

**EViews**  
**Prof. Jonathan B. Hill**  
**University of North Carolina –Chapel Hill**

**Table of Contents**

<b>Topic</b>	<b>Page</b>
<b>I. Data files</b>	3
1. Import Excel data into a <i>workfile</i>	3
2. Save <i>workfile</i>	4
3. Open an existing <i>workfile</i>	4
<b>II. Create, Delete and View Data Series within a Data Workfile</b>	5
1. Creating a Time Trend-Variable	5
2. Quadratic and Exponential Trend	5
3. Create Any Function of Existing Variables	6
4. Viewing Data	6
5. Delete a Data Series	8
<b>III. Functions</b>	9
1. Observation and Date Functions	9
2. Mathematical Functions	9
3. Time Series Functions	9
<b>IV. Ordinary Least Squares Estimation</b>	10
1. Ordinary Least Squares	10
1.1 Define and Estimate a New Regression Equation: <b>Tool Bar</b> ( <i>point-click</i> )	10
1.2 Define and Estimate a New Regression Equation: <b>Program Space</b>	10
1.3 Altering an Existing Regression Equation	11
2. <i>Example #1: OLS and U.S. Mortality Rates</i>	11
3. Weighted Least Squares	12
4. <i>Example #2: WLS and U.S. Mortality Rates</i>	13
5. Tests of Linear/Non-linear Hypotheses ( <i>F</i> -tests of Compound Hypotheses)	14
1.2 Tests of Linear/Nonlinear Hypotheses	14
5.2 <i>Example #3: Education and U.S. Mortality Rates</i>	14
5.3 Chow's <i>F</i> -Test for Structural Change	14
6. Generating Variables: Functions of Regressors and Trends	15
6.1 Creating new variables	15
6.2 Adding functions of existing variables to a regression model	15
6.3 Trend Variables	16
7. Lagged Variables	16
<b>V. Regression Output: Viewing, Storing, Compiling with Test Results, Saving</b>	21
1. Viewing Regression Output: Numerical Output and Tests	21
1.1 VIEW	21
1.2 NAME	21
1.3 FREEZE	21

## Table of Contents Continued

Topic	Page
<b>V. Regression Output: Viewing, Storing, Compiling with Test Results, Saving</b>	
2. Viewing Regression Output: Graphical Output	22
2.1 View Graphical Output	22
2.2 Edit Graphical Output	22
2.3 Copy Graphical Output, Paste into Word	22
<b>VI. Advanced Regression Methods: GLS, SUR, IV, 2SLS, 3SLS</b>	22
1. Heteroscedasticity Robust Estimation	22
2. Seemingly Unrelated Regression	23
2.1 Example: SUR	23
2.2 Estimation of a System of Equations	23
2.3 <i>Example #4: U.S. Mortality Rates and SUR</i>	24
3. Instrumental Variables and Two Stages Least Squares	25
3.1 Endogenous Regressors	25
3.2 Instrumental Variables (IV)	25
3.3 Two Stages Least Squares (2SLS)	25
3.4 Testing for Endogeneity	26
4. <i>Example #5: U.S. Mortality Rates and 2SLS</i>	27
5. Two Stage Least Squares in a SUR System: Three Stages Least Squares	27
6. <i>Example #6: U.S. Morality Rates and 3SLS</i>	28
<b>VII. Limited Dependent Variables</b>	29
1. Binary Response	29
1.1 Binary Maximum Likelihood	29
1.2 Marginal Affects in Binary Response Models	29
1.3 Estimation in EVIEWS	30
1.4 Probit and Logit ML	30
2. <i>Example #7: Binary Choice, Labor Force Participation and Probit</i>	31
2.1 The Regression Model	31
2.2 Marginal Affects	32
3. Censored Regressions Models: The Tobit Model	33
3.1 Censored Regression Model	33
3.2 Tobit Model	34
3.3 Estimating the Tobit Model	34
4. <i>Example #8: Female Work Hours and Tobit Estimation</i>	34

## I. Data files

All EViews files are called “*workfiles*”, and contain imported Excel data, regression equations (if you create any), hypothesis test results, graphs (if you create any), etc. Thus, after working in EViews, if you want to save everything when you are done (data, an equation specification, OLS output, tests results, graphs), you will save it in a “*workfile*”.

All sessions with EViews begins with either opening an existing *workfile*, or creating a new *workfile*.

### 1. Create a new *workfile*

1.1 In the main EViews tool-bar, **FILE, NEW, WORKFILE**

1.2 A pop-up box appears, giving you choices for *Frequency* (time-period denominations: daily, weekly, quarterly, etc.) and *Workfile Range* (first period and last period in the data sample).

*yearly data:* type, for instance “1991”

*quarterly data:* “1991:1” denotes the first quarter of 1991

*monthly data:* “1991:10” denotes Oct., 1991

*weekly data:* For weekly and daily data, we specify the month, colon, the day, colon, then the year.

For example, “10:2:1991” denotes the day Oct. 2, 1991 for daily data, and the first week of Oct. 1991 for weekly data (Oct. 2<sup>nd</sup> occurs in the first week).

*daily data:* See “*weekly data*”.

#### Example

Suppose you have aggregate dividends and profits data with quarterly increments between for first quarter of 1970 and the last quarter of 1991. In the *Workfile Range* box, beneath *Start Date*, type

**1970:1**

and beneath *End Date*, type

**1991:4**

1.3 Once the *start-date* and *end-start* are entered: **OK**.

1.4 When the *workfile* is created, EViews creates a *workfile* box with a list of the current variables. By default, EViews stores the variable “*c*”, which is used to create the intercept in OLS regressions, and “*resid*”, which is where EViews stores regression residuals.

### 2. Import Excel data into a *workfile*

2.1 In the main EViews tool-bar, **FILE, IMPORT, READ-TEXT-LOTUS-EXCEL**.

2.2 Next, in the *Files of Type* box, scroll down and click-on **EXCEL**.

2.3 Next, in the *Look In* box, scroll down to the drive where your Excel data is stored (most likely on a disk, so scroll down to **drive A**), and then navigate until you find the file: double-click on the file name.

2.4 A pop-up box appears labeled *Excel Spreadsheet Import*. All the datasets for the class will be Excel data files with data starting in cell **A2**. The first row will always contain the variable name.

2.4.1 Beneath *Upper-left Data Cell*, type **A2**.

- 2.4.2 Beneath *Names for Series or Number if Name in File*, type the number of variables that the Excel file contains. You will always know this before hand. Finally, **OK**.
- 2.5 Once the Excel data is imported, it will listed in the *workfile* box, along with **c** and **resid**.

### **3. Save *workfile***

- 3.1 In the *workfile* tool-bar

**SAVE**

then choose the appropriate disk drive and create a file-name.

- 3.2 Once the *workfile* is saved, EViews automatically labels the *workfile* box with the name.

### **4. Open an existing *workfile***

Once a *workfile* has been created and saved, it can be opened for future use, with aspects of your recent econometric analysis still intact.

- 4.1 In the main EViews tool-bar

**FILE, OPEN, WORKFILE**

Next, choose the appropriate disk drive and *workfile*.

## II. Create, Delete and View Data Series within a Data Workfile

### 1. Create a Time Trend-Variable

1.1 In the *workfile* tool-bar

**GENR**

Then, type

*variable name* = **@trend**

**Example**

**t = @trend**

1.2 A preferable method is the following: in the *programmable white-area* below the main EViews tool-bar, type<sup>1</sup>

**series t = @trend**

Once the above command is typed in, hit the **ENTER** key, and EViews will perform the command.

1.3 EViews will create a variable called *t*, and will present the variable name *t* in the *workfile* list of variables.

### 2. Quadratic and Exponential Trend

2.1 Assume we have already created a time trend, described above. In the *workfile* tool-bar,

**PROCS, GENERATE SERIES**

Then, type

*Variable name* = **t^2**

*Variable name* = **@exp(t)**

2.2 A preferable method is the following: in the *programmable white-area* below the main EViews tool-bar, type

**series t = @trend**

**series t2 = t^2**

**series exp\_t = @exp(t)**

then **ENTER**. Note that the variable names used here, *t*, *t2*, and *exp\_t*, are arbitrary.

### 3. Create Any Function of Existing Variables

We can create any new variable as a function of existing variables.

3.1 In the *workfile* tool-bar,

**GENR**

Then type in the functional statement using existing function commands<sup>2</sup>.

---

<sup>1</sup> *SERIES* is the EViews command for the generation of a new variable.

**Example**

If *AGE* exists as a variable, age squared can be generated as, for example,

$$\text{AGE\_2} = \text{AGE}^2$$

If *GDP* exists,  $\ln(\text{GDP})$  can be generated as

$$\text{LN\_GDP} = \log(\text{GDP})$$

If *GDP* and *POP* (population) exist, then per-capita *GDP* can be generated as<sup>3</sup>

$$\text{GDP\_PC} = \text{GDP}/\text{POP}$$

3.2 Alternatively, we can use the white-area below the main EViews tool-bar. Use the command **SERIES**.

**Example**

```
series AGE_2 = AGE^2
series LN_GDP = log(GDP)
series GDP_PC = GDP/POP
```

**4. Viewing Data**

## 4.1 View Numerical Data

4.1.1 To view numerical data stored in a *workfile*, double-click on the variable name.

4.1.1.a EViews creates a box called **SERIES: Variable Name** in which the chosen variable is displayed in spreadsheet format.

4.1.2 In order to view *several variables at once*, while holding down the **CTRL** key, click on each variable name; then, click on **SHOW** in the *workfile* tool-bar, then **OK**.

4.1.3 Once the several variables are selected and “*shown*”, EViews creates a **GROUP: UNTITLED** box with the variable data in spread-sheet format on display.

## 4.2 Plotting Data and Sample Statistics

Once data is *shown* numerically (see 4.1, above), we can *visually* view the data and perform basic statistical analysis on the individual variables. All links begin with the *series* or *group* box for the chosen variable(s).

4.2.1 **VIEW, GRAPH** will plot all chosen series on one graph.  
**VIEW, MULTIPLE GRAPH** will plot each series separately if more than one variable was chosen.  
**VIEW, SPREADSHEET** returns you to the actual data for the chosen variables.

4.2.2 **VIEW, DISRIPTIVE STATISTICS, COMMON SAMPLE** derives mean, median standard deviation, and performs a test of normality (*Jarque-Bera* test)

4.2.3 **VIEW, CORRELATIONS, COMMON SAMPLE** derives correlations for all pairs of chosen variables.

---

<sup>2</sup> See Topic III on EViews functions and their associated command forms.

<sup>3</sup> Lower case and upper case are equivalent: *gdp* = *GDP*.



## 5. Delete a Data Series

5.1 Highlight the variable name in the *workfile* box. In the *workfile* tool-bar,

**DELETE**

5.2 Alternatively, click on the series, then right-click the mouse

**DELETE**

### III. Functions

#### 1. Observation and Date Functions

<b>@day</b>	Observation day for daily or weekly workfiles, returns the observation day in the month for each observation.
<b>@elem(x,d)</b>	Element returns the value of the series X, at date (or observation) d. d must be specified in double quotes " " or using the @str function.
<b>@month</b>	Observation month, returns the month of observation (for monthly, daily, and weekly data) for each observation.
<b>@quarter</b>	Observation quarter, returns the quarter of observation (except for annual, semi-annual, and undated data) for each observation.
<b>@year</b>	Observation year returns the year associated with each observation (except for undated data) for each observation.

#### 2. Mathematical Functions

<b>@abs(x), abs(x)</b>	
<b>@fact(x)</b>	factorial
<b>@exp(x), exp(x)</b>	
<b>@inv(x)</b>	inverse of $x = 1/x$
<b>@log(x), log(x)</b>	natural log
<b>@round(x)</b>	
<b>@sqrt(x)</b>	square root

#### 3. Time Series Functions

<b>d(x)</b>	first difference
<b>d(x,n)</b>	$n^{\text{th}}$ -order difference
<b>dlog(x)</b>	first difference of the log
<b>dlog(x,n)</b>	$n^{\text{th}}$ -order difference of the log
<b>@pch(x)</b>	one period percentage change (decimal)
<b>@pchy(x)</b>	one-year percentage change
<b>@seas(n)</b>	seasonal dummy: returns 1 when the quarter or month equals $n$ and 0 otherwise
<b>@trend</b>	generates a trend series, normalized to 0 at the first period/obs in the workfile
<b>@trend(n)</b>	generates a trend series, normalized to 0 at the $n^{\text{th}}$ period/obs in the workfile

## IV. Ordinary Least Squares Estimation

### 1. Ordinary Least Squares

#### 1.1 Define and Estimate a New Regression Equation: **Tool Bar** (*point-click*)

##### 1.1.1 In the main EVIEW tool-bar

#### QUICK, ESTIMATE EQUATION

An *estimation pop-up box* appears.

##### 1.1.2 In the *Equation Specification* space, type in the equation *without* “=”, and *with* “c” if an intercept is to be included.

##### 1.1.3 Beneath *Estimation Settings* and next to *Method*, scroll to **LS** (Least Squares). Usually, **LS** is the default setting. Finally, **OK**.

##### 1.1.4 Beneath *Estimation Settings* and next to *Sample*, alter the sample date-range if you want to use only a portion of the sample.

##### 1.1.5 An *equation* box appears with the OLS results: you can expand it or maximize it.

##### 1.1.5.a The *equation* box output can be *named*: see the topic **NAME EQUATION** below in topic V.3.

##### 1.1.5.b The *equation* box output can be “frozen” for *editing*, and *copying* into Word and Excel: see the topic **FREEZE** in topic V.2, below.

#### 1.2 Define and Estimate a New Regression Equation: **Program Space** (*the white area*)

Rather than pulling-up an estimation pop-up box, we can tell EVIEWS to estimate an equation directly in the programmable white-area beneath the main EVIEWS tool-bar.

The command **LS** tells EVIEWS to perform *least squares* estimation. After the command, type the equation *without* “=”, and *with* “c” if you want an intercept. After everything is typed in, hit **ENTER**.

#### 1.3 Altering an Existing Regression Equation

Once a regression equation is defined and estimated, EVIEWS creates an *equation* box with its own tool-bar. We can remove or add regressors, change the dependent variable, alter functions, and/or change the sample dates employed for estimation.

##### 1.3.1 In the *equation* box tool-bar, **ESTIMATE**: EVIEWS will show you the current regression equation.

##### 1.3.2 Type in the new equation specification. Change the *Method* and *Sample* range if desired.

### 2. Example #1: U.S. Mortality Rates

We have aggregate mortality rates *mort* for each of the 50 U.S. states plus the District of Columbia ( $n = 51$ ). We want to explain mortality rates based on the percent state adults with a college education *ed\_coll*, the number of physicians per 100,000 residents *phys*, and per capita annual expenditure on health care *health\_exp*. Since death

occurs no matter what (!), we should include a constant term. Since the effect education has on mortality may be nonlinear (decreasing, but at a decreasing rate) we will create and include squared *health\_exp*.

The model is

$$mort_i = \beta_0 + \beta_1 ed\_coll_i + \beta_2 ed\_coll_i^2 + \beta_3 phys_i + \beta_4 health + exp_i + \varepsilon_i$$

In the programmable white area type

**series ed\_coll\_2 = ed\_coll^2**

**ls mort c ed\_coll ed\_coll\_2 phys health\_exp**

The results, shown below, somewhat match our intuition about education, but the signs of the other explanatory variables suggests possible endogeneity. States may attract, or simply not repel, people with certain behaviors that are associated both with mortality rates and, say, education or health care expenditure: unobservable information contained in the error may be correlated with the regressors. A Two Stage Least Squares approach may be more appropriate.

The tabulated regression results are

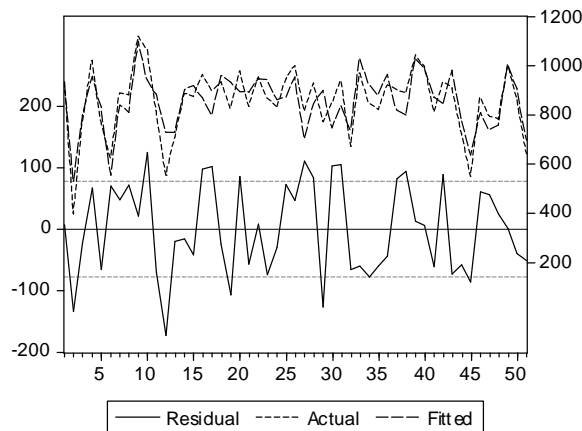
Dependent Variable: MORT  
 Method: Least Squares  
 Sample: 1 51  
 Included observations: 51

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	796.9879	251.7288	3.166058	0.0027
ED_COLL	-776.6662	2636.246	-0.294611	0.7696
ED_COLL_2	-10265.26	7978.920	-1.286548	0.2047
PHYS	1.819145	0.390880	4.653974	0.0000
HEALTH_EXP	0.074272	0.055304	1.342974	0.1859

R-squared	0.704472	Mean dependent var	855.0059
Adjusted R-squared	0.678774	S.D. dependent var	137.9660
S.E. of regression	78.19466	Akaike info criterion	11.64917
Sum squared resid	281262.6	Schwarz criterion	11.83857
Log likelihood	-292.0539	F-statistic	27.41344
Durbin-Watson stat	1.854259	Prob(F-statistic)	0.000000

The fitted values are (in the **equation box: View, Actual/Fitted/Residual:** see Part V)



### 3. Weighted Least Squares

3.1 In the main EVIEW tool-bar

#### QUICK, ESTIMATE EQUATION

3.2 Within the *Equation Specification* space, type in the equation.

3.3 In the *Method* box, scroll to **LS** (Least Squares). Alter the *Sample* range if required.

3.4 Click-on *Options, Weighted LS/TSLs* (check the box).

3.5 A space next to *Weight* appears: type in the variable name that serves as the weighting instrument, then **OK**.

#### Example:

We want to estimate an aggregate state-wide *health care expenditure* model

$$(1) \quad health_t = \beta_1 + \beta_2 senior_t + \beta_3 income_t + \varepsilon_t$$

where  $health_t$  denotes state-wide health care expenditure,  $senior_t$  denotes the percent of the  $t^{\text{th}}$  state's populace that is over the age of 65, and  $income_t$  denotes the  $t^{\text{th}}$  state's aggregate disposable income. We have evidence that the variance of health care expenditure is non-constant, and the proportional to the state's population size squared:

$$(2) \quad \sigma_t^2 = \sigma^2 pop_t^2$$

In this case, OLS is inefficient and standard hypothesis tests are invalid<sup>6</sup>. Employing *Feasible Generalized Least Squares* [FGLS], in this case, is equivalent to Weighted Least Squares, with weights equal to the population size. We want to estimate the transformed model

$$(1') \quad \frac{health_t}{pop_t} = \beta_1 \frac{1}{pop_t} + \beta_2 \frac{senior_t}{pop_t} + \beta_3 \frac{income_t}{pop_t} + \frac{\varepsilon_t}{pop_t}$$

which has a constant variance error,  $\varepsilon_t/pop_t$ .

In the main EVIEWS tool bar

#### QUICK, ESTIMATE EQUATION

Within the *Equation Specification* space, type in

**health c senior income**

Next to *Method*, scroll to **LS**; click-on *Options, Weighted LS/TSLs*.

Next to *Weight*, type

**pop**

Finally, **OK** and again **OK**.

---

<sup>6</sup> See Ramanathan, chapter 8, and Hill, chapter 4.

#### 4. Example #2: U.S. Mortality Rates

Reconsider U.S. mortality rates. There is evidence the dispersion of mortality rates is related to education. If we regress the squares residuals  $\hat{\varepsilon}_i^2$  from

$$mort_i = \beta_0 + \beta_1 ed\_coll_i + \beta_2 ed\_coll_i^2 + \beta_3 phys_i + \beta_4 health + exp_i + \varepsilon_i$$

on the regressors we find

Dependent Variable: E^2  
 Method: Least Squares  
 Sample: 1 51  
 Included observations: 51

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-6761.235	6567.843	-1.029445	0.3085
ED_COLL	57955.33	37372.16	1.550762	0.1277
PHYS	-64.83620	30.06773	-2.156339	0.0362
HEALTH_EXP	9.289032	4.374160	2.123615	0.0390
R-squared	0.104180	Mean dependent var		6716.724
Adjusted R-squared	0.047000	S.D. dependent var		6417.788
S.E. of regression	6265.156	Akaike info criterion		20.39858
Sum squared resid	1.84E+09	Schwarz criterion		20.55010
Log likelihood	-516.1638	F-statistic		1.821959
Durbin-Watson stat	2.231099	Prob(F-statistic)		0.156061

College education and health care expenditure are associated with greater state-to-state differences in mortality rates, on average, while more physicians renders mortality rates more homogenous. In fact, the F-test is a test of heteroscedasticity. The  $p$ -value for  $F$  is about 10%, but it does suggest the homoscedasticity assumption is invalid.

We can use the positively associated factor  $ed\_coll$  in WLS by assuming

$$\sigma_i^2 = \sigma^2 ed\_coll_i^2$$

The results are

Dependent Variable: MORT  
 Weighting series: ED\_COLL\_2

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	987.7339	345.4736	2.859072	0.0064
ED_COLL	-2557.626	3264.201	-0.783538	0.4373
ED_COLL_2	-6204.347	8826.086	-0.702956	0.4856
PHYS	2.075763	0.336755	6.164025	0.0000
HEALTH_EXP	0.040821	0.056756	0.719240	0.4756
Weighted Statistics				
R-squared	0.958101	Mean dependent var		837.3803
Adjusted R-squared	0.954458	S.D. dependent var		400.7874
S.E. of regression	85.53036	Akaike info criterion		11.82851
Sum squared resid	336510.4	Schwarz criterion		12.01791
Log likelihood	-296.6271	F-statistic		55.81226
Durbin-Watson stat	1.756078	Prob(F-statistic)		0.000000

All goodness-of-fit criteria have improved. Note the dependent variable has changed so caution about such comparisons is advised.

## 5. Tests of Linear/Non-linear Hypotheses

### 5.1 Tests of Linear/Nonlinear Hypotheses (i.e. $F$ -tests of compound hypothesis)

#### 5.1.1 In the *equation* box

#### VIEW, COEFFICIENT TESTS, WALD-COEFFICIENT RESTRICTIONS

#### 5.1.2 Beneath *Coefficient Restrictions*, type the restrictions of coefficients $c(i)$ with commas.

### 5.2 Example #3: U.S. Mortality Rates

For the mortality model

$$mort_i = \beta_0 + \beta_1 ed\_coll_i + \beta_2 ed\_coll_i^2 + \beta_3 phys_i + \beta_4 health + exp_i + \varepsilon_i$$

we want to test if education has a zero impact:

$$H_0 : \beta_1 = \beta_2 = 0$$

These are the second and third parameters, hence beneath *Coefficient Restrictions*

$$c(2) = 0, c(3) = 0$$

The results indicate overwhelming rejection of the null in favor of the alternative.

Wald Test: Equation: Untitled			
Test Statistic	Value	df	Probability
F-statistic	49.83546	(2, 46)	0.0000
Chi-square	99.67092	2	0.0000

### 5.2 Chow's F-Test for Structural Change

#### 5.2.1 If the data is cross-sectional, the data needs to be sorted according to the binary quality which is being tested (e.g. female vs. male). Once the data is sorted, the observation number where the binary variable value changes to 1 needs to be obtained.

If the data is a *time series* and we want to test for a *structural change at some point in time*, we need to obtain the precise date.

#### 5.2.2 In the *equation* box

#### VIEW, STABILITY TESTS, CHOW BREAKPOINT TEST

**Then, enter the date (if time series) or observation number (if cross-sectional).**

#### Example

We want to estimate an aggregate quarterly dividends model with lagged profits and *GDP*:

$$DIV_i = \beta_1 + \beta_2 PROFITS_{t-1} + \beta_3 GDP_{t-1} + \varepsilon_i$$

See IV.4, below, for instructions on using lagged values. We estimate the model by least squares, **QUICK, ESTIMATE EQUATION**, and type in the equation

## div c profits(-1) gdp(-1)

We want to test for a change in the underlying structure before and after 1981:1 (when Reagan entered the presidency). The null hypothesis is that there was not a change. In the resulting *equation* box, **VIEW, STABILITY TESTS, CHOW BREAKPOINT TEST**, and type the date

**1981:1**

The resulting *F*-statistic has an associated *p*-value of .0513, thus we reject the null of no-change in regression structure in 1981. We have some evidence that the Reagan presidency inaugurated a corporate era with fundamentally different dividend pay-off trends.

## 6. Generating Variables: Functions of Regressors and Trends

### 6.1 Creating New Variables

- 6.1.1 In the *workfile* box tool-bar, **GENR**, then type in the functional statement using existing function commands<sup>7</sup>.

For example, if *AGE* exists as a variable, age squared can be generated as, for example,

$$\mathbf{AGE\_2 = AGE^2}$$

If *GDP* exists,  $\log(GDP)$  can be generated as

$$\mathbf{LN\_GDP = \log(GDP)}$$

If *GDP* and *POP* (population) exist, then per-capita *GDP* can be generated as

$$\mathbf{GDP\_PC = GDP/POP}$$

- 6.1.2 For a better method, we can use the *programmable white-area* beneath the main **EViews** tool-bar. Use the command **SERIES**:

```
series age_2 = age^2  
series ln_gdp = log(gdp)  
series gdp_pc = gdp/pop
```

Or, for a trend variable,

```
series t = @trend
```

After each line is typed, be sure to hit the **ENTER** key: **EViews** will perform the command only after the **ENTER** key is hit.

### 6.2 Adding Functions of Existing Variables to a Regression Model

Any function of existing variables can be added to a regression equation, whether the variable function was already created or not.

#### Example

Suppose we have the variables *WAGE*, *AGE*, and *ED*, and we want to estimate

$$WAGE_t = \beta_1 + \beta_2 AGE_t + \beta_3 AGE_t^2 + \beta_4 ED_t + \varepsilon_t$$

---

<sup>7</sup> See Part III on **EViews** functions.

Then, in the *programmable white-area*, type

```
ls wage c age age^2 ed
```

and hit **ENTER**. EViews recognizes that it needs to create the function **age^2**.

### Example

Consider estimating a log-model of corporate profits:

$$\log(\text{profits}_t) = \beta_1 + \beta_2 \log(\text{gdp}_t) + \varepsilon_t$$

However, we only have data on *profit* and *GDP*. Then, in the *programmable white-area*, type

```
ls log(profits) c log(gdp)
```

and hit **ENTER**.

## 6.3 Trend Variables

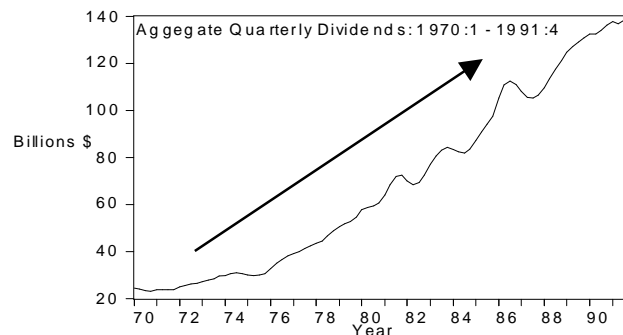
Any *time trend variable* and any function of time trend variable can be added to a regression model to account for deterministic (*non-random*) trend in a time series.

6.3.1 See topics II.1 – II.3 for instructions on how to create linear, quadratic and exponential trend variables.

6.3.2 In order to include a time trend variable or function of such a variable, simply add it to the regression *Equation Specification*.

### Example: Time Trend Model of Aggregate Quarterly Dividends

Aggregate quarterly dividends in the U.S. during the period 1970:1 – 1991:4 display a strong linear time trend:



In order to account for a likely **linear time trend**, we may specify a simple linear trend model

$$\text{div}_t = \beta_1 + \beta_2 t + \varepsilon_t$$

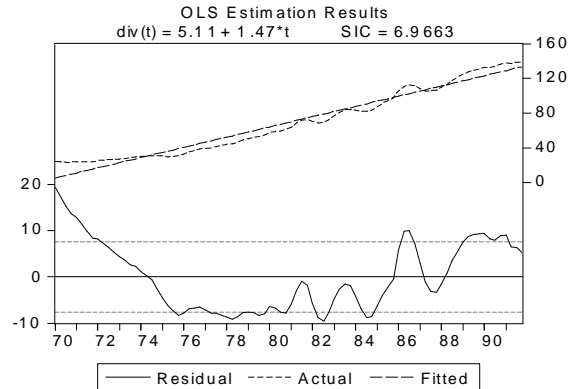
where  $t = 1, 2, \dots, T$ , where  $T = 88$  because we 88 points of data.

In the *programmable white-area*, type

```
series t = @trend  
ls div c t
```

Be sure to type **ENTER** at the end of each line.

The results from OLS estimation of the above model follow:



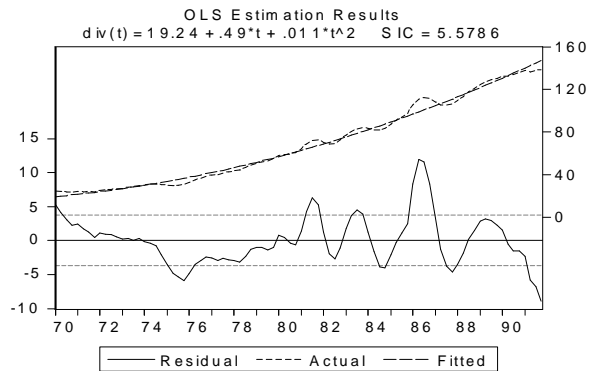
The linear time-trend regression fit is very poor based on residual tests of autocorrelation and the SIC. Consider, instead, a quadratic trend model:

$$div_t = \beta_1 + \beta_2 t + \beta_3 t^2 + u_t$$

In the *programmable white-area*, type

```
series t = @trend
ls div c t t^2
```

Be sure to type **ENTER** at the end of each line. The results follow:



The residuals appear to be more “random” and “noisy”, although, in fact, they are not: there are clear signs of cycles in the residuals suggesting severe autocorrelation and omitted variables (i.e. there exists some neglected dividends structure that we need to model using techniques in Topic VI).

## 7. Lagged Variables

Any existing variable can be lagged<sup>8</sup> for a subsequent regressor. For example, if  $DIV$ ,  $PROFITS$  and  $GDP$  are the existing variables, we can generate related lagged variables by employing, for example,  $GDP(-1)$  or  $GDP(-2)$ , etc.

### Example:

Suppose  $DIV$ ,  $PROFITS$  and  $GDP$  are the existing variables. We are interested in modeling corporate dividend payouts as a function of national income and profits. However, current dividends are paid from *past* profits:

$$DIV_t = \beta_1 + \beta_2 PROFITS_{t-1} + \beta_3 GDP_t + \varepsilon_t$$

In the *programmable white-area*, type

```
ls div c profits(-1) gdp
```

Then, **ENTER**.

### Example:

Consider estimating the same model with logged values:

$$\log(DIV_t) = \beta_1 + \beta_2 \log(PROFITS_{t-1}) + \beta_3 \log(GDP_t) + \varepsilon_t$$

In the *programmable white-area*, type

```
ls log(div) c log(profits(-1)) log(gdp)
```

Then, **ENTER**. Be careful with *parentheses*.

### Example:

We want to estimate an  $AR(3)$  model<sup>9</sup> of corporate dividends:

$$DIV_t = \beta_1 + \beta_2 DIV_{t-1} + \beta_3 DIV_{t-2} + \beta_4 DIV_{t-3} + \varepsilon_t$$

In the *programmable white-area*, type

```
ls div c div(-1) div(-2) div(-3)
```

Then, **ENTER**.

---

<sup>8</sup> A “lagged” variable is a past value of a variable. Thus, for  $GDP_t$ , in the  $t^{\text{th}}$  month, the one-month lagged value of  $GDP$  is  $GDP_{t-1}$ . The 12-month (one-year) lagged value of  $GDP$  is  $GDP_{t-12}$ .

<sup>9</sup> “ $AR(3)$ ” denotes “autoregressive of order 3”: a model which regresses a variables on itself (hence, “auto”, Latinate for “self”) 3 periods into the past (hence, “order 3”). See Topic VI, and Diebold, chapters 6-9.

## V. Regression Output: Viewing, Storing, Compiling with Test Results, Saving

After we estimate a regression model, EVIEWS creates an *equation* box with a *tool-bar*. We can view residuals, perform sophisticated hypothesis tests concerning correlated errors, errors with non-constant variance, ARCH errors, as well as *print*, *save*, and *forecast*.

### 1. Viewing Regression Output: Numerical Output and Tests

#### 1.1 VIEW

Located in the *equation* box tool-bar: Navigates through the equation representation, OLS output and hypothesis tests.

##### 1.1.1 REPRESENTATIONS

The specified model based on the typed equation, and the actual mathematical representation.

##### 1.1.2 ACTUAL, FITTED, RESIDUAL

Self explanatory: plots the dependent variable  $y$ , the fitted values and the residuals.

##### 1.1.3 ESTIMATION OUTPUT

Default: displays the actual OLS output.

##### 1.1.4 COEFFICIENT TESTS

Allows us to perform tests of compound hypotheses on the estimated parameters, including *omitted* variables, *redundant* variables and standard *Wald* tests of *linear coefficient restrictions*.

See **Topic IV.3** on hypothesis tests.

##### 1.1.5 RESIDUALS TESTS

Allows us to perform tests for *autocorrelated* errors, errors with *non-constant variance* (i.e. *heteroscedastic* errors), and a combination of the two in the form of ARCH errors (i.e. *correlated variances*).

#### 1.2 FREEZE

Located in the *equation* box tool-bar.

1.2.1 **FREEZE** stores the regression output in an Excel spread-sheet format called a “*table*”. Once the regression output is “*frozen*”, we can directly edit the results, add titles, and copy-paste the results into Excel or Word. Every EVIEWS graph and table was pasted directly into this document.

#### 1.3 NAME

Located in the *equation* box tool-bar, assigns a name to any *equation* box of results.

1.3.1 Click-on **NAME** in the *equation* box tool-bar. Beneath *Name to identify object*, type the name of your preference.

1.3.2 EVIEWS will place the *equation* name in the list of variables in the *workfile* box. If you name the equation, say, “*eq01*”, then EVIEWS creates the label “=*eq01*” in the *workfile* box.

1.3.3 Once the *workfile* is saved, all named objects will be saved to, including *equations* and *tables* of output.

1.3.4 *Equations need to be named in order for multiple regressions to be performed*. If the Equation is left unnamed as “Untitled”, EVIEWS will attempt to delete the regression results when another equation is estimated.

1.3.5 Once the equation is named, you can click-on the cross in the upper-right corner of the *equation* box in order to remove the equation results from view. EVIEWS, however, stores the equation information: click on the equation name icon in the *workfile* box in order to display the *equation* box once again.

## 2. Viewing Regression Output: Graphical Output

### 2.1 View Graphical Output

In the *equation* box tool-bar, click-on

**RESID**

EViews will display a graphical plot of the variable which is being modeled, the fitted (predicted) values, and the residuals. The plot is called **GRAPH:UNTITLED**.

#### Example:

We are interested in modeling a time trend-model for deaths due to AIDS in the period 1988:1 – 1999:2. Using a polynomial trend model,

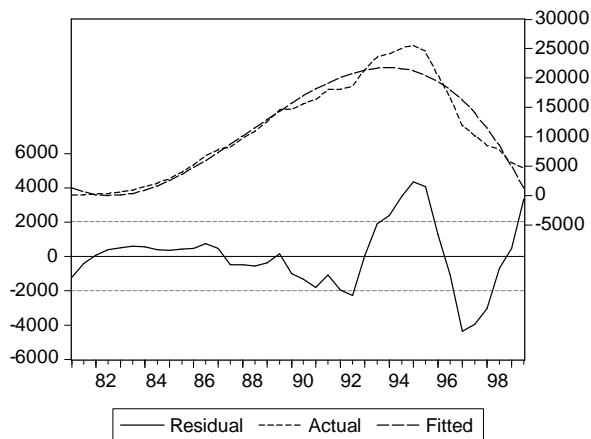
$$y_t = \beta_1 + \beta_2 t + \beta_3 t^2 + \beta_4 t^3 + \varepsilon_t$$

where  $y_t$  denotes the number of deaths due to AIDS in the  $t^{\text{th}}$  period.

In the *programmable white-area*, we type

```
series t = @trend
ls y c t t^2 t^3
```

In the *equation* box, click-on **RESID** to obtain



### 2.2 Edit Graphical Output

Notice that the above graph is not titled. We can create a title as well as commentary inside the graph itself. Moreover, the shape of the lines (thickness, color, symbols) can be edited.

#### 2.2.1 Graph Title

2.2.1.a In the *graph* box

**ADD TEXT.**

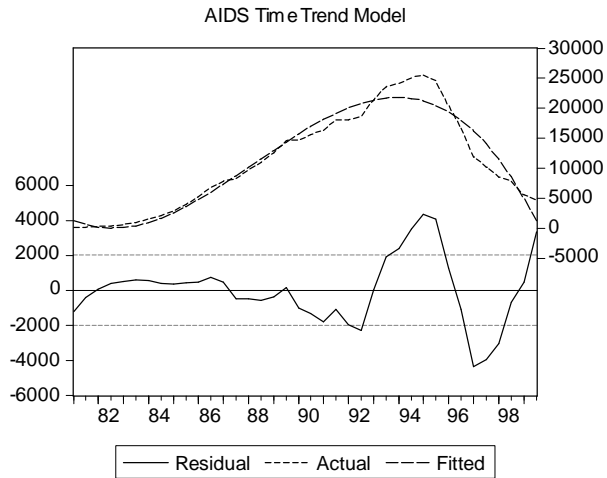
Beneath *Justification*, click-on *Center*.

2.2.1.b Beneath *Position*, click-on *Top*.

2.2.1.c In the *Text for label* area, type the title of your choice.

**Example:**

For the above polynomial time-trend AIDS model, we will use the title *AIDS Time Trend Model*:



2.3 Copy Graphical Output, Paste into Word

Suppose a graph has been created and frozen. For example, the above regression results on AIDS deaths have been frozen and titled as a figure. Open **Word** or **Excel**. Go back to EViews.

5.3.1 In the *graph* box tool-bar, click-on

**PROCS**

Then

**Save graph as metafile**

Finally, click-on

**Copy to clip-board**

2.3.2 Go to **Word** or **Excel**. In **Word** or **Excel**, simply go to the main tool-bar, **EDIT, PASTE**. Because the EViews graph was saved in the clip-board, and EViews is Windows based, **Word** will simply paste the graph itself. Alternatively, hold the Control key and type  $\nu$ : **CNTR  $\nu$** . Once the graph, etc., has been pasted, it will be very large: click-on the object to highlight the corners, then click on the corners, hold and drag to re-shape the object.

## VI. Advanced Regression Methods: GLS, IV, 2SLS, 3SLS

Eviews allows the analyst to perform aspects of Generalized Least Squares including heteroscedasticity and autocorrelation robust standard errors. It allows for a wide array of estimation techniques for systems of equations, in particular when regressors are endogenous. These include Instrumental Variables (IV), Seemingly Unrelated Regression (SUR), Two Stages Least Squares (2SLS) as a two-step IV estimator, and Three Stages Least Squares (3SLS) as a combination of 2SLS with heteroscedasticity or autocorrelation robusification.

### 1. Heteroscedasticity Robust Estimation

Weighted Least Squares (WLS) allows for a direct solution to heteroscedasticity. See Part IV.3. That method, however, requires substantial faith that we chosen the correct weight. Instead we may simply use robust  $t$ -statistics by a method due to H. White (1982). In short, White (1982) suggests using the available regressors to generate standard errors that are robust to an unknown form of heteroscedasticity.

1.1 In the tool bar: **QUICK, ESTIMATE EQUATION.**

1.2 After the equation is typed in the white area click **OPTIONS, HETEROSC. CONSISTENT COVARIANCE, WHITE**, then ok.

1.3 The resulting  $t$ -statistics will be robust to any form of heteroscedasticity that is related to the included regressors.

#### Example: U.S. Mortality Rates

Recall we want to estimate

$$mort_i = \beta_0 + \beta_1 ed\_coll_i + \beta_2 ed\_coll_i^2 + \beta_3 phys_i + \beta_4 health + exp_i + \varepsilon_i$$

We found evidence the regression error variance may depend on the included regressors. If we use White's robust  $t$ -test we find

Dependent Variable: MORT

Method: Least Squares

Sample: 1 51

Included observations: 51

White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.	<b>t-Statistic</b>	<b>Prob.</b>
C	796.9879	184.6299	4.316678	0.0001	<b>3.166058</b>	<b>0.0027</b>
ED_COLL	-776.6662	1901.757	-0.408394	0.6849	<b>-0.294611</b>	<b>0.7696</b>
ED_COLL_2	-10265.26	6320.866	-1.624028	0.1112	<b>-1.286548</b>	<b>0.2047</b>
PHYS	1.819145	0.466715	3.897769	0.0003	<b>4.653974</b>	<b>0.0000</b>
HEALTH_EXP	0.074272	0.057531	1.290986	0.2032	<b>1.342974</b>	<b>0.1859</b>
R-squared	0.704472	Mean dependent var		855.0059		
Adjusted R-squared	0.678774	S.D. dependent var		137.9660		
S.E. of regression	78.19466	Akaike info criterion		11.64917		
Sum squared resid	281262.6	Schwarz criterion		11.83857		
Log likelihood	-292.0539	F-statistic		27.41344		
Durbin-Watson stat	1.854259	Prob(F-statistic)		0.000000		

For comparisons sake we include the non-robust  $t$ -statistics in bold. Education is insignificant at the 15% level, while  $phys$  has gained in significance.

White's (1982) test of heteroscedasticity in fact is little more than a test that the robust standard errors and non-robust standard errors are identical for large samples.

## 2. Seemingly Unrelated Regression

There are many occasions when a system of equations exists which appear to be unrelated because each dependent variable is different, and each set of regressors for each dependent variable is different. The fundamental link is the errors themselves.

### 2.1 Example: SUR

The U.S. state-wide mortality model is

$$mort_i = \beta_0 + \beta_1 ed\_coll_i + \beta_2 ed\_coll_i^2 + \beta_3 phys_i + \beta_4 health + exp_i + \varepsilon_i$$

We also have information on tobacco expenditure per capita<sup>10</sup> (*tob*), percent of adult population with a high school education (*ed\_hs*), per capita income (*inc*) and the percent of the population above the age of 65 (*aged*)<sup>11</sup>. We conjecture that tobacco use is related to income level, high school educatedness and youth:

$$tob_i = \delta_0 + \delta_1 ed\_hs_i + \delta_2 inc_i + \delta_3 aged_i + u_i$$

The single equation results follow:

Dependent Variable: TOB\_PC  
 Method: Least Squares  
 Sample: 1 51  
 Included observations: 51

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	182.6473	37.23020	4.905890	0.0000
ED_HS	-143.1016	45.27598	-3.160652	0.0028
INC	0.003046	0.001511	2.015812	0.0496
AGED	-49.88707	141.2627	-0.353151	0.7256
R-squared	0.185482	Mean dependent var		120.5275
Adjusted R-squared	0.133492	S.D. dependent var		22.13050
S.E. of regression	20.60049	Akaike info criterion		8.963691
Sum squared resid	19945.86	Schwarz criterion		9.115207
Log likelihood	-224.5741	F-statistic		3.567621
Durbin-Watson stat	1.966861	Prob(F-statistic)		0.020888

Tobacco appears to be normal good, negatively related to having a high school education.

The mortality and tobacco use regressions are seemingly unrelated, but *clearly related*. The errors terms  $\varepsilon$  and  $u$  capture unobservable characteristics of state residents, including cultural traits (diet, risk taking) and sociological traits (social networks, religion). Indeed, a state with a high mortality rate may be a state with high tobacco use (see footnote!), hence the errors undoubtedly are related.

It is perfectly fine to estimate equation alone, but we are neglecting possible important information associated with the error term correlation. This implies a potentially more efficient set of estimates may exist if we estimate the equations at the same time and while simultaneously allowing the errors to be correlated. This is *Seemingly Unrelated Regression*.

### 2.2 Estimation of a System of Equations

2.2.1 In the main toolbar click **OBJECTS, NEW OBJECTS, SYSTEM**. Before you click SYSTEM, *name the object* (e.g. mort\_sur).

2.2.2 In the pop-up box type the system of equations using c(1) for the constant, and so on.

<sup>10</sup> Tobacco related products: cigarettes, cigars and chewing tobacco.

<sup>11</sup> Needless to say all this information belongs in the mortality regression! This is up to the student to do during the semester.

**Example:**

The mortality system is typed

$$\text{mort} = \text{c}(1) + \text{c}(2)*\text{ed\_coll} + \text{c}(3)*\text{ed\_coll\_2} + \text{c}(4)*\text{phys} + \text{c}(5)*\text{health\_pc}$$

$$\text{ob\_pc} = \text{c}(6) + \text{c}(7)*\text{ed\_hs} + \text{c}(8)*\text{inc\_pc} + \text{c}(9)*\text{aged}$$

- 2.2.3 On the system is typed, click **ESTIMATE** from the pop-up box toolbar.
- 2.2.4 A new box appears with a list of choices. Click **SEEMINGLY UNRELATED REGRESSION**.
- 2.2.5 There are choices for handling how the correlation between the errors is estimated and these are used to estimate the system of equations. Unfortunately this choice may have a profound impact on the subsequent results.

**2.3 Example #4: U.S. Mortality Rates**

We estimate the above system by SUR. We create the new object: **OBJECTS, NEW OBJECT, SYSTEM**, naming the system *mort\_sur*. The type

$$\text{mort} = \text{c}(1) + \text{c}(2)*\text{ed2\_coll} + \text{c}(3)*\text{ed\_coll\_2} + \text{c}(4)*\text{phys} + \text{c}(5)*\text{health\_pc}$$

$$\text{tob\_pc} = \text{c}(6) + \text{c}(7)*\text{ed\_hs} + \text{c}(8)*\text{inc\_pc} + \text{c}(9)*\text{aged}$$

click **ESTIMATE**, choose **SEEMINGLY UNRELATED REGRESSION** and **ITERATED COEFFICIENTS TO CONVERGENCE**. The results follow:

System: SUR\_MORT  
 Estimation Method: Seemingly Unrelated Regression  
 Sample: 1 51  
 Included observations: 51  
 Total system (balanced) observations 102  
 Linear estimation after one-step weighting matrix

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	786.3284	234.4604	3.353780	0.0012
C(2)	-816.1597	2448.351	-0.333351	0.7396
C(3)	-9834.407	7402.848	-1.328463	0.1873
C(4)	1.795136	0.363124	4.943593	0.0000
C(5)	0.079825	0.051222	1.558407	0.1225
C(6)	208.9991	35.13037	5.949242	0.0000
C(7)	-159.2617	42.70680	-3.729188	0.0003
C(8)	0.003211	0.001430	2.245032	0.0271
C(9)	-198.1056	132.9795	-1.489745	0.1397
<b>Determinant residual covariance</b>		1955256.		
Equation: MORT = C(1) +C(2)*ED_COLL+C(3)*ED_COLL_2+C(4) *PHYS+C(5)*HEALTH_PC				
Observations: 51				
R-squared	0.703673	Mean dependent var	855.0059	
Adjusted R-squared	0.677906	S.D. dependent var	137.9660	
S.E. of regression	78.30028	Sum squared resid	282022.9	
Durbin-Watson stat	1.890212			
Equation: TOB_PC = C(6) + C(7)*ED_HS + C(8)*INC_PC + C(9)*AGED				
Observations: 51				
R-squared	0.166149	Mean dependent var	120.5275	
Adjusted R-squared	0.112925	S.D. dependent var	22.13050	
S.E. of regression	20.84353	Sum squared resid	20419.29	
Durbin-Watson stat	1.931746			

Compare the mortality regression results in bold with the OLS results from Part V.2. There is essentially no difference in the percent of mortality rate variation explained by the regression model, and all coefficient estimates are qualitatively similar. There may, indeed, not be a SUR effect (estimation of the system offers no boost in efficiency over single equation estimation).

### 3. Instrumental Variables (IV) and Two Stages Least Squares (2SLS)

There are several possible reasons a regression model may be poorly specified. In any regressor, for example, is correlated with the error than OLS fails to be product consistent and unbiased estimates.

#### 3.1 Endogenous Regressors

- 3.1.1 If a regressor or regressors  $x_i$  is correlated with the error term  $\varepsilon_i$ , conventional least squares does not deliver a consistent and unbiased estimator. If a set of valid substitute regressors  $z_i$ , or “instruments”, is available, then a least squares can be performed.
- 3.1.2 Validity is determined by *i.* the set  $z_i$  is correlated with  $x_i$  and *ii.*  $z_i$  is uncorrelated with  $\varepsilon_i$ .
- 3.1.3 Straight substitution of  $z_i$  for  $x_i$  is Instrumental Variables. But this begs the questions: if many valid instruments exist, which do we choose?

**Example:**

Reconsider U.S. mortality rates:

$$mort_i = \beta_0 + \beta_1 ed\_coll_i + \beta_2 ed\_coll_i^2 + \beta_3 phys_i + \beta_4 health + exp_i + \varepsilon_i$$

We can easily argue that the unobservable characteristics of each state, which affect mortality rates (e.g. state resident risk taking behavior, cultural information associated with marketable skills) also affect the desire and/or ability to obtain a college education, to seek medical help (e.g. health care expenditure), and to demand medical care (e.g. physician count per 100,000 resident).

#### 3.2 Instrumental Variables (IV)

- 3.2.1 The IV approach is to use a direct variable-by-variable substitute for the endogenous regressors. If a set of regressors exists then there is an optimal method for combining them to form a “best” set of IV’s: simply generate predicted values of the endogenous  $x_i$  by regressing them one by one on the IV’s  $z_i$ .
- 3.2.2 Any variable uncorrelated with the error can be used as an instrument.
- 3.2.3 Creating this “best” set is stage one, and using them as IV’s is stage two of Two Stages Least Squares.
- 3.2.4 EVIEWS’s Two Stages Least Squares routine requires at least as many IV’s as variables in the regression model.

#### 3.3 Two Stages Least Squares (2SLS)

- 3.3.1 In the main toolbar **QUICK, ESTIMATE EQUATION**, type the equation

$$mort = c(1) + c(2)*ed2\_coll + c(3)*ed\_coll\_2 + c(4)*phys + c(5)*health\_exp$$

and scroll through **METHOD** to find **TSLS** (i.e. 2SLS).

- 3.3.2 Since we believe *health\_exp* is endogenous, we include all other regressors and the IV *inc* as the instruments. Type in the instrument box

**ed2\_coll ed\_coll\_2 phys inc\_pc**

Then **ok**.

Dependent Variable: MORT  
Method: Two-Stage Least Squares  
Sample: 1 51  
Included observations: 51  
MORT = C(1)+C(2)\*ED2\_COLL+C(3)\*ED\_COLL\_2+C(4)\*PHYS+C(5)\*HEALTH\_PC  
Instrument list: ED2\_COLL ED\_COLL\_2 PHYS INC\_PC

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	850.9452	326.6824	2.604809	0.0123
C(2)	-1081.035	2891.747	-0.373835	0.7102
C(3)	-9559.261	8452.073	-1.130996	0.2639
C(4)	1.976187	0.719358	2.747154	0.0086
C(5)	0.043648	0.130032	0.335667	0.7386
R-squared	0.702502	Mean dependent var		855.0059
Adjusted R-squared	0.676633	S.D. dependent var		137.9660
S.E. of regression	78.45485	Sum squared resid		283137.5
Durbin-Watson stat	1.846369			

### 3.4 Testing for Endogeneity

- 3.4.1 The Hausman (1978) test allows us to compare two estimators for one regression model, where one estimator is guaranteed to be consistent and efficient.
- 3.4.2 In the 2SLS case, if the suspected endogenous regressor  $x_i$  is NOT endogenous, then OLS and 2SLS should be approximately identical. Otherwise, in the presence of endogenous regressors OLS is not consistent so OLS and 2SLS must produce significantly different estimates.
- 3.4.3 EViews allows us to perform the Hausman test by a sequence of regressions (Davidson and MacKinnon 1989, 1993):
- i. Regress the suspected endogenous variable (e.g. *health\_exp*) on all exogenous variables and available instruments  $z_i$ . Collect the residuals, say  $w_i$ .
  - ii. In the case of *health\_exp*, regression residuals  $w_i$  represent *health\_exp* after controlling for association with other variables.
  - iii. Now regress  $y_i$  on  $x_i$  as usual, only include  $w_i$  from the first auxiliary regression. If the suspected endogenous variable is truly endogenous then the slope on  $w_i$  will be significant.

#### 4. Example #5: U.S. Morality Rates and 2SLS

4.1 We suspect *health\_exp* is endogenous. Regress *health\_exp* on all other explanatory variables plus the income instrument *inc*. Save the residuals

```
ls health_exp c ed_coll ed_coll_2 phys inc
series w = resid
```

4.2 Regress *mort* on the usual regressors plus *u*:

```
ls mort c ed_coll ed_coll_2 phys health_exp w
```

The results follow:

Dependent Variable: MORT  
Method: Least Squares  
Sample: 1 51  
Included observations: 51

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	850.9452	328.9525	2.586833	0.0130
ED_COLL	-1081.035	2911.842	-0.371255	0.7122
ED_COLL_2	-9559.261	8510.806	-1.123191	0.2673
PHYS	1.976187	0.724357	2.728195	0.0091
HEALTH_EXP	0.043648	0.130936	0.333351	0.7404
W	0.037442	0.144780	0.258615	0.7971
R-squared	0.704911	Mean dependent var		855.0059
Adjusted R-squared	0.672123	S.D. dependent var		137.9660
S.E. of regression	79.00003	Akaike info criterion		11.68690
Sum squared resid	280845.2	Schwarz criterion		11.91418
Log likelihood	-292.0161	F-statistic		21.49926
Durbin-Watson stat	1.843317	Prob(F-statistic)		0.000000

The results support our finding that 2SLS did not generate estimates very different from OLS. Here, the coefficient on *u* is not significant at any level, so we fail to reject the null that *health\_exp* is exogenous.

#### 5. Two Stage Least Squares in a SUR System: Three Stages Least Squares

Three Stages Least Squares is 2SLS applied to a Seemingly Unrelated System. The three steps concern *i.* controlling for correlation between the different equation error terms; *ii.* controlling for endogenous regressors; and *iii.* estimating the robustified system.

5.1 Follow the SUR instructions: **OBJECTS, NEW OBJECT, SYSTEM** (name the system, say *mort\_3sls*).

5.2 In the white pop-up box type the equations as before. Below the last equation type the instrument set. I include all exogenous variables included in the regression and all instruments that were left out:

```
@inst [exogenous regressors] [instruments]
```

There is no “=” and there are no commas.

5.3 Click **ESTIMATE, Three Stages Least Squares**.

## 6. Example #6: U.S. Morality Rates and 3SLS

Recall the system of equations is for mortality rates and tobacco use:

$$mort_i = \beta_0 + \beta_1 ed\_coll_i + \beta_2 ed\_coll_i^2 + \beta_3 phys_i + \beta_4 health + exp_i + \varepsilon_i$$

$$tob_i = \delta_0 + \delta_1 ed\_hs_i + \delta_2 inc_i + \delta_3 aged_i + u_i$$

6.1 In the main toolbar **OBJECTS, NEW OBJECT, SYSTEM** (name *mort\_3sls*).

6.2 Type

**mort = c(1) +c(2)\*ed2\_coll +c(3)\*ed\_coll\_2 +c(4)\*phys +c(5)\*health\_pc**

**tob\_pc = c(6) + c(7)\*ed\_hs + c(8)\*inc\_pc + c(9)\*aged**

**@inst inc\_pc ed2\_coll ed\_coll\_2 phys health\_pc ed\_hs**

Click **ESTIMATE, THREE STAGE LEAST SQUARES**. The results are

System: MORT\_3SLS

Estimation Method: Three-Stage Least Squares

Sample: 1 51

Included observations: 51

Total system (balanced) observations 102

Linear estimation after one-step weighting matrix

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	769.0567	235.4522	3.266297	0.0015
C(2)	-643.7900	2458.311	-0.261883	0.7940
C(3)	-10497.68	7442.833	-1.410442	0.1617
C(4)	1.817089	0.364334	4.987429	0.0000
C(5)	0.081796	0.051370	1.592275	0.1147
C(6)	182.3184	43.50059	4.191172	0.0001
C(7)	-146.6055	44.42703	-3.299916	0.0014
C(8)	0.003231	0.001433	2.255129	0.0265
C(9)	-47.82365	195.9449	-0.244067	0.8077

Determinant residual covariance 2032528.

Equation: MORT = C(1) +C(2)\*ED\_COLL+C(3)\*ED\_COLL\_2+C(4) \*PHYS+C(5)\*HEALTH\_EXP

Instruments: INC\_PC ED2\_COLL ED\_COLL\_2 PHYS HEALTH\_PC ED\_HS C

Observations: 51

R-squared	0.703737	Mean dependent var	855.0059
Adjusted R-squared	0.677975	S.D. dependent var	137.9660
S.E. of regression	78.29189	Sum squared resid	281962.6
Durbin-Watson stat	1.889835		

Equation: TOB\_PC = C(6) + C(7)\*ED\_HS + C(8)\*INC\_PC + C(9)\*AGED

Instruments: INC\_PC ED2\_COLL ED\_COLL\_2 PHYS HEALTH\_PC

ED\_HS C

Observations: 51

R-squared	0.185201	Mean dependent var	120.5275
Adjusted R-squared	0.133193	S.D. dependent var	22.13050
S.E. of regression	20.60404	Sum squared resid	19952.75
Durbin-Watson stat	1.975971		

## VII. Limited Dependent Variables

There are a variety of situations where the dependent variable range of possible values is limited. It may be 0/1-binary (e.g. 1 = is in labor force), it may be an integer (e.g. number of loans outstanding), it may be categorical (e.g. education level *high school*, *college 4 years*, *college 6 years*) and it may be truncated (e.g. work hours > 0 if employed).

In this part we review two model scenarios: Binary Response and Censored Regression.

### 1. Binary Response

In this case  $y_i = 0$  or  $1$ . Typically the approach is to assume  $y_i$  depends on observable  $x_i$  and unobservable  $\varepsilon_i$  traits:

$$(*) \quad y_i = 1 \text{ if } \varepsilon_i > -\sum_{j=1}^k \beta_j x_{i,j} \text{ and } y_i = 0 \text{ if } \varepsilon_i \leq -\sum_{j=1}^k \beta_j x_{i,j}$$

We estimate the coefficients  $\beta$  by binary Maximum Likelihood.

#### 1.1 Binary Maximum Likelihood

1.1.1 We assume the errors  $\varepsilon_j$  are iid with some known cumulative distribution function  $F$ :

$$F(\varepsilon) := P(\varepsilon_i \leq \varepsilon)$$

1.1.2 The Binary **Likelihood Function**  $L(Y|\beta)$  is the joint probability a sample of binary responses  $Y = [y_1, \dots, y_n]$ .

In order to represent the Likelihood Function it helps to re-order the sample as a thought experiment.

**WE DO NOT NEED TO RE-ORDER THE SAMPLE WHEN WE USE EVIEWS.**

This is merely for representing the concept of Binary Maximum Likelihood. We can arbitrarily order the observations so that  $y_i = 0$  occur first in the sample and all  $y_i = 1$  occur last:  $Y = [0, 0, \dots, 0, 1, 1, \dots, 1]$ .

There are  $n_0$  observations with response 0 and  $n_1$  observations with response 1. Note:

$$n_0 + n_1 = n$$

Under independence and using (\*) the natural log of the Likelihood Function is

$$\ln L(y|\beta) = \sum_{i=1}^{n_0} F\left(-\sum_{j=1}^k \beta_j x_{i,j}\right) + \sum_{i=n_0+1}^n \left(1 - F\left(-\sum_{j=1}^k \beta_j x_{i,j}\right)\right)$$

#### 1.2 Marginal Affects in Binary Response Models

The coefficients  $\beta_j$  need to be carefully interpreted. They do NOT represent the marginal impact of  $x_{i,j}$  on  $y_i$ . Rather, notice by the definition of a probability density  $f(x) = (\partial/\partial x)F(x)$ :

$$\frac{\partial}{\partial x_{j,j}} P(y_i = 1) = \frac{\partial}{\partial x_{j,j}} \left(1 - F\left(-\sum_{j=1}^k \beta_j x_{i,j}\right)\right) = -\frac{\partial}{\partial x_{j,j}} F\left(-\sum_{j=1}^k \beta_j x_{i,j}\right) = f\left(-\sum_{j=1}^k \beta_j x_{i,j}\right) \beta_j$$

So,  $\beta_j$  scaled by the density, represents the marginal impact of  $x_{i,j}$  on the likelihood of response  $y_i = 1$ .

- i. Perhaps most importantly, notice the marginal impact IS NOT A CONSTANT. It depends on *each individual's* observable information  $x_{j,i}$ .
- ii. Since it is individual specific, typically we plot out the marginal affects, or analyze the descriptive statistics, including its mean:

$$MEAN\left\{\frac{\partial}{\partial x_{j,i}} P(y_i=1)\right\} = \left(\frac{1}{n} \sum_{i=1}^n f\left(-\sum_{j=1}^k \beta_j x_{i,j}\right)\right) \times \beta_j$$

Alternatively, we can compute the marginal affect for the average individual:

$$\frac{\partial}{\partial x_{j,i}} P(\text{mean}\{y_i\}=1) = f\left(-\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \beta_j x_{i,j}\right) \times \beta_j$$

### 1.3 Estimation in EViews

We can now  $\beta_j$  using EViews. We simply denote what the cdf  $F$  is. The most popular choices in practice are the **standard normal** and the **logistic**.

If we assume  $F$  is the **standard normal** then the estimation method is called **Probit Maximum Likelihood** (i.e. Binary ML with standard normal cdf).

If we assume  $F$  is the **logistic** then the estimation method is called Logit Maximum Likelihood (i.e. Binary ML with logistic cdf).

### 1.4 Probit and Logit ML

1.4.1 In the main toolbar **QUICK, ESTIMATE EQUATION**, scroll through the options for **BINARY CHOICE**, choose **PROBIT** or **LOGIT**.

1.4.2 In the white are type the equation, using the 0/1 variable on the left:

$$y = c \ x1 \ x2 \ x3 \ \dots$$

1.4.3 There are two options: we can select ways to robustify against the fact that we may chosen the wrong  $F$ ; and we may choose the numerical estimation method use for estimating this highly nonlinear model ( $\ln(L)$  is itself very nonlinear).

The true cdf may not be the standard (Probit) or logistic (LOGIT). After all, we are merely guessing.

#### i. **Huber/White**

Under **OPTIONS**, click **ROBUST COVARIANCE MATRIX**, and then **HUBER/WHITE** in order to generate standard errors, and therefore  $t$ -statistics, that are robust to the fact that we may have chosen then wrong cdf  $F$ .

This should be done whenever possible.

#### ii. **GLM Robust Covariance Matrix**

If we make some general assumptions about the true distribution  $F$  then the **GLM** choice for robust covariance matrix is another option.

## 2. Example #7: Binary Choice and Labor Force Participation and Probit

We have a sample of women who in the labor force ( $lfp_i = 1$ ) or not ( $lfp_i = 0$ ). Available regressors are *age*, husband's age *age\_h*, and the number of children under the age of 6 *child\_6*.

### 2.1 The Regression Model

The model is

$$\text{works if } \varepsilon_i > -\beta_0 - \beta_1 \text{age}_i - \beta_2 \text{age\_h}_i - \beta_3 \text{child\_6}_i$$

$$\text{does not work if } \varepsilon_i \leq -\beta_0 - \beta_1 \text{age}_i - \beta_2 \text{age\_h}_i - \beta_3 \text{child\_6}_i$$

We will estimate the model by Probit ML, using Huber-White robust *t*-tests.

2.1.1 In the main toolbar **QUICK, ESTIMATE EQUATION**, scroll through the options for **BINARY CHOICE, PROBIT**.

2.1.2 In the white are type the model

**lfp c age age\_h child\_6**

2.1.3 Now, **OPTIONS, ROBUST COVARIANCE MATRIX, HUBER/WHITE**.

Then, ok twice.

The results follow:

Dependent Variable: LFP  
Method: ML - Binary Logit (Quadratic hill climbing)  
Sample: 1 753  
Included observations: 753  
Convergence achieved after 4 iterations  
QML (Huber/White) standard errors & covariance

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	3.337181	0.533849	6.251172	0.0000
AGE	-0.036233	0.020628	-1.756550	0.0790
AGE_H	-0.026326	0.020776	-1.267107	0.2051
CHILD_6	-1.355352	0.195827	-6.921162	0.0000
Mean dependent var	0.568393	S.D. dependent var		0.495630
S.E. of regression	0.474904	Akaike info criterion		1.289399
Sum squared resid	168.9249	Schwarz criterion		1.313962
Log likelihood	-481.4587	Hannan-Quinn criter.		1.298862
Restr. log likelihood	-514.8732	Avg. log likelihood		-0.639387
LR statistic (3 df)	66.82902	McFadden R-squared		0.064899
Probability(LR stat)	2.03E-14			
Obs with Dep=0	325	Total obs		753
Obs with Dep=1	428			

## 2.2 Marginal Affects

In order to interpret the estimated coefficients, we want to generate the series

$$\frac{\partial}{\partial x_{j,i}} P(y_i = 1) = f\left(-\sum_{j=1}^k \hat{\beta}_j x_{i,j}\right) \hat{\beta}_j$$

Using the estimated values, we will compute

$$\frac{\partial}{\partial x_{j,i}} \hat{P}(y_i = 1) = f\left(-\sum_{j=1}^k \hat{\beta}_j x_{i,j}\right) \hat{\beta}_j$$

EViews does not provide this in a simple way, so we will compute in order

$$\sum_{j=1}^k \hat{\beta}_j x_{i,j} \Rightarrow f\left(-\sum_{j=1}^k \hat{\beta}_j x_{i,j}\right) \Rightarrow f\left(-\sum_{j=1}^k \hat{\beta}_j x_{i,j}\right) \hat{\beta}_j$$

### 2.2.1 We obtain

$$\sum_{j=1}^k \hat{\beta}_j x_{i,j}$$

by clicking within the equation popup box **FORECAST, INDEX-WHERE PROB-F(-INDEX)**. Then ok.

Since the 0/1 dependent variable is called *lfp* the forecast value will given the automatic name *lfpf*, or change the name.

### 2.2.2 Now use *lfpf* to generate

$$f\left(-\sum_{j=1}^k \hat{\beta}_j x_{i,j}\right)$$

In the main white-area type

**series f\_xb = @dnorm(-lfpf)**

Then enter. The function **dnorm** represents the *standard normal density*.

### 2.2.3 In our case the mean of **f\_xb** is .325381. So,

$$MEAN\left\{\frac{\partial}{\partial x_{j,i}} \hat{P}(y_i = 1)\right\} = .325381 \times \hat{\beta}_j$$

We can now inspect the marginal impact of each explanatory variable on the likelihood of entering the labor force.

### 3. Censored Regression Models: The Tobit Model

The female labor force participation data set contains information on work hours and wages. For each person that is not in the labor force annual hours  $h = 0$  and wage  $w = 0$  (of course). But that is because they do not work nor receive a wage. It is *not* that they *have a job* and work  $h = 0$  hours per week and get paid  $w = 0$ /hour.

There are two ways to think about this. First, we may discard people not working and use only those with  $h > 0$  and  $w > 0$  to generate labor supply curve. The people in the sample who work are the ones whose information is used to generate a supply curve. This neglects all the individuals who do not work: labor supply and therefore the relationship between work hours and wages, is influenced by those not working (their non-presence helps to dampen wages) as much as by those who are working. If we neglect this fact then our labor supply coefficient estimates will be biased. This is **Sample Selection Bias**.

The second way to think about this is to allow all individuals to stay in the sample. It may be that some people would choose to work  $h < 0$  (have someone do their job for them) and would love to receive  $w > 0$  (get paid for less than nothing!). We cannot observe this because it is unlikely that such a lazy person would find someone else willing to complete this odd relationship (I do your work, and I pay you to do it!). Thus, within any labor supply sample wherever we see  $h = 0$  we must assume that the person would actually choose, if they could,  $h^* \leq 0$ . This is data censoring and such models are **Censored Regression Models**.

#### 3.1 Censored Regression Model

The problem with a regression models with a censored dependent variable is the model the model does not account for what individuals would prefer, rather than what they have. There is, ultimately, a missing variable.

We must differentiate between the chosen  $y^*$  and the observed  $y$ . For the sake of simplicity we assume truncation occurs at zero. The censored regression model is

$$\begin{aligned} \text{Unobserved/Chosen } y_i^* &= \sum_{j=1}^k \beta_j x_{i,j} + \varepsilon_i \\ \text{Observed } y_i &= y_i^* \text{ if } y_i^* \geq 0 \\ &= 0 \text{ if } y_i^* < 0 \end{aligned}$$

Since we do not observe  $y^*$  (e.g. work hours  $h < 0$ !), we must, of course, use the observed  $y$  (e.g.  $h = 0$ ):

$$y_i = \sum_{j=1}^k \beta_j x_{i,j} + \varepsilon_i$$

But it can be shown that OLS estimates will be biased because there is a missing variable accounting for the truncation ( $y^* < y$ ).

#### 3.2 Tobit Model

Since we do not observe  $y^*$  (e.g. work hours  $h < 0$ !), we must, of course, use the observed  $y$  (e.g.  $h = 0$ ):

$$y_i = \sum_{j=1}^k \beta_j x_{i,j} + \varepsilon_i$$

But it can be shown that OLS estimates will be biased because there is a missing variable accounting for the truncation ( $y^* < y$ ). If the errors are iid normally distributed  $N(0, \sigma^2)$  then the correct model is

$$y_i = \sum_{j=1}^k \beta_j x_{i,j} + \sigma \frac{\phi\left(\sum_{j=1}^k \beta_j x_{i,j}\right)}{\Phi\left(\sum_{j=1}^k \beta_j x_{i,j}\right)} + \varepsilon_i$$

where  $\phi(z)$  is the standard normal density and  $\Phi(z)$  the standard normal cdf. This is called the **Tobit Regression Model**, after Tobin (1958).

### 3.3 Estimating the Tobit Model

EViews offers a Tobit routine. In the main toolbar **QUICK, ESTIMATE EQUATION**, type the equation, scroll and choose **CENSORED – TOBIT**.

Choose the way the dependent variable is censored via **LEFT** and **RIGHT**. In the annual work hour case hours are truncated at zero and 8736 (168 hours/week times 52 weeks). Leave either space blank if there is no censoring.

Next, **OPTIONS, ROBUST COVARIANCES, HUBER-WHITE**.

Then ok twice.

## 4. Example #8: Female Work Hours and Tobit Estimation

We want to estimate the following annual work hour model:

$$hours_i = \beta_0 + \beta_1 wage_i + \beta_2 age_i + \beta_3 hours\_h_i + \beta_4 wage\_h_i + \beta_5 age\_h_i + \beta_6 child\_6_i + \varepsilon_i$$

where  $hours\_h$  is the female's husband's work hours, etc., and  $child\_6$  the number of children under the age of 6 in the family.

OLS results follow:

Dependent Variable: HOURS

Method: Least Squares

Sample: 1 753

Included observations: 753

White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1612.222	259.0128	6.224486	0.0000
WAGE	106.1719	19.48620	5.448570	0.0000
AGE	-5.532008	6.992067	-0.791184	0.4291
HOURS_H	-0.101367	0.048597	-2.085887	0.0373
WAGE_H	-26.57645	5.395773	-4.925421	0.0000
AGE_H	-8.230411	6.727591	-1.223382	0.2216
CHILD_6	-371.8635	62.04534	-5.993416	0.0000
R-squared	0.237422	Mean dependent var		740.5764
Adjusted R-squared	0.231289	S.D. dependent var		871.3142
S.E. of regression	763.9350	Akaike info criterion		16.12410
Sum squared resid	4.35E+08	Schwarz criterion		16.16708
Log likelihood	-6063.722	F-statistic		38.71011
Durbin-Watson stat	1.606736	Prob(F-statistic)		0.000000

Next, Tobit results:

Dependent Variable: HOURS

Method: ML - Censored Normal (TOBIT) (Quadratic hill climbing)

Sample: 1 753

Included observations: 753

Left censoring (value) series: 0

Right censoring (value) series: 8736

Convergence achieved after 6 iterations

QML (Huber/White) standard errors & covariance

	Coefficient	Std. Error	z-Statistic	Prob.
C	2157.747	413.1214	5.223035	0.0000
WAGE	204.3639	30.26070	6.753443	0.0000
AGE	-12.21144	12.22641	-0.998775	0.3179
HOURS_H	-0.220945	0.080955	-2.729226	0.0063
WAGE_H	-59.52449	12.00630	-4.957772	0.0000
AGE_H	-15.81633	11.65812	-1.356680	0.1749
CHILD_6	-825.4865	130.1804	-6.341098	0.0000
Error Distribution				
SCALE:C(8)	1146.301	51.03048	22.46306	0.0000
R-squared	0.106779	Mean dependent var		740.5764
Adjusted R-squared	0.098387	S.D. dependent var		871.3142
S.E. of regression	827.3419	Akaike info criterion		10.13556
Sum squared resid	5.10E+08	Schwarz criterion		10.18468
Log likelihood	-3808.037	Hannan-Quinn criter.		10.15448
Avg. log likelihood	-5.057154			
Left censored obs	325	Right censored obs		0
Uncensored obs	428	Total obs		753

Since no one works all hours on all days, right censoring is irrelevant: we can leave **RIGHT** blank and receive the same results.

Notice the stark coefficient estimate differences. By not accounting for censorship all marginal affects are underestimated. By not controlling for the numerous  $hours = wages = 0$ , least squares under estimates the marginal affect a one dollar differential has on annual work hours by a factor of two! Similarly, the presence of young children is overwhelming associated with dampened work hours, but that effect is far stronger once truncation is controlled for.