

Functional Form

So far considered models written in *linear* form

$$Y = b_0 + b_1X + u \quad (1)$$

Implies a straight line relationship between y and X

Functional Form

So far considered models written in *linear* form

$$Y = b_0 + b_1X + u \quad (1)$$

Implies a straight line relationship between y and X

Sometimes economic theory and/or observation of data will not suggest that there is a linear relationship between variables

Functional Form

So far considered models written in *linear* form

$$Y = b_0 + b_1X + u \quad (1)$$

Implies a straight line relationship between y and X

Sometimes economic theory and/or observation of data will not suggest that there is a linear relationship between variables

One way to model a non-linear relationship is the equation

Functional Form

So far considered models written in *linear* form

$$Y = b_0 + b_1X + u \quad (1)$$

Implies a straight line relationship between y and X

Sometimes economic theory and/or observation of data will not suggest that there is a linear relationship between variables

One way to model a non-linear relationship is the equation

$$Y = a + b/X + e \quad (2)$$

Functional Form

So far considered models written in *linear* form

$$Y = b_0 + b_1X + u \quad (1)$$

Implies a straight line relationship between y and X

Sometimes economic theory and/or observation of data will not suggest that there is a linear relationship between variables

One way to model a non-linear relationship is the equation

$$Y = a + b/X + e \quad (2)$$

(where the line asymptotes to the value “ a ” as $X \uparrow$ - from below if $b < 0$, from above if $b > 0$)

Functional Form

So far considered models written in *linear* form

$$Y = b_0 + b_1X + u \quad (1)$$

Implies a straight line relationship between y and X

Sometimes economic theory and/or observation of data will not suggest that there is a linear relationship between variables

One way to model a non-linear relationship is the equation

$$Y = a + b/X + e \quad (2)$$

(where the line asymptotes to the value “ a ” as $X \uparrow$ - from below if $b < 0$, from above if $b > 0$)

However this is not a linear equation, unlike (1), since it does not trace out a straight line between Y and X and OLS only works (ie minimise RSS) if can somehow make (2) linear.

The solution is to use algebra to transform equations like (2) so appear like (1)

The solution is to use algebra to transform equations like (2) so appear like (1)

In the above example do this by creating a variable equal to the reciprocal of X , $1/X$, so that the relationship between y and $1/X$ is linear (ie a straight line)

So that
$$Y = a + b/X + e \quad (2)$$

The solution is to use algebra to transform equations like (2) so appear like (1)

In the above example do this by creating a variable equal to the reciprocal of X , $1/X$, so that the relationship between y and $1/X$ is linear (ie a straight line)

So that
$$Y = a + b/X + e \quad (2)$$

becomes

$$Y = a + b*(1/X) + e \quad (3)$$

(3) is now "*linear in parameters*"

ie regress Y on $1/X$ rather than Y on X

The solution is to use algebra to transform equations like (2) so appear like (1)

In the above example do this by creating a variable equal to the reciprocal of X , $1/X$, so that the relationship between y and $1/X$ is linear (ie a straight line)

So that
$$Y = a + b/X + e \quad (2)$$

becomes

$$Y = a + b*(1/X) + e \quad (3)$$

(3) is now "*linear in parameters*"

ie regress Y on $1/X$ rather than Y on X

The only thing now need to be careful about is how to interpret the coefficients from this specification

$$dY/d((1/X)) = b \quad \text{but} \quad dY/dX = -b/X^2$$

Log Linear Models

A much more common and useful transformation is based on the realisation that some economic relationships can be modelled as

$$Y = b_0 X^{b_1} \exp(u)$$

Log Linear Models

A much more common and useful transformation is based on the realisation that some economic relationships can be modelled as

$$Y = b_0 X^{b_1} \exp(u)$$

Can't estimate this by OLS since this is not a straight line

Log Linear Models

A much more common and useful transformation is based on the realisation that some economic relationships can be modelled as

$$Y = b_0 X^{b_1} \exp(u)$$

Can't estimate this by OLS since this is not a straight line

To make this model linear in parameters take (natural) logs so that

$$\ln Y = \ln b_0 + b_1 \ln X + u \quad (4)$$

Log Linear Models

A much more common and useful transformation is based on the realisation that some economic relationships can be modelled as

$$Y = b_0 X^{b_1} \exp(u)$$

Can't estimate this by OLS since this is not a straight line

To make this model linear in parameters take (natural) logs so that

$$\ln Y = \ln b_0 + b_1 \ln X + u \quad (4)$$

(Note can only do this if Y & $X > 0$ since log of a negative number does not exist)

Log Linear Models

A much more common and useful transformation is based on the realisation that some economic relationships can be modelled as

$$Y = b_0 X^{b_1} \exp(u)$$

Can't estimate this by OLS since this is not a straight line

To make this model linear in parameters take (natural) logs so that

$$\ln Y = \ln b_0 + b_1 \ln X + u \quad (4)$$

(Note can only do this if Y & $X > 0$ since log of a negative number does not exist)

which traces out a straight line between $\ln Y$ (not Y) and $\ln X$ (not X) and so can be estimated by OLS

This is a useful specification because the estimated coefficients can be interpreted as *elasticities*

This is a useful specification because the estimated coefficients can be interpreted as *elasticities*

Since $d\ln Y/dY = 1/Y$

This is a useful specification because the estimated coefficients can be interpreted as *elasticities*

Since $d\ln Y/dY = 1/Y$ then $d\ln Y = dY/Y$

This is a useful specification because the estimated coefficients can be interpreted as *elasticities*

Since $d\ln Y/dY = 1/Y$ then $d\ln Y = dY/Y$

which is the % change in $y \div 100$

This is a useful specification because the estimated coefficients can be interpreted as *elasticities*

Since $d\ln Y/dY = 1/Y$ then $d\ln Y = dY/Y$

which is the % change in $y \div 100$

Similarly

$d\ln X = dX/X$ is the % change in $X \div 100$

This is a useful specification because the estimated coefficients can be interpreted as *elasticities*

Since $d\ln Y/dY = 1/Y$ then $d\ln Y = dY/Y$

which is the % change in $y \div 100$

Similarly

$d\ln X = dX/X$ is the % change in $X \div 100$

From (4) $\ln Y = \ln b_0 + b_1 \ln X + u$

So $d\ln Y/d\ln X = b_1$

This is a useful specification because the estimated coefficients can be interpreted as *elasticities*

Since $d\ln Y/dY = 1/Y$ then $d\ln Y = dY/Y$

which is the % change in $y \div 100$

Similarly

$d\ln X = dX/X$ is the % change in $X \div 100$

From (4) $\ln Y = \ln b_0 + b_1 \ln X + u$

So $d\ln Y/d\ln X = b_1$

Sub. in from above

$$(dY/Y)/(dX/X) = b_1$$

This is a useful specification because the estimated coefficients can be interpreted as *elasticities*

Since $d\ln Y/dY = 1/Y$ then $d\ln Y = dY/Y$

which is the % change in $y \div 100$

Similarly

$d\ln X = dX/X$ is the % change in $X \div 100$

From (4) $\ln Y = \ln b_0 + b_1 \ln X + u$

So $d\ln Y/d\ln X = b_1$

Sub. in from above

$$(dY/Y)/(dX/X) = b_1$$

so $b_1 = \% \Delta \text{ in } Y \div \% \Delta \text{ in } X$

This is a useful specification because the estimated coefficients can be interpreted as *elasticities*

Since $d\ln Y/dY = 1/Y$ then $d\ln Y = dY/Y$

which is the % change in $y \div 100$

Similarly

$d\ln X = dX/X$ is the % change in $X \div 100$

From (4) $\ln Y = \ln b_0 + b_1 \ln X + u$

So $d\ln Y/d\ln X = b_1$

Sub. in from above

$$(dY/Y)/(dX/X) = b_1$$

so $b_1 = \% \Delta \text{ in } Y \div \% \Delta \text{ in } X$

= elasticity of y wrt X

Semi-Log Models

Another common functional form is the semi-log model (log-lin model) in which the dependent variable is measured in logs and the X variables in levels

$$y = \beta_0 \exp^{\beta_1 X}$$

Semi-Log Models

Another common functional form is the semi-log model (log-lin model) in which the dependent variable is measured in logs and the X variables in levels

$$y = \beta_0 \exp^{\beta_1 X}$$

Taking (natural) logs gives

$$\text{Log}_e Y = \text{Log}_e \beta_0 + \beta_1 X \text{log}_e(\exp)$$

Semi-Log Models

Another common functional form is the semi-log model (log-lin model) in which the dependent variable is measured in logs and the X variables in levels

$$y = \beta_0 \exp^{\beta_1 X}$$

Taking (natural) logs gives

$$\text{Log}_e Y = \text{Log}_e \beta_0 + \beta_1 X \log_e(\exp)$$

which since $\log_e(\exp) = 1$ gives

$$\text{Log}_e Y = \text{Log}_e \beta_0 + \beta_1 X$$

Semi-Log Models

Another common functional form is the semi-log model (log-lin model) in which the dependent variable is measured in logs and the X variables in levels. This typically applies to economic variables which are related exponentially

$$y = \beta_0 \exp^{\beta_1 X}$$

Taking (natural) logs gives

$$\text{Log}_e Y = \text{Log}_e \beta_0 + \beta_1 X \log_e(\exp)$$

which since $\log_e(\exp) = 1$ gives

$$\text{Log}_e Y = \text{Log}_e \beta_0 + \beta_1 X$$

The interpretation of the estimated coefficient β_1 is then

$$\frac{d\text{Log}(y)}{dX} = \beta_1 = \frac{dy/Y}{X}$$

$$\frac{d\text{Log}(y)}{dX} = \beta_1 = \frac{dy/Y}{X}$$

so $\beta_1 = \% \text{ change in } y / 100 \text{ w.r.t. unit change in } X$

$$\frac{d\text{Log}(y)}{dX} = \beta_1 = \frac{dy/Y}{X}$$

so $\beta_1 = \% \text{ change in } y / 100 \text{ w.r.t. unit change in } X$

and $\% \text{ change in } y = \beta_1 * dX * 100$

$$\frac{d\text{Log}(y)}{dX} = \beta_1 = \frac{dy/Y}{X}$$

so $\beta_1 = \% \text{ change in } y / 100 \text{ w.r.t. unit change in } X$

and $\% \text{ change in } y = \beta_1 * dX * 100$

This is called a *semi-elasticity*

$$\frac{d\text{Log}(y)}{dX} = \beta_1 = \frac{dy/Y}{X}$$

so $\beta_1 = \% \text{ change in } y / 100 \text{ w.r.t. unit change in } X$

and $\% \text{ change in } y = \beta_1 * dX * 100$

This is called a *semi-elasticity*

Commonly used in estimation of wage determination and growth

$$\frac{d\text{Log}(y)}{dX} = \beta_1 = \frac{dy/Y}{X}$$

so $\beta_1 = \% \text{ change in } y / 100 \text{ w.r.t. unit change in } X$

$$\text{and } \% \text{ change in } y = \beta_1 * dX * 100$$

This is called a *semi-elasticity*

Commonly used in estimation of wage determination and growth

Eg if wage and age are related by

$$\hat{\text{Log}}(\text{wage}) = \beta_0 + \beta_1 \text{Age}$$

$$\frac{d\text{Log}(y)}{dX} = \beta_1 = \frac{dy/Y}{X}$$

so $\beta_1 = \% \text{ change in } y / 100 \text{ w.r.t. unit change in } X$

$$\text{and } \% \text{ change in } y = \beta_1 * dX * 100$$

This is called a *semi-elasticity*

Commonly used in estimation of wage determination and growth

Eg if wage and age are related by

$$\hat{\text{Log}}(\text{wage}) = \beta_0 + \beta_1 \text{Age}$$

β_1 gives the % change in wages following an increase in age by 1 year

$$\frac{d\text{Log}(y)}{dX} = \beta_1 = \frac{dy/Y}{X}$$

so $\beta_1 = \% \text{ change in } y / 100 \text{ w.r.t. unit change in } X$

$$\text{and } \% \text{ change in } y = \beta_1 * dX * 100$$

This is called a *semi-elasticity*

Commonly used in estimation of wage determination and growth

Eg if wage and age are related by

$$\hat{\text{Log}}(\text{wage}) = \beta_0 + \beta_1 \text{Age}$$

β_1 gives the % change in wages following an increase in age by 1 year

Or if

$$\text{Log}(\text{GDP}) = a + b\text{Year} + u$$

b gives the annual % growth rate in GDP:

$$\frac{d\text{Log}(y)}{dX} = \beta_1 = \frac{dy/Y}{X}$$

so $\beta_1 = \% \text{ change in } y / 100 \text{ w.r.t. unit change in } X$

$$\text{and } \% \text{ change in } y = \beta_1 * dX * 100$$

This is called a *semi-elasticity*

Commonly used in estimation of wage determination and growth

Eg if wage and age are related by

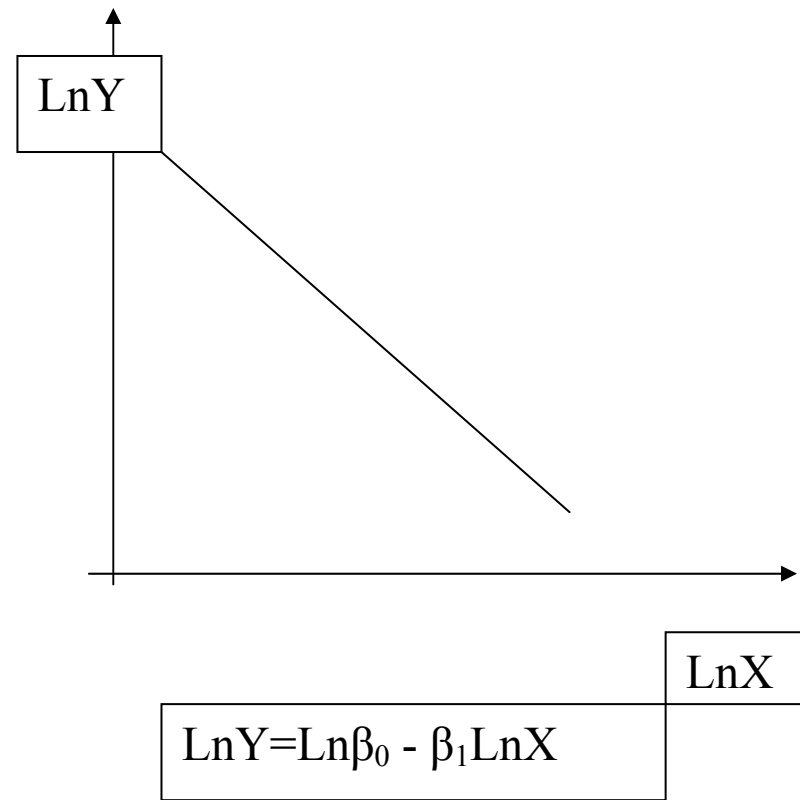
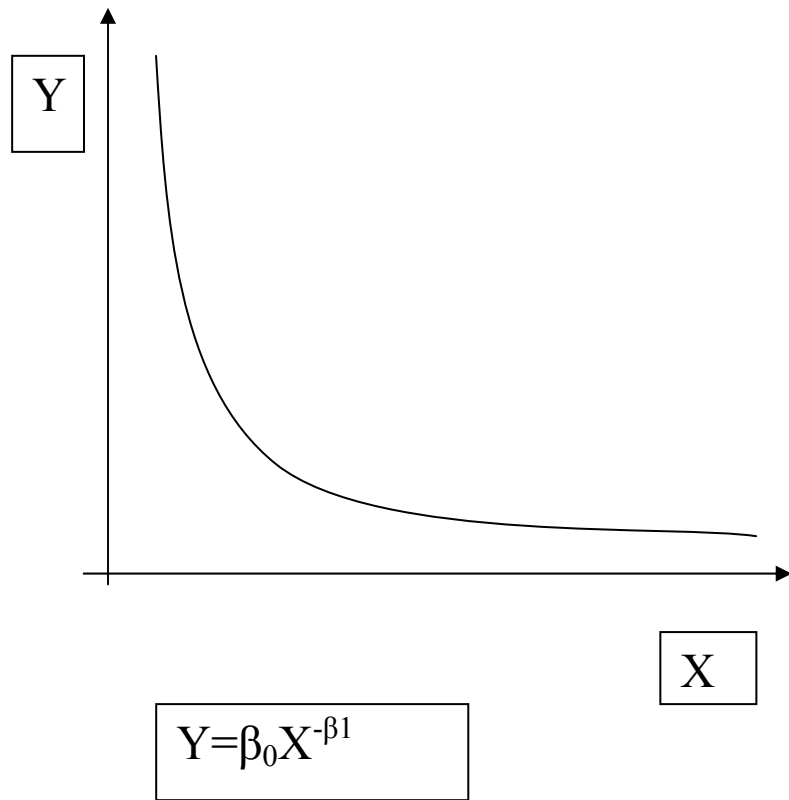
$$\hat{\text{Log}}(\text{wage}) = \beta_0 + \beta_1 \text{Age}$$

β_1 gives the % change in wages following an increase in age by 1 year

Or if

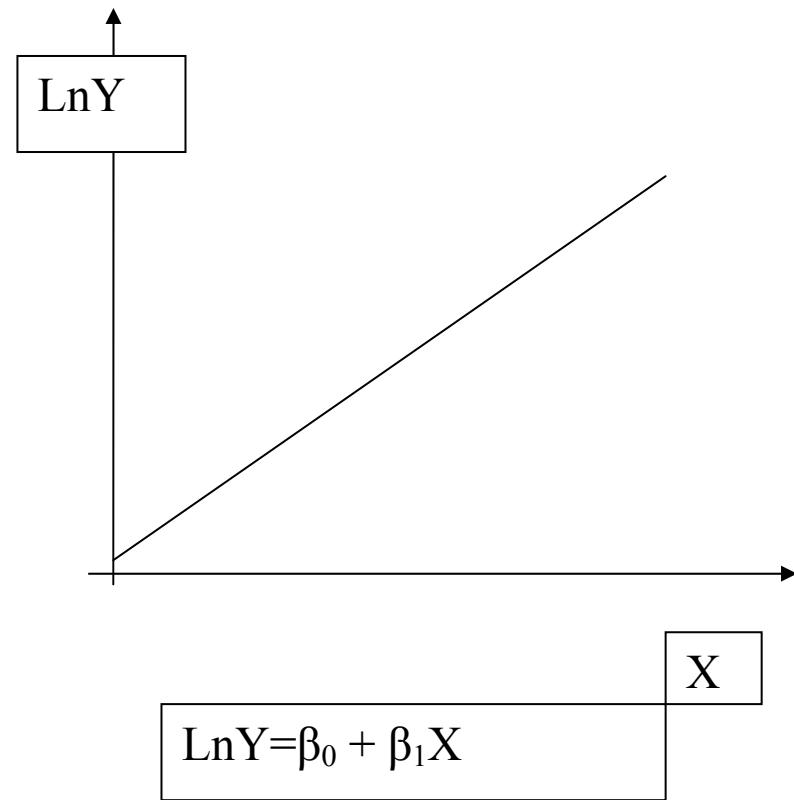
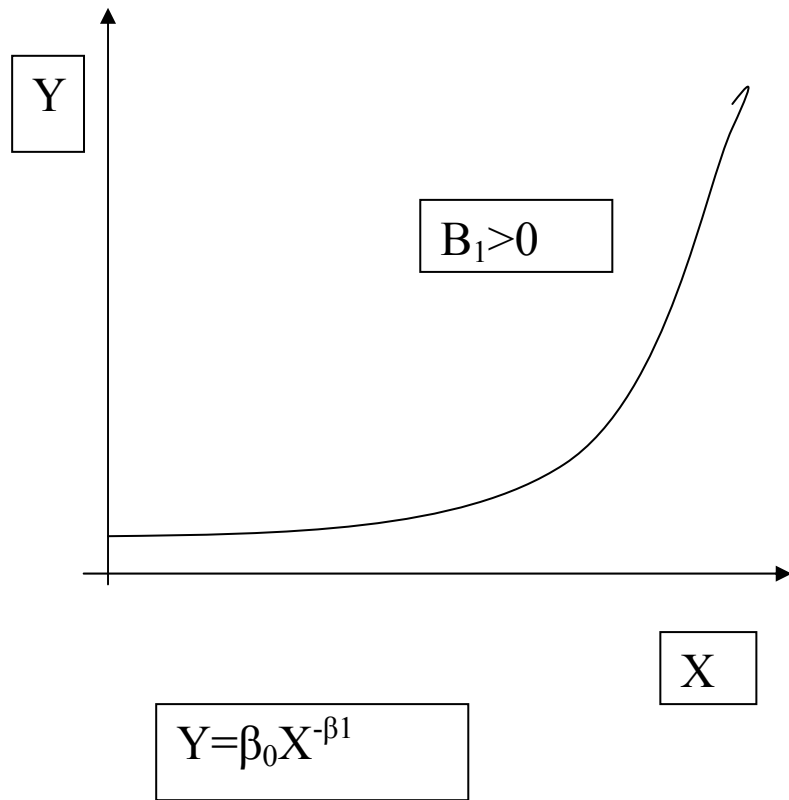
$$\text{Log}(\text{GDP}) = a + b\text{Year} + u$$

b gives the annual % growth rate in GDP:



Logarithmic transformations can turn non-linear relationships into linear models that can be estimated by OLS

The estimated coefficients can then be interpreted as elasticities (as in the log-log specification above)

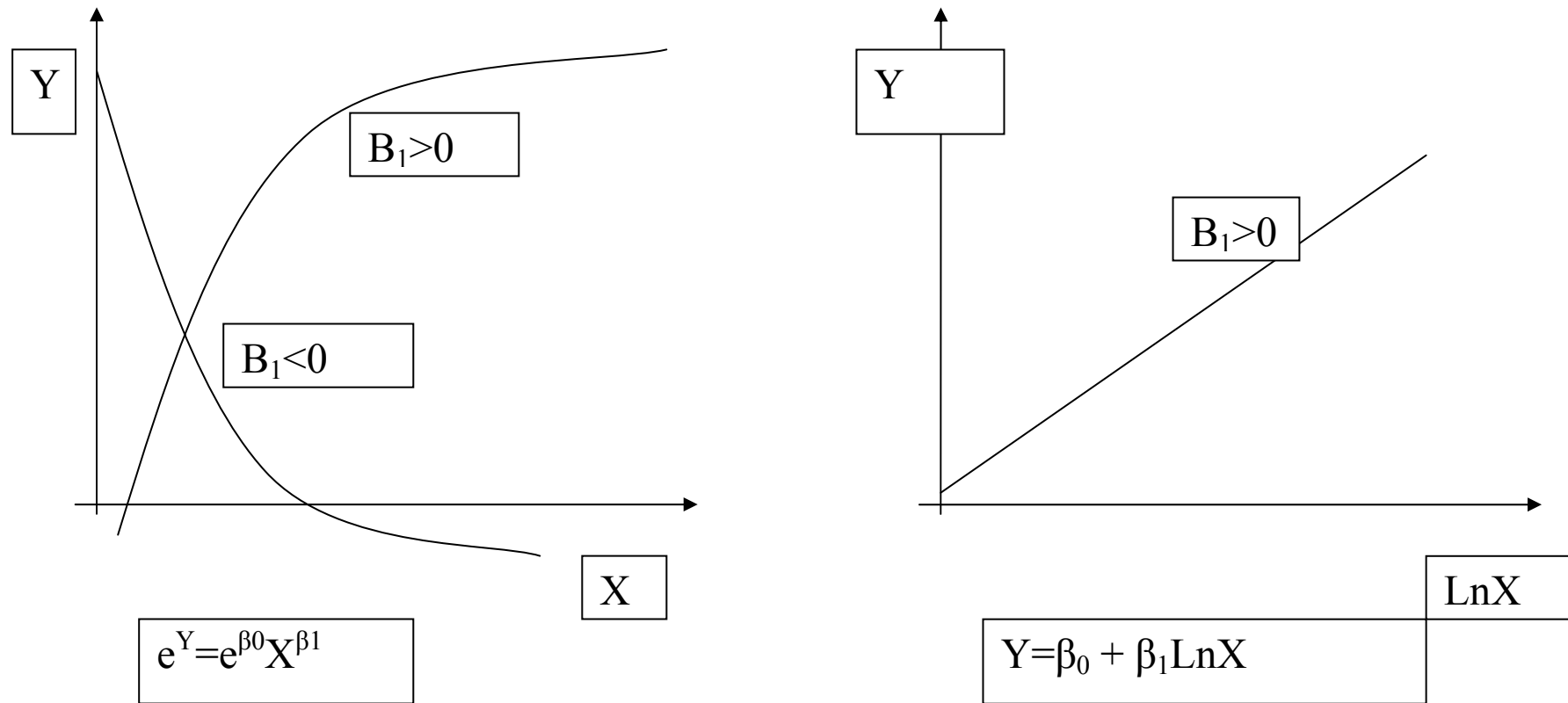


Or as growth rates in the case of the log-lin specification here
(useful for estimating growth rates or rates of return)

It is also possible to have a lin-log model of the form:

$$Y = \beta_0 + \beta_1 \text{Log}_e X$$

(based on the model $e^y = e^{\beta_0} X^{\beta_1}$)



(Eg model food expenditure as a function of log income)

It is also possible to have a lin-log model of the form:

$$Y = \beta_0 + \beta_1 \text{Log}_e X$$

(based on the model $e^y = e^{\beta_0} X^{\beta_1}$)

and now $\frac{dy}{d\text{Log}_e X} = \beta_1 = \frac{Y}{dX / X}$

It is also possible to have a lin-log model of the form:

$$Y = \beta_0 + \beta_1 \text{Log}_e X$$

(based on the model $e^y = e^{\beta_0} X^{\beta_1}$)

$$Y = \beta_0 + \beta_1 \text{Log}_e X$$

and now $\frac{dy}{d\text{Log}_e X} = \beta_1$

$$Y = \beta_0 + \beta_1 \text{Log}_e X$$

and now $\frac{dy}{d\text{Log}_e X} = \beta_1 = \frac{dY}{dX / X}$

$$Y = \beta_0 + \beta_1 \text{Log}_e X$$

and now $\frac{dy}{d\text{Log}_e X} = \beta_1 = \frac{dY}{dX / X}$

so $\beta_1 =$ unit change in y divided by % change in $X / 100$

$$Y = \beta_0 + \beta_1 \text{Log}_e X$$

(based on the model $e^y = e^{\beta_0} X^{\beta_1}$)

and now $\frac{dy}{d\text{Log}_e X} = \beta_1 = \frac{dY}{dX / X}$

so β_1 = unit change in y divided by % change in $X / 100$

and

$$\text{change in } y = \beta_1 * \% \text{ change in } X / 100$$

$$Y = \beta_0 + \beta_1 \text{Log}_e X$$

(based on the model $e^y = e^{\beta_0} X^{\beta_1}$)

and now $\frac{dy}{d\text{Log}_e X} = \beta_1 = \frac{dY}{dX / X}$

so β_1 = unit change in y divided by % change in $X / 100$

and

$$\text{change in } y = \beta_1 * \% \text{ change in } X / 100$$

so $\beta_1/100$ is the unit change in y when X increases by 1%

(this form could be used for the Engle Curves outlined earlier)

Testing Functional Form

How do you know whether to use logs or levels for the dependent variable?

If want to compare goodness of fit of models in which the dependent variable is in logs or levels then can not use the R^2 .

Testing Functional Form

How do you know whether to use logs or levels for the dependent variable?

If want to compare goodness of fit of models in which the dependent variable is in logs or levels then can not use the R^2 .

The TSS in Y is not the same as the TSS in $\text{Ln}Y$,

$$\sum_{i=1}^N (Y_i - \bar{Y})^2 \neq \sum_{i=1}^N (\text{Ln}Y_i - \text{Ln}\bar{Y})^2$$

so comparing R^2 is not valid.

Testing Functional Form

How do you know whether to use logs or levels for the dependent variable?

If want to compare goodness of fit of models in which the dependent variable is in logs or levels then can not use the R^2 .

The TSS in Y is not the same as the TSS in $\text{Ln}Y$,

$$\sum_{i=1}^N (Y_i - \bar{Y})^2 \neq \sum_{i=1}^N (\text{Ln}Y_i - \text{Ln}\bar{Y})^2$$

so comparing R^2 is not valid.

Instead the basic idea behind testing for the appropriate functional form of the *dependent* variable is to transform the data so as to make the **RSS** comparable

Do this by :

Do this by :

1. Calculating the geometric mean

where geometric (rather than arithmetic) mean

$$= (y_1 * y_2 * \dots * y_n)^{1/n} = \exp^{1/n \text{Ln}(y_1 * y_2 \dots y_n)}$$

Do this by :

1. Calculating the geometric mean

where geometric (rather than arithmetic) mean

$$= (y_1 * y_2 * \dots * y_n)^{1/n} = \exp^{1/n \text{Ln}(y_1 * y_2 \dots y_n)}$$

2. rescale each y observation by dividing by this value

$$y_i^* = y_i / \text{geometric mean}$$

Do this by :

1. Calculating the geometric mean

where geometric (rather than arithmetic) mean

$$= (y_1 * y_2 * \dots * y_n)^{1/n} = \exp^{1/n \text{Ln}(y_1 * y_2 * \dots * y_n)}$$

2. rescale each y observation by dividing by this value

$$y_i^* = y_i / \text{geometric mean}$$

3. regress y^* (rather than y) on X, save RSS
regress $\text{Ln}y^*$ (rather than $\text{Ln}y$) on X, save RSS

Do this by :

1. Calculate the geometric mean

where geometric (rather than arithmetic) mean

$$= (y_1 * y_2 * \dots * y_n)^{1/n} = \exp^{1/n \text{Ln}(y_1 * y_2 * \dots * y_n)}$$

2. rescale each y observation by dividing by this value

$$y_i^* = y_i / \text{geometric mean}$$

3. regress y^* (rather than y) on X, save RSS

regress $\text{Ln}y^*$ (rather than $\text{Ln}y$) on X, save RSS

(in practice the RSS is the same whether you use $\text{Ln}Y$ or $\text{Ln}y^*$)

the model with the lowest RSS is the one with the better fit

More formally can show that

$$\text{BoxCox} = N/2 * \log(\text{RSS}_{\text{largest}} / \text{RSS}_{\text{smallest}}) \sim \chi^2_{(1)}$$

More formally can show that

$$\text{BoxCox} = N/2 * \log(\text{RSS}_{\text{largest}} / \text{RSS}_{\text{smallest}}) \sim \chi^2_{(1)}$$

Follows a Chi-Squared distribution with one degree of freedom (one because there is one statistic being tested)

More formally can show that

$$\text{BoxCox} = N/2 * \log(\text{RSS}_{\text{largest}} / \text{RSS}_{\text{smallest}}) \sim \chi^2_{(1)}$$

Follows a Chi-Squared distribution with one degree of freedom (one because there is one statistic being tested)

If estimated value exceeds critical value (from tables Chi-squared at 5% level with 1 degree of freedom is 3.84) *reject* the null hypothesis that the models are the same

(ie there is a significantly different in terms of goodness of fit).

More formally can show that

$$\text{BoxCox} = N/2 * \log(\text{RSS}_{\text{largest}} / \text{RSS}_{\text{smallest}}) \sim \chi^2_{(1)}$$

Follows a Chi-Squared distribution with one degree of freedom (one because there is one statistic being tested)

If estimated value exceeds critical value (from tables Chi-squared at 5% level with 1 degree of freedom is 3.84) *reject* the null hypothesis that the models are the same

(ie there is a significantly different in terms of goodness of fit).

So choose the one with the ***lowest*** RSS

More formally can show that

$$\text{BoxCox} = N/2 * \log(\text{RSS}_{\text{largest}} / \text{RSS}_{\text{smallest}}) \sim \chi^2_{(1)}$$

Follows a Chi-Squared distribution with one degree of freedom (one because there is one statistic being tested)

If estimated value exceeds critical value (from tables Chi-squared at 5% level with 1 degree of freedom is 3.84) *reject* the null hypothesis that the models are the same

(ie there is a significantly different in terms of goodness of fit).

So choose the one with the ***lowest*** RSS

But do **not** use the transformed model to look at the coefficients of the model – use the originals

```
. u boxcox /* read in data */
```

The data contains info on GDP and employment growth for 21 countries

```
. su empl gdp
Variable | Obs Mean Std. Dev. Min Max
-----+-----
empl | 21 1.108095 .8418647 .02 3.02
gdp | 21 3.059524 1.625172 1.15 7.73
```

The data show that gdp and employment growth are measured in percentage points, with a maximum of 7.73 %point annual GDP growth and a minimum 1.15% points.

A linear regression gives

```
. reg empl gdp
Source | SS df MS Number of obs = 21
-----+----- F( 1, 19) = 26.97
Model | 8.31618159 1 8.31618159 Prob > F = 0.0001
Residual | 5.85854191 19 .308344311 R-squared = 0.5867
-----+----- Adj R-squared = 0.5649
Total | 14.1747235 20 .708736175 Root MSE = .55529

empl | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-----+-----
gdp | .396778 .0764018 5.193 0.000 .2368672 .5566888
_cons | -.1058566 .2632937 -0.402 0.692 -.6569367 .4452235
```

Gdp is measured in percentage points, $dempl/dgdp = \beta_{gdp}$
and hence $dempl = \beta_{gdp} * dgdp$ so a **1 % point** rise in gdp growth raises employment growth by 0.4 points a year

and a log-lin specification gives

```
g lempl=log(empl) /* generate log of dep. Variable */
```

```
. reg lempl gdp
Source | SS df MS Number of obs = 21
-----+----- F( 1, 19) = 5.89
Model | 6.84252682 1 6.84252682 Prob > F = 0.0253
Residual | 22.0706507 19 1.1616132 R-squared = 0.2367
-----+----- Adj R-squared = 0.1965
Total | 28.9131775 20 1.44565888 Root MSE = 1.0778

lempl | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-----+-----
gdp | .35991 .1482915 2.427 0.025 .0495322 .6702877
_cons | -1.436343 .5110381 -2.811 0.011 -2.505958 -.3667282
```

log-lin model so coefficients are growth rates. This time $dlemp/dgdp = \beta_{gdp}$
and hence $dlemp = \beta_{gdp} * dgdp$ where $dlemp = \% \text{ change in gdp}/100$.
So a **1% point (not a 1 %)** rise in gdp growth raises emp growth by 36% a year
(from table of means above, can see a 35% increase in gdp amounts to around 0.36 percentage points of extra growth a year - which is similar to estimate in levels)

Looks like linear specification is preferred, but since R^2 or RSS not comparable use Box-Cox test to test formally

Get geometric mean

```
. means empl
```

Variable	Type	Obs	Mean	[95% Conf. Interval]	
empl	Arithmetic	21	1.108095	.724883	1.491307
	Geometric	21	.7152021	.413749	1.236291

Rescale linear dependent variable and log of dependent variable

```
. g empadj=empl/.715
. g lempadj=log(empadj)
```

Regress adjusted dependent variables on gdp and log(gdp) respectively

```
. reg empadj gdp
```

Source	SS	df	MS	Number of obs = 21	
Model	16.2671653	1	16.2671653	F(1, 19) = 26.97	Prob > F = 0.0001
Residual	11.4598119	19	.603147995	R-squared = 0.5867	Adj R-squared = 0.5649
Total	27.7269772	20	1.38634886	Root MSE = .77663	

empadj	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gdp	.5549343	.1068557	5.193	0.000	.3312828	.7785858
_cons	-.1480511	.368243	-0.402	0.692	-.9187925	.6226903

```
. reg lempadj gdp
```

Source	SS	df	MS	Number of obs = 21	
Model	6.84252671	1	6.84252671	F(1, 19) = 5.89	Prob > F = 0.0253
Residual	22.0706501	19	1.16161317	R-squared = 0.2367	Adj R-squared = 0.1965
Total	28.9131769	20	1.44565884	Root MSE = 1.0778	

lempadj	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gdp	.35991	.1482915	2.427	0.025	.0495322	.6702877
_cons	-1.100871	.5110381	-2.154	0.044	-2.170486	-.0312554

Now RSS are comparable, and can see linear is preferred.

Formal test of significant difference between the 2 specifications

```
. g test=(21/2)*log(22.1/11.5) = N/2log(RSSlargest/RSSsmallest) ~  $\chi^2_{(1)}$ 
```

```
/* stata recognises "log" as Ln or loge */
```

```
. di test
```

6.86

Given test is Chi-Squared with 1 degree of freedom. Estimated value exceeds critical value (from tables Chi-squared at 5% level with 1 degree of freedom is 3.84) so models are significantly different in terms of goodness of fit.

Test for Normality of Residuals

All the hypotheses, tests and confidence intervals done so far are based on the assumption that the (unknown true) residuals are normally distributed. If not then tests are invalid

Test for Normality of Residuals

All the hypotheses, tests and confidence intervals done so far are based on the assumption that the (unknown true) residuals are normally distributed. If not then tests are invalid

When choosing a functional form better to choose one which gives normally distributed errors

Since never observe true residuals can instead look at the *OLS residuals*

Test for Normality of Residuals

All the hypotheses, tests and confidence intervals done so far are based on the assumption that the (unknown true) residuals are normally distributed. If not then tests are invalid

When choosing a functional form better to choose one which gives normally distributed errors

Since never observe true residuals can instead look at the *OLS residuals*

Why?

Test for Normality of Residuals

All the hypotheses, tests and confidence intervals done so far are based on the assumption that the (unknown true) residuals are normally distributed. If not then tests are invalid

When choosing a functional form better to choose one which gives normally distributed errors

Since never observe true residuals can instead look at the *OLS residuals*

Why? Can show that *if* all Gauss-Markov assumptions are satisfied (see earlier notes) then the OLS residuals are also *asymptotically* normally distributed (ie approximately normal if sample size is large)

A normal distribution should have following properties:

A normal distribution should have following properties:

1. symmetric about its mean (in the case of OLS residuals the mean is zero)

A normal distribution should have following properties:

1. symmetric about its mean (in the case of OLS residuals the mean is zero)

A Non-symmetric distribution is said to be *skewed*.

A normal distribution should have following properties:

1. symmetric about its mean (in the case of OLS residuals the mean is zero)

A Non-symmetric distribution is said to be *skewed*.

Can measure this by looking at the 3rd moment of the normal distribution relative to the 2nd (mean is the 1st moment, variance is the second moment)

$$\text{Skewness} = \frac{[E(X - \mu_X)^3]^2}{[E(X - \mu_X)^2]^3} = \frac{\text{square of } 3^{\text{rd}} \text{ moment}}{\text{cube of } 2^{\text{nd}} \text{ moment}}$$

A normal distribution should have following properties:

1. symmetric about its mean (in the case of OLS residuals the mean is zero)

A Non-symmetric distribution is said to be *skewed*.

Can measure this by looking at the 3rd moment of the normal distribution relative to the 2nd (mean is the 1st moment, variance is the second moment)

$$\text{Skewness} = \frac{[E(X - \mu_X)^3]^2}{[E(X - \mu_X)^2]^3} = \frac{\text{square of } 3^{\text{rd}} \text{ moment}}{\text{cube of } 2^{\text{nd}} \text{ moment}}$$

Symmetry is represented by a value of 0 for the skewness coefficient

A normal distribution should have following properties:

1. symmetric about its mean (in the case of OLS residuals the mean is zero)

A Non-symmetric distribution is said to be *skewed*.

Can measure this by looking at the 3rd moment of the normal distribution relative to the 2nd (mean is the 1st moment, variance is the second moment)

$$\text{Skewness} = \frac{[E(X - \mu_X)^3]^2}{[E(X - \mu_X)^2]^3} = \frac{\text{square of } 3^{\text{rd}} \text{ moment}}{\text{cube of } 2^{\text{nd}} \text{ moment}}$$

Symmetry is represented by a value of 0 for the skewness coefficient

Right skewness gives a value > 0

(more values clustered to close to left of mean and a few values a long way to the right of the mean tend to make the value >0)

A normal distribution should have following properties:

1. symmetric about its mean (in the case of OLS residuals the mean is zero)

A Non-symmetric distribution is said to be *skewed*.

Can measure this by looking at the 3rd moment of the normal distribution relative to the 2nd (mean is the 1st moment, variance is the second moment)

$$\text{Skewness} = \frac{[E(X - \mu_X)^3]^2}{[E(X - \mu_X)^2]^3} = \frac{\text{square of } 3^{\text{rd}} \text{ moment}}{\text{cube of } 2^{\text{nd}} \text{ moment}}$$

Symmetry is represented by a value of zero for the skewness coefficient

Right skewness gives a value > 0

(more values clustered to close to left of mean and a few values a long way to the right of the mean tend to make the value >0)

Left skewness gives a value < 0

(kdensity age, normal)

2. A distribution is said to display *kurtosis* if the height of the distribution is unusual

2. A distribution is said to display *kurtosis* if the height of the distribution is unusual

(suggests observations more bunched or more spread out than should be).

2. A distribution is said to display *kurtosis* if the height of the distribution is unusual

(suggests observations more bunched or more spread out than should be).

Measure this by

$$\textit{Kurtosis} = \frac{E(X - \mu_X)^4}{[E(X - \mu_X)^2]^2} = \frac{4^{\textit{th}} \textit{moment}}{\textit{square of } 2^{\textit{nd}} \textit{moment}}$$

2. A distribution is said to display *kurtosis* if the height of the distribution is unusual

(suggests observations more bunched or more spread out than should be).

Measure this by

$$Kurtosis = \frac{E(X - \mu_X)^4}{[E(X - \mu_X)^2]^2} = \frac{4^{th} \text{ moment}}{\text{square of } 2^{nd} \text{ moment}}$$

A normal distribution should have a kurtosis value of 3

Can combine both these features to give the **Jarque-Bera Test for Normality**
(in the residuals)

$$JB = N * \left[\frac{Skewness^2}{6} + \frac{(Kurtosis - 3)^2}{24} \right]$$

Can combine both these features to give the **Jarque-Bera Test for Normality** (in the residuals)

$$JB = N * \left[\frac{Skewness^2}{6} + \frac{(Kurtosis - 3)^2}{24} \right]$$

Can show that this is *asymptotically* Chi² distributed with 2 degrees of freedom (1 for skewness and 1 for kurtosis)

Can combine both these features to give the **Jarque-Bera Test for Normality** (in the residuals)

$$JB = N * \left[\frac{Skewness^2}{6} + \frac{(Kurtosis - 3)^2}{24} \right]$$

Can show that this is *asymptotically* Chi² distributed with 2 degrees of freedom (1 for skewness and 1 for kurtosis)

If estimated chi-squared > chi-squared_{critical}

reject null that residuals are normally distributed

Can combine both these features to give the **Jarque-Bera Test for Normality** (in the residuals)

$$JB = N * \left[\frac{Skewness^2}{6} + \frac{(Kurtosis - 3)^2}{24} \right]$$

Can show that this is *asymptotically* Chi² distributed with 2 degrees of freedom (1 for skewness and 1 for kurtosis)

If estimated chi-squared > chi-squared_{critical}

reject null that residuals are normally distributed

(If not suggests should try another functional form to try and make residuals normal, otherwise t stats may be invalid).

Example: **Jarque-Bera Test for Normality (in residuals)**

```
. u wage /* read in data */
1st regress hourly pay on years of experience and get residuals
. reg hourpay xper
```

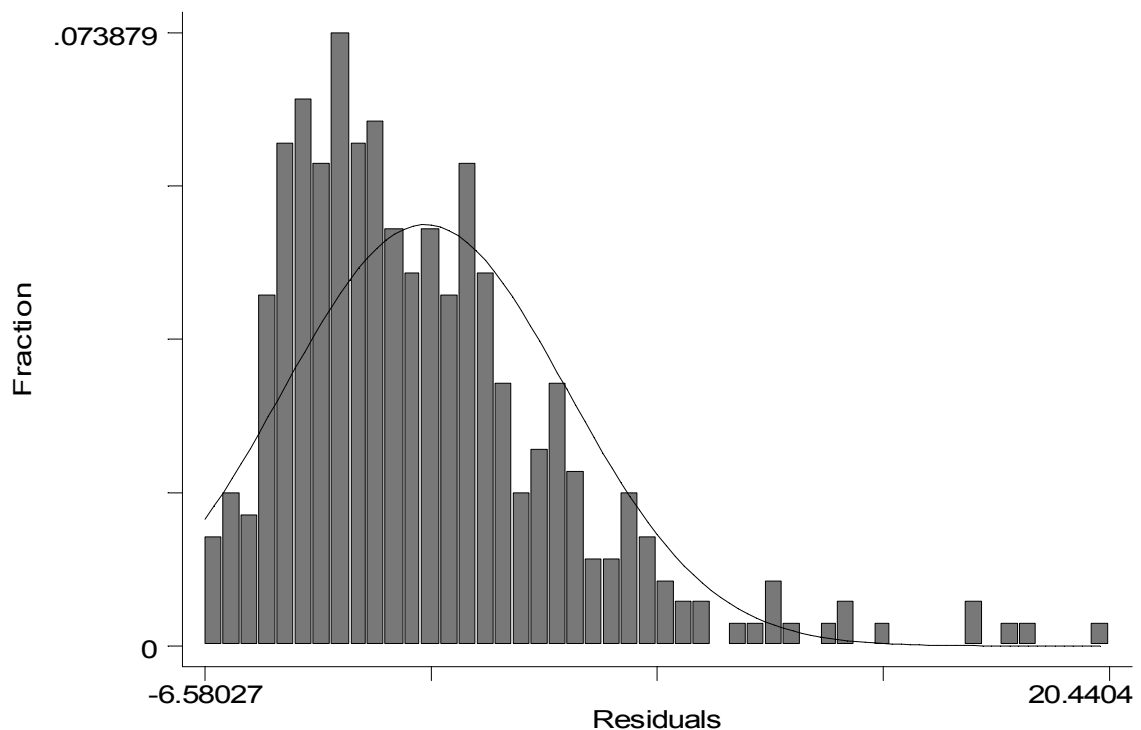
Source	SS	df	MS	Number of obs =	379
Model	136.061219	1	136.061219	F(1, 377) =	7.53
Residual	6815.41926	377	18.0780352	Prob > F =	0.0064
				R-squared =	0.0196
				Adj R-squared =	0.0170
Total	6951.48048	378	18.39016	Root MSE =	4.2518

hourpay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
xper	.0487259	.017761	2.743	0.006	.0138028 .083649
_cons	7.26455	.4333534	16.764	0.000	6.412457 8.116642

```
. predict res, resid
```

Check histogram of residuals using the following stata command

```
. gra res, normal bin(50)
/* normal option superimposes a normal distribution on the graph */
```



Residuals show signs of right skewness (residuals bunched to left - not symmetric) and kurtosis (leptokurtic - since peak of distribution higher than expected for a normal distribution)

To test more formally

```
. su res, detail
```

Residuals				

	Percentiles	Smallest		
1%	-6.253362	-6.580268		
5%	-4.919813	-6.372607		
10%	-4.27017	-6.313276	Obs	379
25%	-3.011451	-6.253362	Sum of Wgt.	379
50%	-.9261839		Mean	1.11e-08
		Largest	Std. Dev.	4.246199
75%	1.869452	16.5097		
90%	5.383683	17.73377	Variance	18.03021
95%	7.480312	17.9211	Skewness	1.50555
99%	16.5097	20.44043	Kurtosis	6.432967

Construct Jarque-Bera test

$$. jb = (379/6)*((1.50555^2)+((6.43-3)^2)/4)$$

$$= 328.9$$

The statistic has a χ^2 distribution with 2 degrees of freedom, (one for skewness one for kurtosis).

From tables critical value at 5% level for 2 degrees of freedom is 5.99

*So $JB > \chi^2_{critical}$, so **reject** null that residuals are normally distributed.*

Suggests should try another functional form to try and make residuals normal, otherwise t stats may be invalid.

Remember this test is only valid asymptotically, so it relies on having a large sample size. Users with data sets smaller than 50 observations should be wary about using this test.