Quantitative Methods for Economics

Tutorial 12

Katherine Eyal

TUTORIAL 12

25 October 2010

ECO3021S

Part A: Problems

1. State with *brief reason* whether the following statements are true, false or uncertain:

   (a) In the presence of heteroskedasticity OLS estimators are biased as well as inefficient.

   (b) If heteroskedasticity is present, the conventional t and F tests are invalid.

   (c) If a regression model is mis-specified (e.g., an important variable is omitted), the OLS residuals will show a distinct pattern.

   (d) If a regressor that has nonconstant variance is (incorrectly) omitted from a model, the (OLS) residuals will be heteroskedastic.

2. In a regression of average wages, (W, in Rands) on the number of employees (N) for a random sample of 30 firms, the following regression results were obtained (t-statistics in parentheses):

$$\widehat{W} = \underset{(N/A)}{7.5} + \underset{(16.10)}{0.009}\ N \qquad R^2 = 0.90$$

$$\widehat{W}/N = \underset{(14.43)}{0.008} + \underset{(76.58)}{7.8}\ (1/N) \qquad R^2 = 0.99$$

   (a) How do you interpret the two regressions?

   (b) What is the researcher assuming in going from the first to the second equation? Was he worried about heteroskedasticity? How do you know?

   (c) Can you relate the slopes and intercepts of the two models?

   (d) Can you compare the R^2 values of the two models? Why or why not?

3. In 1914, the South African Robert Lehfeldt published what has become a well-known estimate of a price elasticity of demand that relied on a double logarithmic specification. Lehfeldt reasoned that variations in weather drive fluctuations in wheat production from one year to the next, and that the price of wheat adjusts so that buyers are willing to buy all the wheat produced. Hence, he argued, the price of wheat observed in any year reflects the demand for wheat. Consider the equation

$$\log(price) = \beta_0 + \beta_1 \log(wheat) + u$$

where *price* denotes the price of wheat, *wheat* denotes the quantity of wheat, and $1/\beta_1$ is the price elasticity of demand for wheat. Demonstrate that if the first four Gauss-Markov assumptions apply, the inverse of the OLS estimator of the slope in the above equation is a consistent estimator of the price elasticity of demand for wheat.

4. Suppose that the model

$$pctstck = \beta_0 + \beta_1 funds + \beta_2 risktol + u$$

satisfies the first four Gauss-Markov assumptions, where *pctstck* is the percentage of a worker's pension invested in the stock market, $funds$ is the number of mutual funds that the worker can choose from, and $risktol$ is some measure of risk tolerance (larger $risktol$ means the person has a higher tolerance for risk). If $funds$ and $risktol$ are positively correlated, what is the inconsistency in $\widetilde{\beta}_1$, the slope coefficient in the simple regression of *pctstck* on $funds$?

Part B: Computer Exercises

1. Does the separation of corporate control from corporate ownership lead to worse firm performance? George Stigler and Claire Friedland have addressed this question empirically using a sample of U.S. firms. A subset of their data are in the file EXECCOMP.DTA. The variables in the file are as follows:

ecomp	Average total annual compensation in thousands of dollars for a firm's top three executives
assets	Firm's assets in millions of dollars
profits	Firm's annual profits in millions of dollars
mcontrol	Dummy variable indicating management control of the firm

Consider the following equation:

$$profits = \beta_0 + \beta_1 assets + \beta_2 mcontrol + u \tag{1}$$

(a) Estimate the equation in (1). Construct a 95% confidence interval for β_2.

(b) Estimate the equation in (1) but now compute the heteroskedasticity-robust standard errors. (Use the command: `reg profits assets mcontrol, robust`) Construct a 95% confidence interval for β_2 and compare it with the nonrobust confidence interval from part (a).

(c) Compute the Breusch-Pagan test for heteroskedasticity. Form both the F and LM statistics and report the p values. (You can use the command `hettest` to perform the LM version of the Breusch-Pagan test in Stata.)

(d) Compute the special case of the White test for heteroskedasticity. Form both the F and LM statistics and report the p values. How does this compare to your results from (c)? (You can use the command `imtest, white` to perform the LM version of the original White test in Stata. Note that this test statistic will be different from the one you compute for the special case of the White test.)

(e) Divide all of the variables (including the intercept term) in equation (1) by the square root of *assets*. Re-estimate the parameters of (1) using these data. What are these estimates known as? These new estimates are BLUE if what is true about the disturbances in (1)?

(f) Conduct White's test for the disturbances in (e) being homoskedastic. What do you conclude at the 5% level?

(g) Construct a 95% confidence interval for β_2 using the estimates from the regression in (e). Are you sympathetic to the claim that managerial control has no large effect on corporate profits?

2. Let $arr86$ be a binary variable equal to unity if a man was arrested during 1986, and zero otherwise. The population is a group of young men in California born in 1960 or 1961 who have at least one arrest prior to 1986. A linear probability model for describing $arr86$ is

$$arr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u,$$

where

$pcnv$	$=$	the proportion of prior arrests that led to a conviction
$avgsen$	$=$	the average sentence served from prior convictions (in months)
$tottime$	$=$	months spent in prison since age 18 prior to 1986
$ptime86$	$=$	months spent in prison in 1986
$qemp86$	$=$	the number of quarters (0 to 4) that the man was legally employed in 1986

The data set is in CRIME1.DTA.

(a) Estimate this model by OLS and verify that all fitted values are strictly between zero and one. What are the smallest and largest fitted values?

(b) Estimate the equation by weighted least squares, as discussed in Section 8.5 of Wooldridge.

(c) Use the WLS estimates to determine whether $avgsen$ and $tottime$ are jointly significant at the 5% level.

3. Use the data in WAGE1.DTA for this question.

(a) Estimate the equation

$$wage = \beta_0 + \beta_1\ educ + \beta_2\ exper + \beta_3\ tenure + u.$$

Save the residuals and plot a histogram.

(b) Repeat part (a) but with $\log(wage)$ as the dependent variable.

(c) Would you say that Assumption MLR.6 in Wooldridge (Normality) is closer to being satisfied for the level-level model or the log-level model?

4. Use the data in CHARITY.DTA for this question.

(a) Using all 4,268 observations, estimate the equation

$$gift = \beta_0 + \beta_1 mailsyear + \beta_2 giftlast + \beta_3 propresp + u$$

and interpret your results in full.

(b) Reestimate the equation in part (a), using the first 2,134 observations.

(c) Find the ratio of the standard errors on $\widehat{\beta}_2$ from parts (a) and (b). Compare this with the result from equation (5.10) on page 175 of Wooldridge.

TUTORIAL 12 SOLUTIONS

25 October 2010

ECO3021S

Part A: Problems

1. State with *brief reason* whether the following statements are true, false or uncertain:

   (a) In the presence of heteroskedasticity OLS estimators are biased as well as inefficient.

   (b) If heteroskedasticity is present, the conventional t and F tests are invalid.

   (c) If a regression model is mis-specified (e.g., an important variable is omitted), the OLS residuals will show a distinct pattern.

   (d) If a regressor that has nonconstant variance is (incorrectly) omitted from a model, the (OLS) residuals will be heteroskedastic.

   SOLUTION:

   (a) False. In the presence of heteroskedasticity OLS estimators are **unbiased**, but are inefficient.

   (b) True. If heteroskedasticity is present, the OLS standard errors are biased and are no longer valid for constructing confidence intervals and t statistics. The usual OLS t statistics do not have t/distributions, and F statistics are no longer F distributed in the presence of heteroskedasticity.

   (c) True. If there are specification errors, the residuals will exhibit noticeable patterns.

   (d) True. Often, what looks like heteroskedasticity may be due to the fact that some important variables are omitted from the model.

2. In a regression of average wages, (W, in Rands) on the number of employees (N) for a random sample of 30 firms, the following regression results were obtained (t-statistics in parentheses):

$$\widehat{W} = \underset{(N/A)}{7.5} + \underset{(16.10)}{0.009}\, N \qquad R^2 = 0.90$$

$$\widehat{W}/N = \underset{(14.43)}{0.008} + \underset{(76.58)}{7.8}\, (1/N) \qquad R^2 = 0.99$$

(a) How do you interpret the two regressions?

(b) What is the researcher assuming in going from the first to the second equation? Was he worried about heteroskedasticity? How do you know?

(c) Can you relate the slopes and intercepts of the two models?

(d) Can you compare the R^2 values of the two models? Why or why not?

SOLUTION:

(a) The first regression is an OLS estimation of the relationship between average wages and number of employees. The results suggest that if a firm hires ten new employees, the average wage increases by R0.09. The second regression is weighted least squares (WLS) estimation of the relationship between average wages and number of employees. The results suggest that if a firm hires ten new employees, the average wage increases by R0.08.

(b) The researcher is assuming that the error variance is proportional to N^2 : $\text{Var}(u_i \mid N) = \sigma^2 N_i^2$. He was worried about heteroskedasticity because if $\text{Var}(u_i \mid N) = \sigma^2 N_i^2$, the error variance is clearly non-constant. If this is a correct specification of the form of the variance, then WLS estimation (the second equation) is more efficient than OLS (the first equation).

(c) The intercept of the second equation is the slope coefficient of the first equation, and the slope coefficient of the second equation is the intercept of the first equation.

(d) You cannot compare the R^2 values because the dependent variables are different in the two models.

3. In 1914, the South African Robert Lehfeldt published what has become a well-known estimate of a price elasticity of demand that relied on a double logarithmic specification. Lehfeldt reasoned that variations in weather drive fluctuations in wheat production from one year to the next, and that the price of wheat adjusts so that buyers are willing to buy all the wheat produced. Hence, he argued, the price of wheat observed in any year reflects the demand for wheat. Consider the equation

$$\log(price) = \beta_0 + \beta_1 \log(wheat) + u$$

where *price* denotes the price of wheat, *wheat* denotes the quantity of wheat, and $1/\beta_1$ is the price elasticity of demand for wheat. Demonstrate that if the first four Gauss-Markov assumptions apply, the inverse of the OLS estimator of the slope in the above equation is a consistent estimator of the price elasticity of demand for wheat.

SOLUTION:

To show that $1/\widehat{\beta}_1$ is a consistent estimator of $1/\beta_1$, we have to show that $\widehat{\beta}_1$ is a consistent estimator of β_1. This is just the proof of Theorem 5.1 (see page 169 in Wooldridge).

4. Suppose that the model

$$pctstck = \beta_0 + \beta_1 funds + \beta_2 risktol + u$$

satisfies the first four Gauss-Markov assumptions, where *pctstck* is the percentage of a worker's pension invested in the stock market, *funds* is the number of mutual funds that the worker can choose from, and *risktol* is some measure of risk tolerance (larger *risktol* means the person has a higher tolerance for risk). If *funds* and *risktol* are positively correlated, what is the inconsistency in $\widetilde{\beta}_1$, the slope coefficient in the simple regression of *pctstck* on *funds*?

SOLUTION:

A higher tolerance of risk means more willingness to invest in the stock market, so $\beta_2 > 0$. By assumption, *funds* and *risktol* are positively correlated. Now we use equation (5.5) of Wooldridge, where $\delta_1 > 0$: $\text{plim}(\widetilde{\beta}_1) = \beta_1 + \beta_2\delta_1 > \beta_1$, so $\widetilde{\beta}_1$ has a positive inconsistency (asymptotic bias). This makes sense: if we omit *risktol* from the regression and it is positively correlated with *funds*, some of the estimated effect of *funds* is actually due to the effect of *risktol*.

Part B: Computer Exercises

1. Does the separation of corporate control from corporate ownership lead to worse firm performance? George Stigler and Claire Friedland have addressed this question empirically using a sample of U.S. firms. A subset of their data are in the file EXECCOMP.DTA. The variables in the file are as follows:

ecomp	Average total annual compensation in thousands of dollars for a firm's top three executives
assets	Firm's assets in millions of dollars
profits	Firm's annual profits in millions of dollars
mcontrol	Dummy variable indicating management control of the firm

Consider the following equation:

$$profits = \beta_0 + \beta_1 assets + \beta_2 mcontrol + u \quad (1)$$

(a) Estimate the equation in (1). Construct a 95% confidence interval for β_2.

(b) Estimate the equation in (1) but now compute the heteroskedasticity-robust standard errors. (Use the command: `reg profits assets mcontrol, robust`) Construct a 95% confidence interval for β_2 and compare it with the nonrobust confidence interval from part (a).

(c) Compute the Breusch-Pagan test for heteroskedasticity. Form both the F and LM statistics and report the p values. (You can use the command `hettest` to perform the LM version of the Breusch-Pagan test in Stata.)

(d) Compute the special case of the White test for heteroskedasticity. Form both the F and LM statistics and report the p values. How does this compare to your results from (c)? (You can use the command `imtest, white` to perform the LM version of the original White test in Stata. Note that this test statistic will be different from the one you compute for the special case of the White test.)

(e) Divide all of the variables (including the intercept term) in equation (1) by the square root of *assets*. Re-estimate the parameters of (1) using these data. What are these estimates known as? These new estimates are BLUE if what is true about the disturbances in (1)?

(f) Conduct White's test for the disturbances in (e) being homoskedastic. What do you conclude at the 5% level?

(g) Construct a 95% confidence interval for β_2 using the estimates from the regression in (e). Are you sympathetic to the claim that managerial control has no large effect on corporate profits?

SOLUTION:

(a)

```
      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  2,    66) =   13.49
       Model |  6843.42169     2  3421.71085           Prob > F      =  0.0000
    Residual |  16735.2881    66  253.564972           R-squared     =  0.2902
-------------+------------------------------           Adj R-squared =  0.2687
       Total |  23578.7098    68  346.745733           Root MSE      =  15.924

------------------------------------------------------------------------------
     profits |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      assets |   .0274626   .0055847     4.92   0.000     .0163124    .0386128
    mcontrol |  -6.186284   3.844436    -1.61   0.112    -13.86195     1.48938
       _cons |   8.087969   3.053642     2.65   0.010     1.991175    14.18476
------------------------------------------------------------------------------
```

The 95% confidence interval for β_2 is given in the Stata output: $[-13.86195, 1.48938]$. (You should also make sure that you can construct this CI by hand.)

(b)

```
Regression with robust standard errors                 Number of obs =      69
                                                       F(  2,    66) =    2.92
                                                       Prob > F      =  0.0607
                                                       R-squared     =  0.2902
                                                       Root MSE      =  15.924

------------------------------------------------------------------------------
             |               Robust
     profits |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      assets |   .0274626   .0157674     1.74   0.086    -.0040181    .0589433
    mcontrol |  -6.186284   3.765382    -1.64   0.105    -13.70411    1.331542
       _cons |   8.087969   3.796616     2.13   0.037     .5077806    15.66816
------------------------------------------------------------------------------
```

The 95% confidence interval for β_2 is given in the Stata output: $[-13.70411, 1.331542]$.
(You should also make sure that you can construct this CI by hand.)

This CI is narrower than the nonrobust one from part (a).

(c) Stata commands:

reg profits assets mcontrol	(This is the usual OLS regression)
predict u, resid	(This saves the residuals from the regression in a variable named u)
gen u2=u^2	(This creates the squared residuals and saves them as $u2$)
reg u2 assets mcontrol	(This is regression (8.14) of Wooldridge)

Stata output:

```
      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  2,    66) =   10.77
       Model |  21041422.4     2  10520711.2           Prob > F      =  0.0001
    Residual |  64482732.8    66  977011.104           R-squared     =  0.2460
-------------+------------------------------           Adj R-squared =  0.2232
       Total |  85524155.3    68  1257708.17           Root MSE      =  988.44

------------------------------------------------------------------------------
          u2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      assets |   1.596643   .3466613     4.61   0.000     .9045113    2.288775
    mcontrol |  -120.8976    238.637    -0.51   0.614    -597.3517    355.5566
       _cons |  -143.9968   189.5498    -0.76   0.450    -522.4451    234.4515
------------------------------------------------------------------------------
```

Breusch-Pagan test for heteroskedasticity (BP test):

- F statistic:

$$\begin{aligned} F &= \frac{R^2_{\widehat{u}^2}/k}{\left(1 - R^2_{\widehat{u}^2}\right)/(n-k-1)} \\ &= 10.77 \end{aligned}$$

  This F statistic is easily obtained from the Stata output, as it is the F statistic for overall significance of a regression.
  The p value is 0.0001 indicating that we can reject the null hypothesis of homoskedasticity.

- LM statistic:

$$\begin{aligned} LM &= n \cdot R^2_{\widehat{u}^2} \\ &= 69\,(0.2460) \\ &= 16.974 \end{aligned}$$

  The p value is less than 0.01 (the 1% critical value from the χ^2 distribution with 2 degrees of freedom is 9.21), indicating that we can reject the null of homoskedasticity.

Performing the BP test in Stata:

```
reg profits assets mcontrol
hettest
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of profits

         chi2(1)      =   171.40
         Prob > chi2  =   0.0000
```

Stata performs a different version of the BP test to the one discussed in Wooldridge, but the conclusion is the same. We reject the null of homoskedasticity.

(d) Stata commands:

`reg profits assets mcontrol`	(This is the usual OLS regression)
`predict u, resid`	(This saves the residuals from the regression in a variable named u)
`gen u2=u^2`	(This creates the squared residuals and saves them as $u2$)
`predict profitshat`	(This calculates the predicted values of profits and saves them as $profitshat$)
`gen profitshat2=profitshat^2`	(This squares the predicted values of profits saves them as $profitshat2$)
`reg u2 profitshat profitshat2`	(This is regression (8.20) of Wooldridge)

(**Note** that you do not need to recreate the squared residuals if you already did so in (c).)
Stata output:

```
      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  2,    66) =   10.32
       Model |  20378554.1     2  10189277.1           Prob > F      =  0.0001
    Residual |  65145601.1    66  987054.563           R-squared     =  0.2383
-------------+------------------------------           Adj R-squared =  0.2152
       Total |  85524155.3    68  1257708.17           Root MSE      =  993.51

------------------------------------------------------------------------------
          u2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  profitshat |   70.87986   37.64392     1.88   0.064    -4.278665    146.0384
 profitshat2 |  -.2983731   .6413758    -0.47   0.643    -1.578921     .982175
       _cons |  -588.2597   348.3295    -1.69   0.096    -1283.722    107.2025
------------------------------------------------------------------------------
```

Special case of White test for heteroskedasticity:

- F statistic $= 10.32, p$ value $= 0.0001$ (easily obtained from Stata output). We can reject the null of homoskedasticity.
- LM statistic $= n \cdot R^2_{\widehat{u}^2} = 69\,(0.2383) = 16.443$, and the p value is less than 0.01 (the 1% critical value from the χ^2 distribution with 2 degrees of freedom is 9.21). We can reject the null of homoskedasticity.

Performing the White test in Stata:

```
reg profits assets mcontrol
imtest, white
```

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(4)      =     18.05
         Prob > chi2  =    0.0012

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      18.05      4    0.0012
            Skewness |       3.69      2    0.1581
            Kurtosis |       1.48      1    0.2230
---------------------+-----------------------------
               Total |      23.22      7    0.0016
---------------------------------------------------
```

Stata performs the original version of the White test (equation 8.19 of Wooldridge). The test statistic is different to the one we computed for the special case of the White test, but the conclusion is the same. We can reject the null of homoskedasticity.

(e)

$$profits/\sqrt{assets} = \beta_0/\sqrt{assets} + \beta_1 assets/\sqrt{assets} + \beta_2 mcontrol/\sqrt{assets} + u/\sqrt{assets}$$
$$profits/\sqrt{assets} = \beta_0\left(1/\sqrt{assets}\right) + \beta_1\sqrt{assets} + \beta_2 mcontrol/\sqrt{assets} + u^*$$

Stata commands:

```
gen sqrtassets=assets^(1/2)
gen profits_sqrtassets = profits/sqrtassets
gen one_sqrtassets = 1/sqrtassets
gen mcontrol_sqrtassets = mcontrol/sqrtassets
reg profits_sqrtassets one_sqrtassets sqrtassets mcontrol_sqrtassets,
noconstant
```

Results:

```
      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  3,    66) =   33.56
       Model |  41.8974441     3  13.9658147           Prob > F      =  0.0000
    Residual |  27.4621735    66  .416093538           R-squared     =  0.6041
-------------+------------------------------           Adj R-squared =  0.5861
       Total |  69.3596176    69  1.00521185           Root MSE      =  .64505

------------------------------------------------------------------------------
profits_sq~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
one_sqrtas~s |   2.492719   1.708424     1.46   0.149    -.9182605    5.903698
  sqrtassets |   .0406921   .0068165     5.97   0.000     .0270825    .0543017
mcontrol_s~s |  -2.029233   1.886417    -1.08   0.286    -5.795586    1.737119
------------------------------------------------------------------------------
```

These estimates are known as the *weighted least squares (WLS) estimates*. The WLS estimates are BLUE if $\text{Var}(u_i \mid assets_i) = \sigma^2 assets_i$.

(f) Special case of the White test:

```
predict ustar, resid
gen ustar2 = ustar^2
predict profits_sqrtassetshat
gen profits_sqrtassetshat2 = profits_sqrtassets^2
```

```
reg ustar2 profits_sqrtassetshat profits_sqrtassetshat2

      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  2,    66) =   94.16
       Model |  47.6341764     2  23.8170882           Prob > F      =  0.0000
    Residual |  16.6939312    66  .252938352           R-squared     =  0.7405
-------------+------------------------------           Adj R-squared =  0.7326
       Total |  64.3281077    68  .946001583           Root MSE      =  .50293

------------------------------------------------------------------------------
      ustar2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
profits_sq~t |   .5869394   .2360104     2.49   0.015     .1157295    1.058149
profits_sq~2 |   .3605197   .0306082    11.78   0.000     .2994084    .4216309
       _cons |  -.3926059   .1748171    -2.25   0.028    -.7416394   -.0435724
------------------------------------------------------------------------------
```

The F statistic $= 94.16$, with a p value of 0.0000. We can reject the null of homoskedasticity. This may be because our assumed form for the error variance is incorrect, or the model is mis-specified.

Original version of the White test:

```
reg profits_sqrtassets one_sqrtassets sqrtassets mcontrol_sqrtassets,
noconstant
imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(7)      =     16.49
         Prob > chi2  =    0.0210

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      16.49      7    0.0210
            Skewness |       2.92      3    0.4039
            Kurtosis |       2.13      1    0.1440
---------------------+-----------------------------
               Total |      21.54     11    0.0282
---------------------------------------------------
```

We can reject the null of homoskedasticity at the 5% level of significance.

(g) The 95% confidence interval for β_2 is given in the Stata output: $[-5.795586, 1.737119]$. (You should also make sure that you can construct this CI by hand.)

We cannot reject the null that management control has no effect on profits.

2. Let $arr86$ be a binary variable equal to unity if a man was arrested during 1986, and zero otherwise. The population is a group of young men in California born in 1960 or 1961 who have at least one arrest prior to 1986. A linear probability model for describing $arr86$ is

$$arr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u,$$

where

$$\begin{aligned} pcnv &= \text{the proportion of prior arrests that led to a conviction} \\ avgsen &= \text{the average sentence served from prior convictions (in months)} \\ tottime &= \text{months spent in prison since age 18 prior to 1986} \\ ptime86 &= \text{months spent in prison in 1986} \\ qemp86 &= \text{the number of quarters (0 to 4) that the man was legally employed in 1986} \end{aligned}$$

The data set is in CRIME1.DTA.

(a) Estimate this model by OLS and verify that all fitted values are strictly between zero and one. What are the smallest and largest fitted values?

(b) Estimate the equation by weighted least squares, as discussed in Section 8.5 of Wooldridge.

(c) Use the WLS estimates to determine whether $avgsen$ and $tottime$ are jointly significant at the 5% level.

SOLUTION:

You will have to first create the $arr86$ variable:

```
gen arr86 = 0
replace arr86 = 1 if narr86 > 0
```

(a)

```
      Source |       SS       df       MS              Number of obs =    2725
-------------+------------------------------           F(  5,  2719) =   27.03
       Model |  25.8452455     5  5.16904909           Prob > F      =  0.0000
    Residual |  519.971268  2719  .191236215           R-squared     =  0.0474
-------------+------------------------------           Adj R-squared =  0.0456
       Total |  545.816514  2724   .20037317           Root MSE      =  .43731

------------------------------------------------------------------------------
       arr86 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv | -.1624448   .0212368    -7.65   0.000    -.2040866    -.120803
      avgsen |  .0061127    .006452     0.95   0.344    -.0065385     .018764
     tottime | -.0022616   .0049781    -0.45   0.650    -.0120229    .0074997
     ptime86 | -.0219664   .0046349    -4.74   0.000    -.0310547   -.0128781
      qemp86 | -.0428294   .0054046    -7.92   0.000    -.0534268   -.0322319
       _cons |  .4406154   .0172329    25.57   0.000     .4068246    .4744063
------------------------------------------------------------------------------
```

Stata commands to get the fitted values and check that they lie between zero and one:

```
predict arr86hat
sum arr86hat
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    arr86hat |      2725    .2770642    .0974062   .0066431   .5576897
```

All the fitted values are between zero and one. The smallest fitted value is 0.0066431 and the largest fitted value is 0.5576897.

(b) The estimated heteroskedasticity function for each observation i is $\widehat{h}_i = \widehat{arr86}_i \left(1 - \widehat{arr86}_i\right)$, which is strictly between zero and one because $0 < \widehat{arr86}_i < 1$ for all i. The weights for WLS are $1/\widehat{h}_i$. We transform all the variables (including the intercept) by multiplying each variable by $1/\sqrt{\widehat{h}_i}$. Then we run OLS using the transformed variables to get the WLS estimates:
Stata commands to do this easily:

```
gen hhat = arr86hat * (1 - arr86hat)
reg arr86 pcnv avgsen tottime ptime86 qemp86 [aw = 1/hhat]
```

```
(sum of wgt is   1.5961e+04)

      Source |       SS       df       MS              Number of obs =    2725
-------------+------------------------------           F(  5,  2719) =   43.70
       Model |   36.631818     5  7.3263636            Prob > F      =  0.0000
    Residual |  455.826896  2719  .167645052           R-squared     =  0.0744
-------------+------------------------------           Adj R-squared =  0.0727
       Total |  492.458714  2724  .180785137           Root MSE      =  .40944

------------------------------------------------------------------------------
       arr86 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.1678436   .0189122    -8.87   0.000    -.2049272     -.13076
      avgsen |   .0053665   .0051146     1.05   0.294    -.0046624    .0153954
     tottime |  -.0017615   .0032514    -0.54   0.588     -.008137     .004614
     ptime86 |  -.0246188   .0030451    -8.08   0.000    -.0305898   -.0186479
      qemp86 |  -.0451885   .0054225    -8.33   0.000    -.0558212   -.0345558
       _cons |   .4475965   .0179922    24.88   0.000     .4123167    .4828763
------------------------------------------------------------------------------
```

(c) Stata command: `test avgsen tottime`

```
 ( 1)  avgsen = 0
 ( 2)  tottime = 0

       F(  2,  2719) =    0.88
            Prob > F =    0.4129
```

We cannot reject the null hypothesis and conclude that *avgsen* and *tottime* are not jointly significant at the 5% level of significance.

3. Use the data in WAGE1.DTA for this question.

(a) Estimate the equation

$$wage = \beta_0 + \beta_1 \ educ + \beta_2 \ exper + \beta_3 \ tenure + u.$$

Save the residuals and plot a histogram.

(b) Repeat part (a) but with $\log(wage)$ as the dependent variable.

(c) Would you say that Assumption MLR.6 in Wooldridge (Normality) is closer to being satisfied for the level-level model or the log-level model?

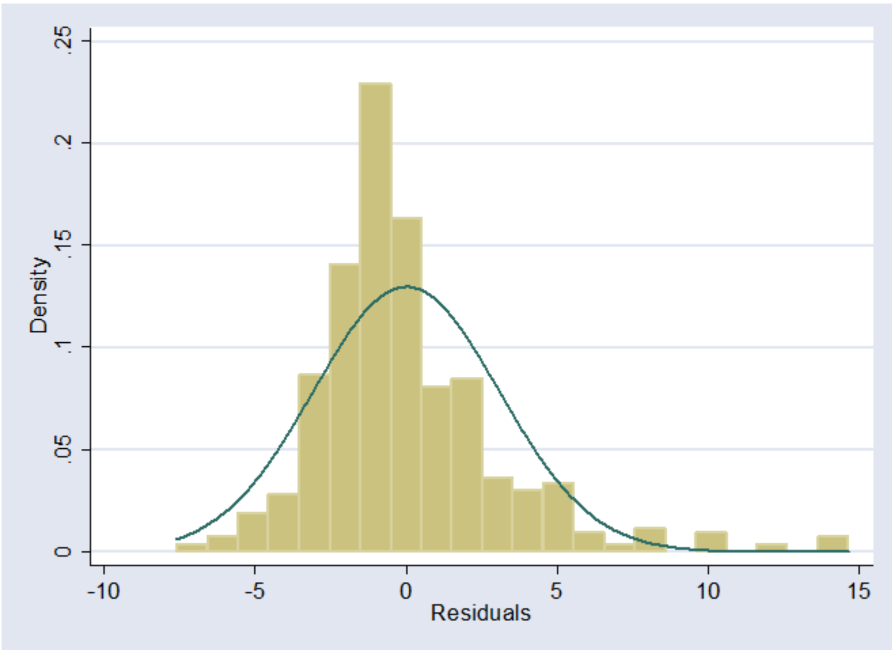
SOLUTION:

(a) `reg wage educ exper tenure`
`predict uhat, resid`
`histogram uhat, normal`

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  3,   522) =   76.87
       Model |   2194.1116     3  731.370532           Prob > F      =  0.0000
    Residual |  4966.30269   522  9.51398984           R-squared     =  0.3064
-------------+------------------------------           Adj R-squared =  0.3024
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.0845

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5989651   .0512835    11.68   0.000     .4982176    .6997126
       exper |   .0223395   .0120568     1.85   0.064    -.0013464    .0460254
      tenure |   .1692687   .0216446     7.82   0.000     .1267474    .2117899
       _cons |  -2.872735   .7289643    -3.94   0.000    -4.304799   -1.440671
------------------------------------------------------------------------------
```

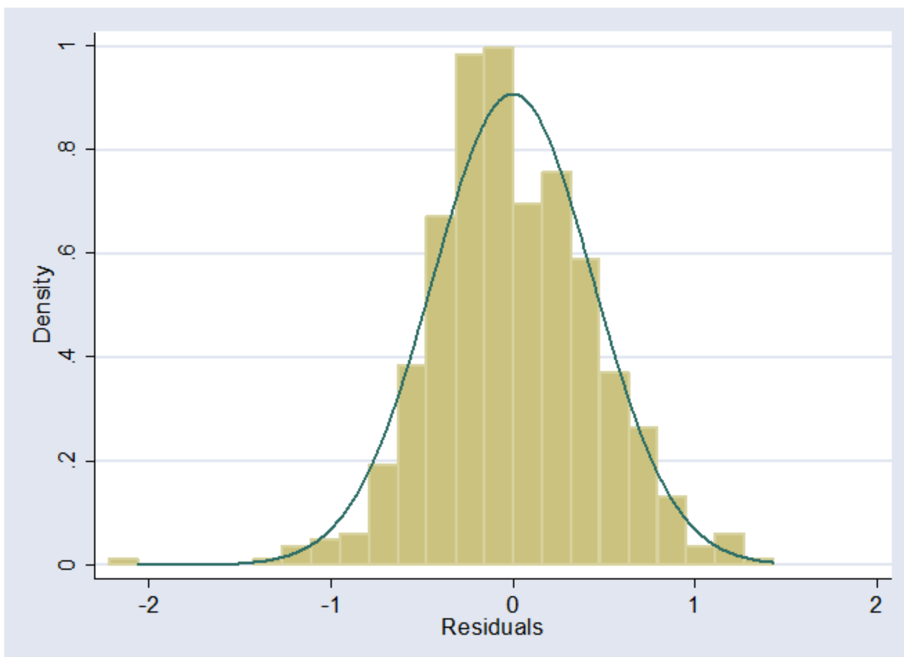
Use the Stata command **histogram varname, normal** to generate a histogram of **varname** with the normal density overlaid.



(b)

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  3,   522) =   80.39
       Model |  46.8741776     3  15.6247259           Prob > F      =  0.0000
    Residual |  101.455574   522  .194359337           R-squared     =  0.3160
-------------+------------------------------           Adj R-squared =  0.3121
       Total |  148.329751   525   .28253286           Root MSE      =  .44086

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .092029   .0073299    12.56   0.000     .0776292    .1064288
       exper |   .0041211   .0017233     2.39   0.017     .0007357    .0075065
      tenure |   .0220672   .0030936     7.13   0.000     .0159897    .0281448
       _cons |   .2843595   .1041904     2.73   0.007     .0796756    .4890435
------------------------------------------------------------------------------
```



(c) The residuals from the log(*wage*) regression appear to be more normally distributed. Certainly the histogram in part (b) fits under its comparable normal density better than in part (a), and the histogram for the *wage* residuals is notably skewed to the left. In the *wage* regression there are some very large residuals (roughly equal to 15) that lie almost five estimated standard deviations

($\widehat{\sigma} = 3.085$) from the mean of the residuals, which is identically zero, of course. Residuals far from zero does not appear to be nearly as much of a problem in the $\log(wage)$ regression.

4. Use the data in CHARITY.DTA for this question.

   (a) Using all 4,268 observations, estimate the equation

$$gift = \beta_0 + \beta_1 mailsyear + \beta_2 giftlast + \beta_3 propresp + u$$

   and interpret your results in full.

   (b) Reestimate the equation in part (a), using the first 2,134 observations.

   (c) Find the ratio of the standard errors on $\widehat{\beta}_2$ from parts (a) and (b). Compare this with the result from equation (5.10) on page 175 of Wooldridge.

SOLUTION:

(a)

```
      Source |       SS       df       MS              Number of obs =    4268
-------------+------------------------------           F(  3,  4264) =  129.26
       Model |  80700.7052     3  26900.2351           Prob > F      =  0.0000
    Residual |  887399.134  4264  208.114244           R-squared     =  0.0834
-------------+------------------------------           Adj R-squared =  0.0827
       Total |   968099.84  4267  226.880675           Root MSE      =  14.426

------------------------------------------------------------------------------
        gift |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   mailsyear |   2.166259   .3319271     6.53   0.000     1.515509    2.817009
    giftlast |   .0059265   .0014324     4.14   0.000     .0031184    .0087347
    propresp |   15.35861   .8745394    17.56   0.000     13.64405    17.07316
       _cons |  -4.551518   .8030336    -5.67   0.000    -6.125882   -2.977155
------------------------------------------------------------------------------
```

(b) Stata command: `reg gift mailsyear giftlast propresp in 1/2134`

```
      Source |       SS       df       MS              Number of obs =    2134
-------------+------------------------------           F(  3,  2130) =   44.95
       Model |  37014.4346     3  12338.1449           Prob > F      =  0.0000
    Residual |  584626.368  2130  274.472473           R-squared     =  0.0595
-------------+------------------------------           Adj R-squared =  0.0582
       Total |  621640.803  2133  291.439664           Root MSE      =  16.567

------------------------------------------------------------------------------
        gift |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   mailsyear |   4.975756   .9807182     5.07   0.000     3.052491    6.899021
    giftlast |   .0040329   .0016511     2.44   0.015      .000795    .0072708
    propresp |   14.03846   1.566064     8.96   0.000     10.96729    17.10963
       _cons |  -10.40117   2.419694    -4.30   0.000    -15.14638   -5.655958
------------------------------------------------------------------------------
```

(c) The ratio of the standard error of $\widehat{\beta}_2$ using 2,134 observations to that using 4,268 observations is $\dfrac{0.0016511}{0.0014324} = 1.1527$. From (5.10) we compute $\sqrt{\dfrac{4268}{2134}} = 1.4142$, which is somewhat below the ratio of the actual standard errors.