only when $k=0$. The above plot gives you the power function of the tests. A good test should have 5% rejection rate when $k=0$, and high rejection rate when $k \neq 0$. Comparing the plots for i) and ii), which is a more reliable *t*-test?

Repeat the above exercise with a sample of 1,000. Any change to your conclusion?

B. Plotting the Power Functions of Two Tests for Heteroskedasticity

In this exercise we compare the small-sample performance of the White test and the Breusch-Pagan test for heteroskedasticity. Consider the following DGP:

$$y_t = 1 + 2x_t + u_t,\ x_t \sim N(0,1),\ u_t \sim N\left(0, \left((kx_t)^2 + 1\right)^2\right)$$

Use different true values k: -1, -0.9, …, 0, …, 0.9, 1. **For each true value**, do the following: Generate 100 observations from the DGP, run the regression $y_t = \alpha + \beta x_t + u_t$, and do two different tests for heteroskedasticity: i) the White test and ii) the Breusch-Pagan test. Repeat it 5000 times. Record the % that the null hypothesis is rejected.

When the above experiment is done for all true values, plot the rejection % on the y-axis, and the true values on the x-axis, in the same graph for i) and ii).

Again, notice that unlike our previous exercises, the null is wrong except only when $k=0$. The above plot gives you the power function of the tests. A good test should have 5% rejection rate when $k=0$, and high rejection rate when $k \neq 0$. Comparing the plots for i) and ii), which is a more reliable test for heteroskedasticity?

Repeat the above exercise for $u_t \sim \left((kx_t)^2 + 1\right) * U(0,1) - 0.5$, i.e. the error is not normal.

Do the tests get worse?

<u>QUESTION 8</u> – Borjas's 1980 JLE Paper

a) Prove result (4).

b) Explain footnote 7 intuitively.

c) Explain footnote 10 intuitively. Why is selection bias a bad thing in this case?

d) Prove result (8) (Trivial).

e) Prove result (9) (Take a look at footnote 20 first).

f) Explain what Table 3 is trying to say.

<u>QUESTION 9</u> – Rose's 1987 JPE Paper

a) What are hypotheses to be tested in this paper for union workers in the trucking industry, nonunion workers in the trucking industry and drivers outside the regulated trucking industry?

b) What is Figure 1 trying to show? Why is the evidence from Figure 1 not conclusive?

c) On page 1158, though it will be covered later this quarter, explain intuitively what the author means by "occupational fixed effect" and "industry fixed effect".

d) On page 1159, why is not having information on the firms employing respondents a problem?

e) On page 1159, what does the author mean by "union status should be modeled as an endogenous variable"? Why will it overstate the level of the union premium?

f) On page 1160, why does the author claim that "they are likely to bias the results against finding rent sharing"?

g) What is a "representative driver"?

h) Why is the small sample size for later years in Table 2 a problem?

i) Explain footnote 32 carefully.

j) Explain what Table 3 is trying to show.