SSSII

Gwilym Pryce

Gpryce.com

Lecture 5: Omitted Variables & Measurement Errors

Plan:

- (1) Regression Assumptions
- (2) Omitted variables [I(b)]
- (3) Inclusion of Irrelevant Variables [1(c)]
- (4) Errors in variables [1(d)]
- (5) Error term with non zero mean [2]

(1) Regression assumptions

For estimation of *a* and *b* and for regression inference to be correct:

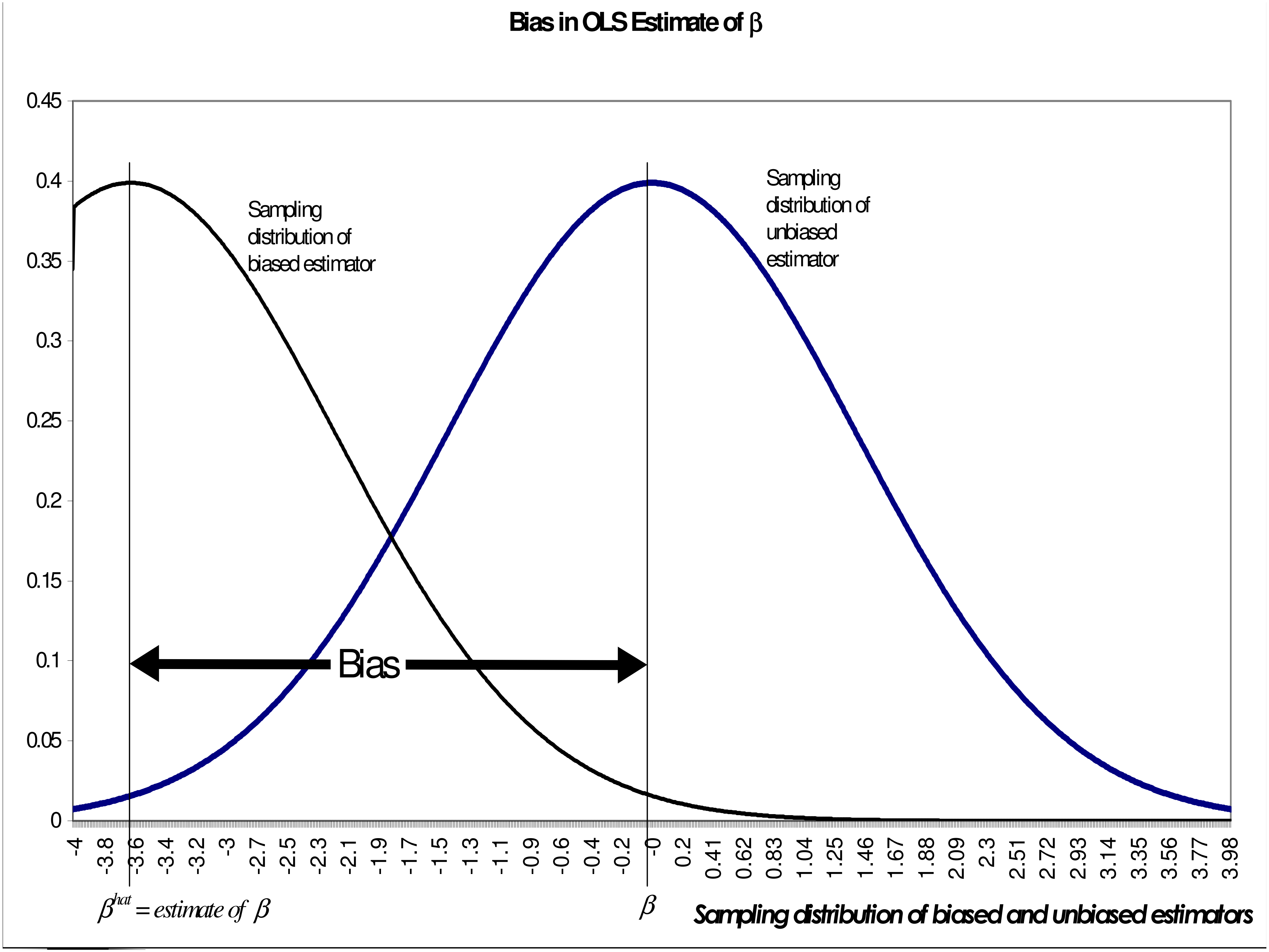
- 1. Equation is correctly specified:
  - (a) Linear in parameters (can still transform variables)
  - (b) Contains all relevant variables
  - (c) Contains no irrelevant variables
  - (d) Contains no variables with measurement errors
- 2. Error Term has zero expected mean – i.e. long run average value of error term is zero:
  - this simply says that in the population (or repeated samples), the distribution of the unobservables (captured by the error term) has zero mean.
- 3. Error Term has constant variance
- 4. Error Term is not autocorrelated
  - I.e. correlated with error term from previous time periods
- 5. Explanatory variables are fixed
  - observe normal distribution of *y* for repeated fixed values of *x*
- 6. No linear relationship between RHS variables
  - I.e. no “multicolinearity”

Diagnostic Tests and Analysis of Residuals

- Diagnostic tests are tests that are meant to "diagnose" problems with the models we are estimating.
  - Least squares residuals play an important role in many diagnostic tests -- some of which we have already looked at.
    - E.g. F-tests of parameter stability
- For each violation we shall look at the Consequences, Diagnostic Tests, and Solutions.

(2) Omitted variables [violation 1(b)]

- Consequences:
  - usually the OLS estimator of the coefficients of the remaining variables will be biased
    - bias = (coefficient of the excluded variable) × (regression coefficient in a regression of the excluded variable on the included variable)
  - where we have several included variables and several omitted variables:
    - the bias in each of the estimated coefficients of the included variables will be a weighted sum of the coefficients of all the excluded variables
      - the weights are obtained from (hypothetical) regressions of each of the excluded variables on all the included variables.


Bias in OLS Estimate of β
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Sampling distribution of biased estimator
Sampling distribution of unbiased estimator
Bias
-4
-3.8
-3.6
-3.4
-3.2
-3
-2.7
-2.5
-2.3
-2.1
-1.9
-1.7
-1.5
-1.3
-1.1
-0.9
-0.6
-0.4
-0.2
0
0.2
0.41
0.62
0.83
1.04
1.25
1.46
1.67
1.88
2.09
2.3
2.51
2.72
2.93
3.14
3.35
3.56
3.77
3.98
$\beta^{hat} = estimate\ of\ \beta$
β
Sampling distribution of biased and unbiased estimators

- also inferences based on these estimates will be inaccurate because estimates of the standard errors will be biased
    - so t-statistics etc. will not be reliable.
- Where there is an excluded variable, the variance of coefficients of variables that are included will actually be lower than if there were no excluded variables.

- Diagnostic Tests:
  - (i) a **low R^2** is the most obvious sign that explanatory variables are missing, but this can also be caused by incorrect functional form (I.e. non-linearities).
  - (ii) If the omitted variable is known/measurable, you can enter the variable and check the **t-value** to see if it should be in.
  - (iii) Ramsey's regression specification error test (**RESET**) for omitted variables:
    - Ramsey (1969) suggested using y^{hat2}, y^{hat3} and y^{hat4} as proxies for the omitted and unknown variable *z:*

RESET test procedure:

- 1. Regress y on the known explanatory variable(s) x:

  $$y = b1 + b2x + u$$

  and obtain the predicted values, y^{hat}

- 2. Regress y on x, y^{hat2}, y^{hat3} and y^{hat4}:

  $$y = g_1 + g_2\, x + g_3\, y^{hat2} + g_4\, y^{hat3} + g_5 y^{hat4} + e$$

- 3. Do an F-test on whether the coefficients on y^{hat2}, y^{hat3} and y^{hat4} are all equal to zero.
  - If the significance level is low and you can reject the null, then there is evidence of an omitted variable(s):

    H_0: no omitted variables: $g_3 = g_4 = g_5 = 0$.

    H_1: there are omitted variables: g_3, g_4, g_5 do not all equal zero.

- Solutions:
  - Use/create proxies
  - As a general rule it is better to include too many variables than have omitted variables because inclusion of irrelevant variables does not bias the OLS estimators of the slope coefficients.

(3) Inclusion of Irrelevant Variables [violation 1(c)]

- Consequences:
  - OLS estimates of the slope coefficient of the standard errors will **not** be biased
  - however, the OLS estimate will not be “best” (cf BLUE) because the standard errors will be larger than if irrelevant variables had been excluded (I.e. the OLS will not be as *“efficient”*).
  - This means that the t-values will be lower than they should be, and the confidence intervals for the slope coefficients larger than would be the case if only relevant variables were included.

- Diagnostic tests:
  - – t-tests (Backward and Forward methods) but use with care:
    - better to make reasoned judgements
  - – F-tests on groups of variables
  - – compare adjusted R^2 of model with the variable included with the adjusted R^2 of the model without the variable.

- Hierarchical (or *sequential*) regression:
  - Allows you to add in variables one at a time and consider the contribution it makes to the R^2
    - in SPSS Linear Regression window, enter the first block of independent variables
    - then click **Next** and enter your second block of independent variables.
    - Click on the **Statistics** button and tick the boxes marked **Model Fit**, and **R squared change**.
    - Click **Continue**

- <u>Solutions:</u>
  - inclusion of irrelevant variables is not as severe as the consequences of omitting relevant variables, so the temptation is to include "*everything but the kitchen sink*".
  - There is a balancing act between <u>*bias*</u> and <u>*efficiency*</u>.
    - A small amount of bias may be preferable to a great deal of inefficiency.
  - The best place to start is with good theory.
    - Then include all the variables available that follow from this theory
    - and then exclude variables that add least to the model and are of least theoretical importance.

(4) Errors in variables
[violation 1(d)]

- <u>Consequences:</u>
  - “The Government are very keen on amassing statistics -- they collect them, add them, raise them to the *n*th power, take the cube root and prepare wonderful diagrams. But what you must never forget is that every one of those figures comes in the first instance from the village watchman, who just puts down what he damn pleases”

(Stamp, 1929, pp. 258-9; quoted in Kennedy, p. 140)

- Errors in the dependent variable are not usually a problem since such errors are incorporated in the disturbance term.
- Errors in explanatory variables are more problematic, however.
  - The consequences of measurement errors in explanatory variables depend on whether or not the variables mismeasured are **independent of the disturbance term**.
    - If not independent of the error term, OLS estimates of slope coefficients will be biased.

- Diagnostic Tests:
  - no simple tests for general mismeasurement
    - correlations between error term and explanatory variables may be caused by other factors such as simultaneity.
  - Errors in the measurement of specific observations can be tested for, however, by looking for outliers
    - but again, outliers may be caused by factors other than measurement errors.
    - Whole raft of measures and means for searching for outliers and measuring the influence of particular observations -- we'll look at some of these in the lab.

- Solutions:
  - if there are different measures of the same variable, present results for both to see how sensitive the results are.
  - If there are clear outliers, examine them to see if they should be omitted.
  - If you know what the measure error is, you can weight the regression accordingly (see p. 141 of Kennedy) but since we rarely know the error, this method is not usually much use.

– In time series analysis there are instrumental variable methods to address errors in measurement (not covered in this course)

– if you know the variance of the measurement error, Linear Structural Relations methods can be used (see Kennedy), but again, these methods are rarely used since we don’t usually know the variance of measurement errors.

(5) Non normal & Nonzero Mean Expected Errors

[violation 2]

- Consequences:
  - note that the OLS estimation procedure is such as to automatically create residuals whose mean is zero.
    - OLS adjusts the value of the coefficient to ensure the errors have zero mean.
  - So we cannot formally test for non-zero mean residuals
    - But be aware of theoretical reasons why a particular model might theoretically produce non-zero means in repeated samples

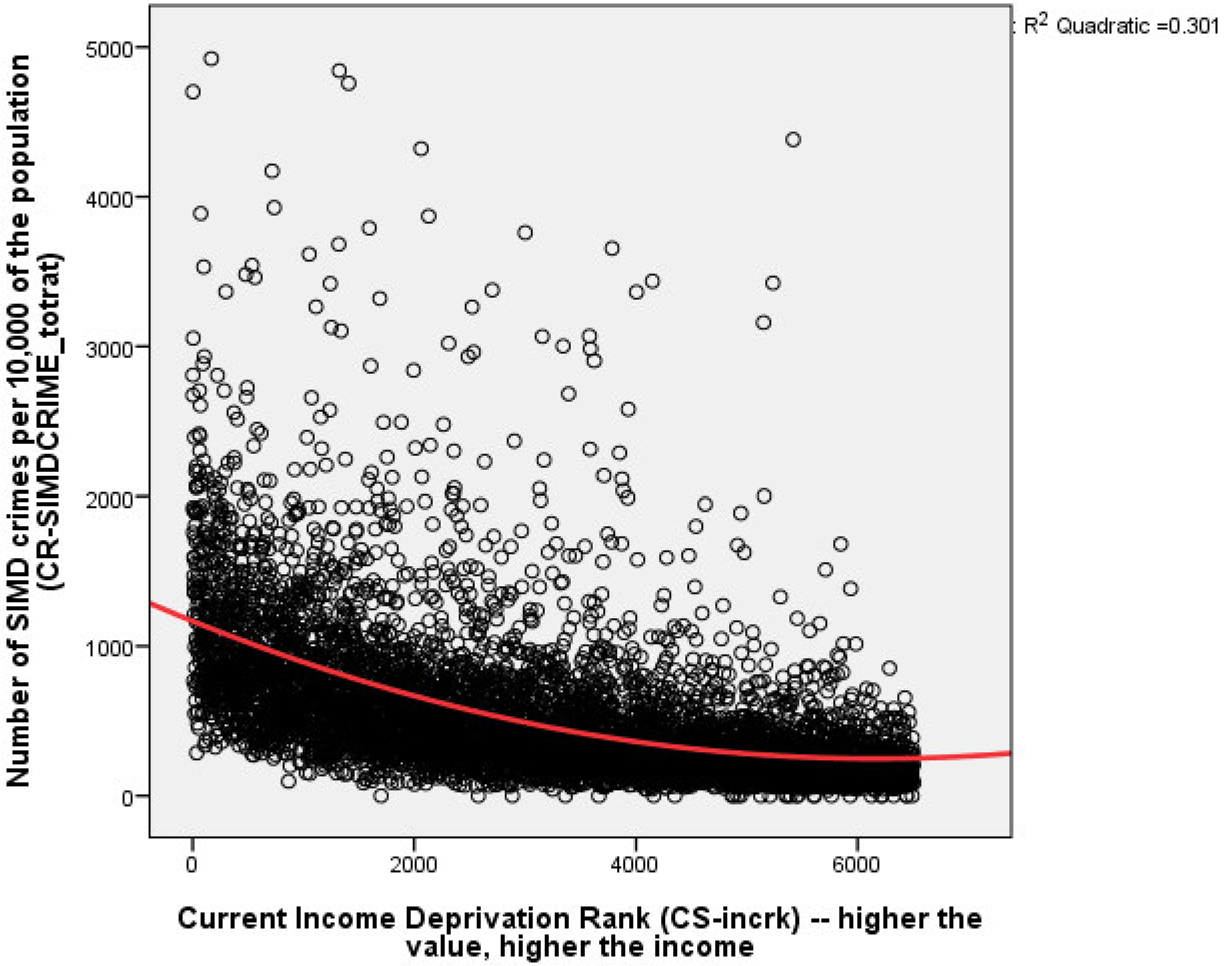
- if the nonzero mean is constant (due, for example, to systematically positive or systematically negative errors of measurement in the dependent variable)
  - then the OLS estimation of the intercept will be biased
- if the non-zero mean is due to omitted variables, then the error of the misspecified equation will not have a constant, zero mean.
  - This should be viewed as a violation of the first assumption of OLS (ommitted varialbles), not the second.
- We don’t need to assume normally distributed errors in order for OLS estimates to be BLUE.
  - However, we do need them to be normally distributed in order for the t-tests and F-tests to be reliable.
- Non-normal errors are usually due to other mispecification errors
  - such as non-linearities in the relationships between variables.

- Diagnostic Tests:
  - Shape of the distribution of errors can be examined visually by doing a histogram or *normal probability plot*:
    - Normal probability plots (also called normal quantile plots) are calculated for a variable *x* as follows:

1. Arrange the data values from smallest to largest.
   - Record what percentile of data each value occupies.
   - E.g. the smallest observation in a set of 20 is at the 5% point, the second smallest is at the 10% point, and so on”
2. Do normal distribution calculations to find the *z*-score values at these same percentiles.
   - E.g. $z = -1.645$ is the 5% point of the standard normal distribution, and $z = -1.282$ is the 10% point.
3. Plot each data point *x* against the corresponding *z*.
   - If the data distribution is close to standard normal, the plotted points will lie close to the 45 degree line $x = z$.
   - If the data distribution is close to ***any*** normal distribution, the plotted points will lie close to some straight line
     - (this is because standardising turns any normal distribution into a standard normal and standardising is a linear transformaiton -- affects slope and intercept but cannot turn a line into a curved pattern)

(Moore and McCabe)

E.g.1 Crime & Income (SNS)



Impose a linear line of best fit:

```
REGRESSION
/DEPENDENT crime_rate_04
/METHOD=ENTER income
/RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
/SAVE RESID(U) PRED(crime_hat).
```

Model Summary[b]

R Square	Adjusted R Square
.278	.278

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1003.204	10.807		92.832	.000
	Current Income Deprivation Rank (CS-incrk) -- higher the value, higher the income	-.143	.003	-.528	-48.433	.000

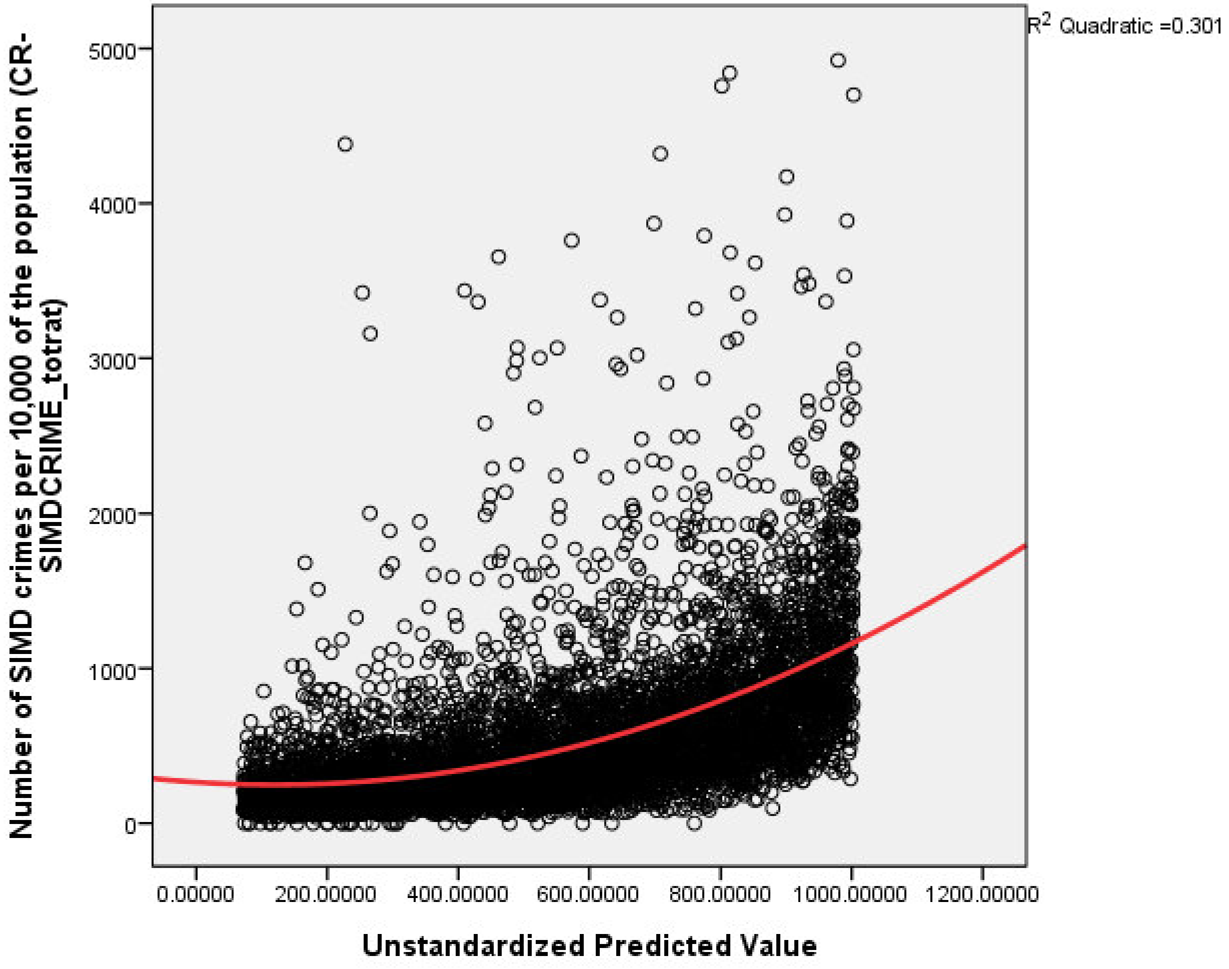
a. Dependent Variable: Number of SIMD crimes per 10,000 of the population (CR-SIMDCRIME_totrat)

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	72.02	1003.06	552.81	266.639	6082
Residual	-782.218	4154.636	.000	429.278	6082
Std. Predicted Value	-1.803	1.689	.000	1.000	6082
Std. Residual	-1.822	9.677	.000	1.000	6082

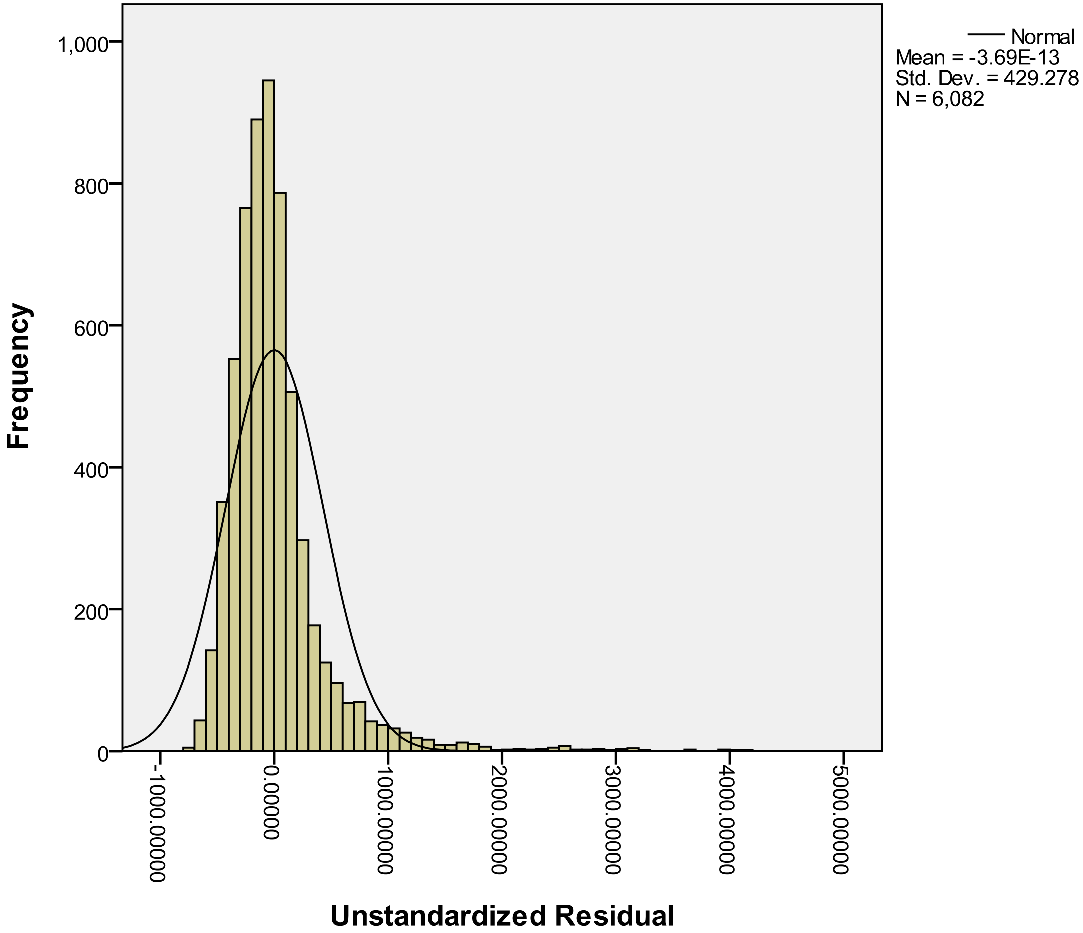
a. Dependent Variable: Number of SIMD crimes per 10,000 of the population (CR-SIMDCRIME_totrat)

Scatter of y on yhat should be linear:

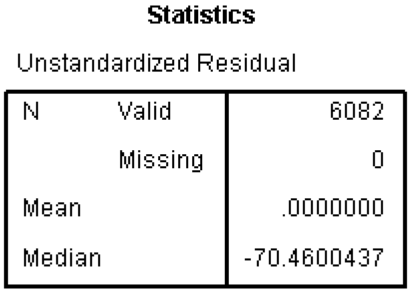


Are the residuals normal?

- Positive skew: Fat upper tail



Mean > median:

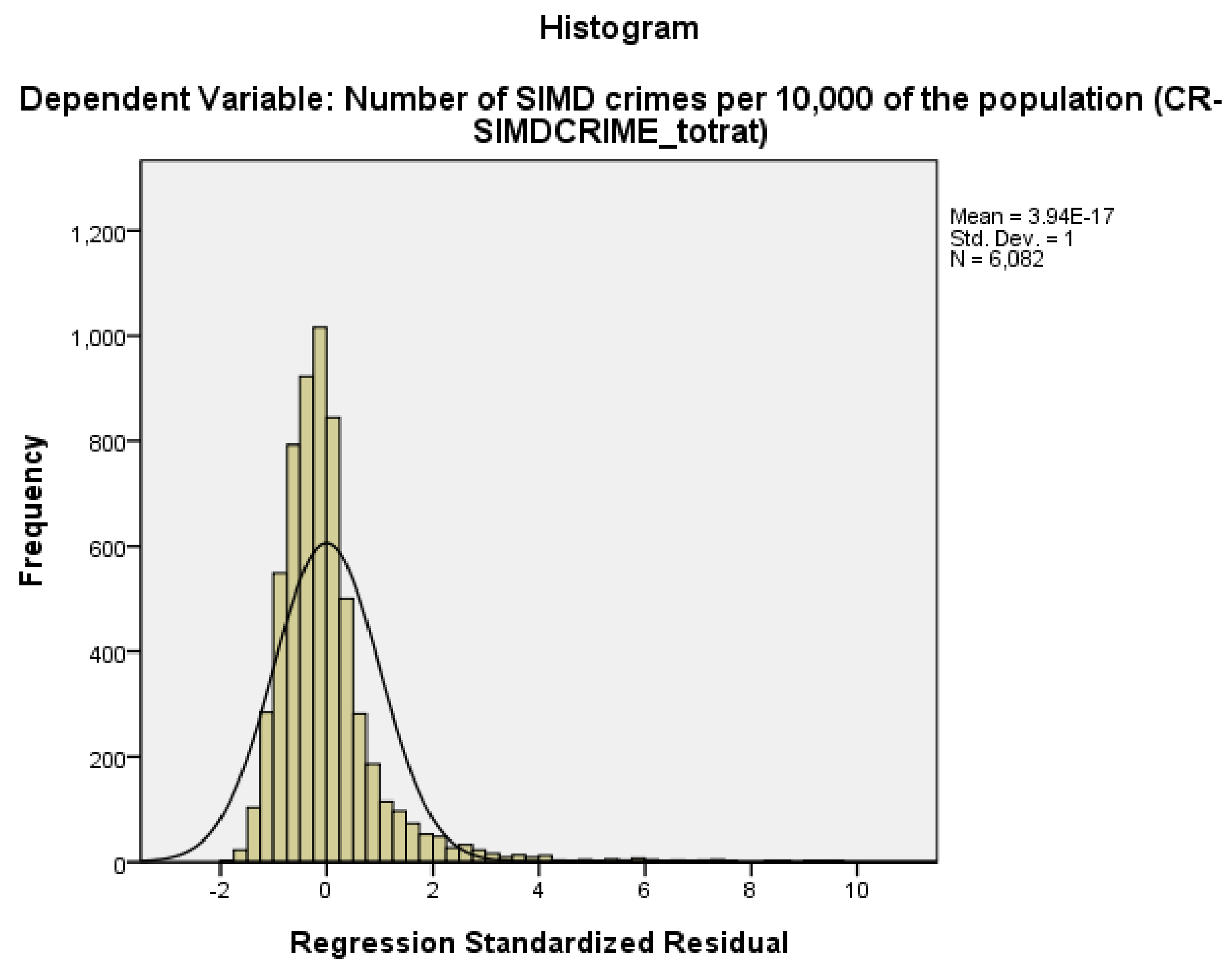


Statistics

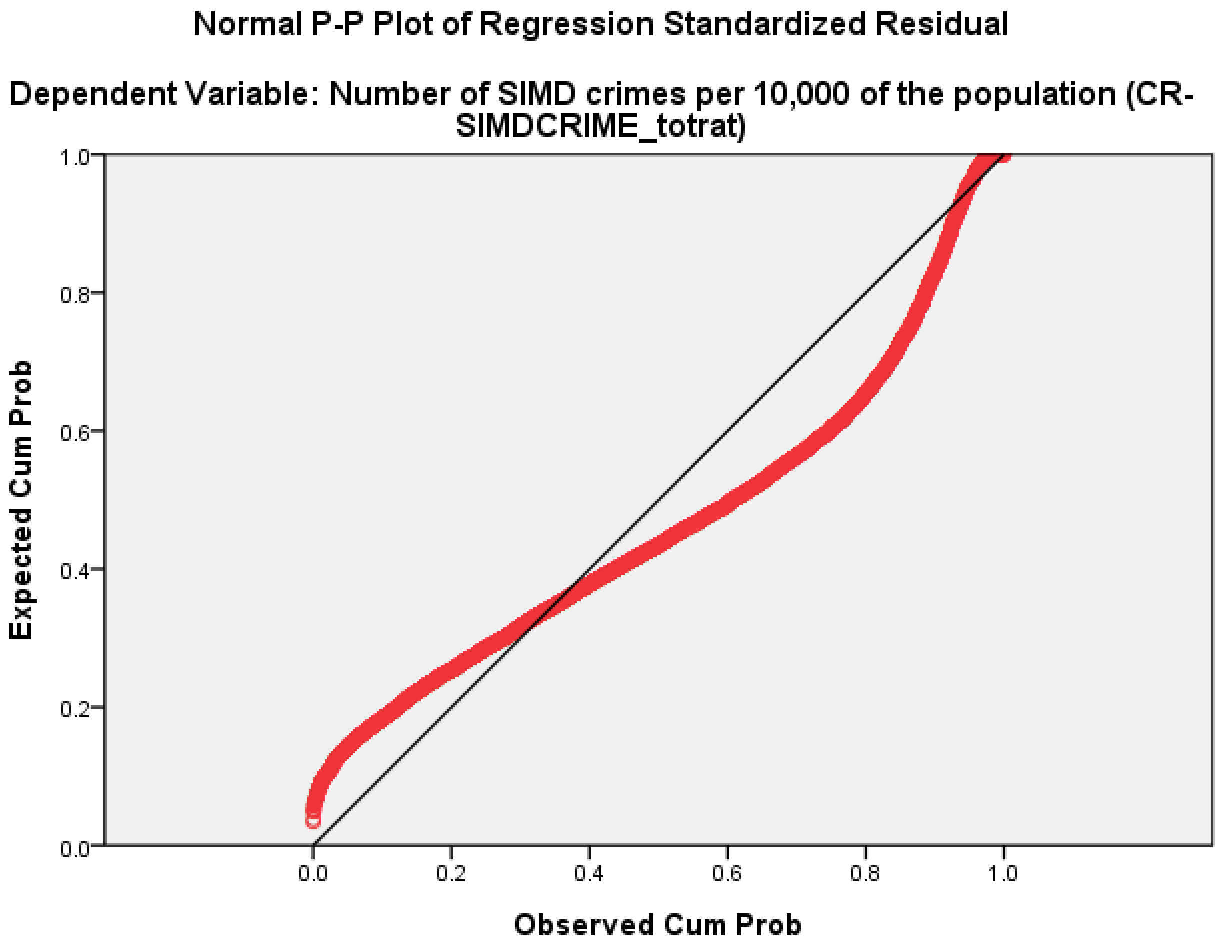
Unstandardized Residual

N	Valid	6082
	Missing	0
Mean		.0000000
Median		-70.4600437

Standardised residuals:



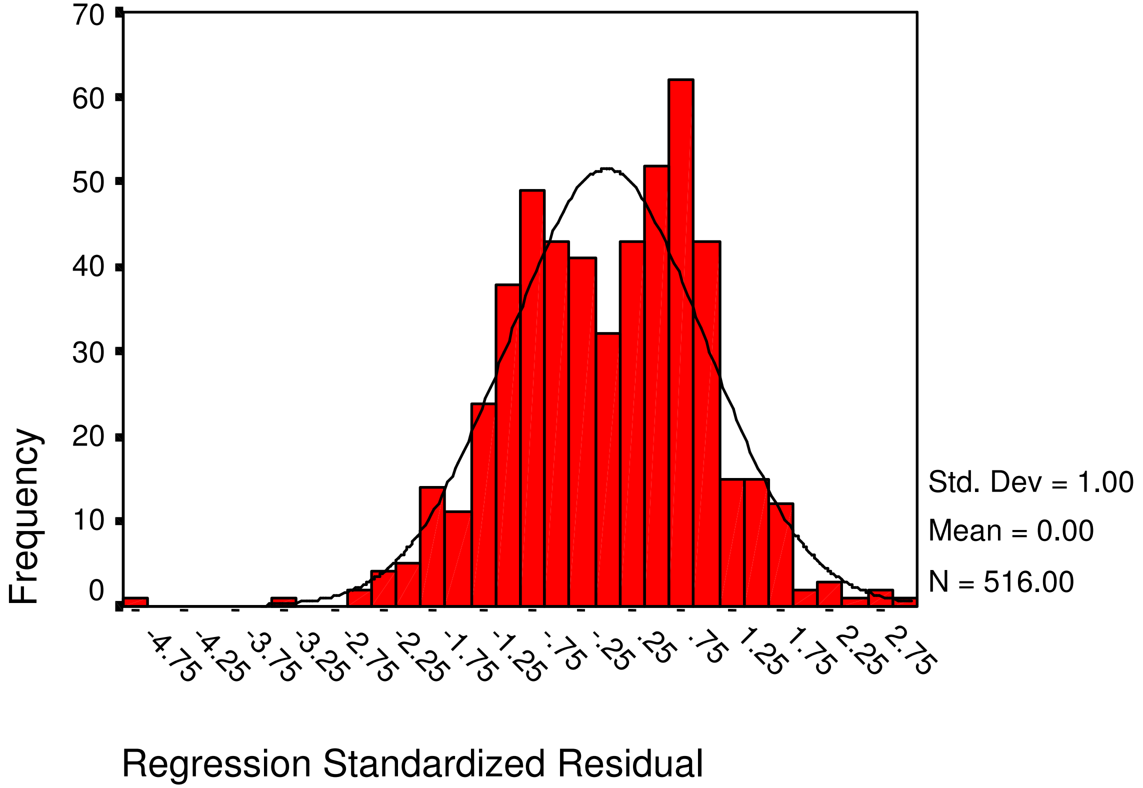
Normal Probability Plot:



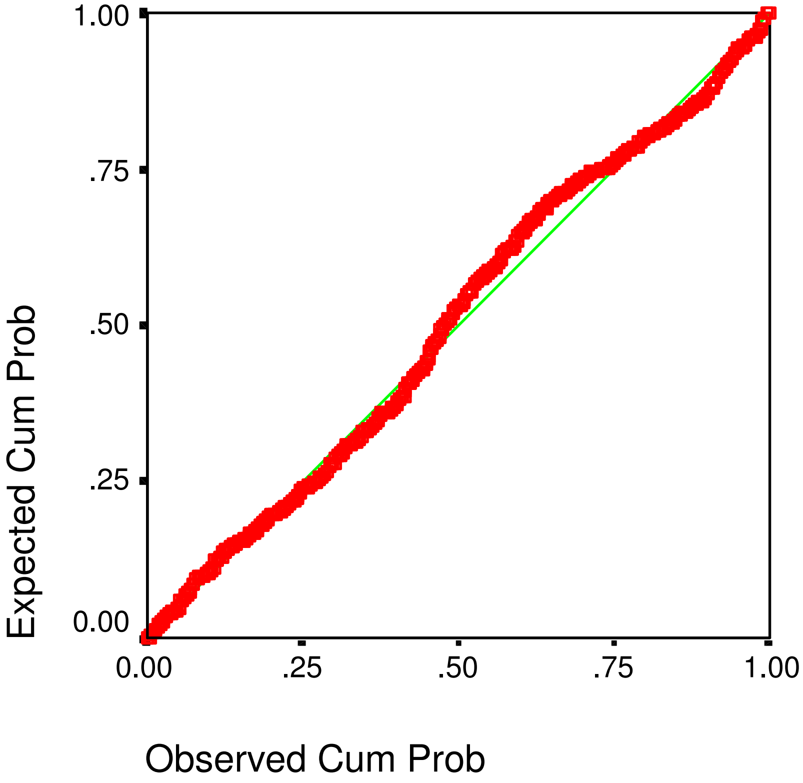
E.g.2 Imports Regression:

Histogram

Dependent Variable: Imports per capita



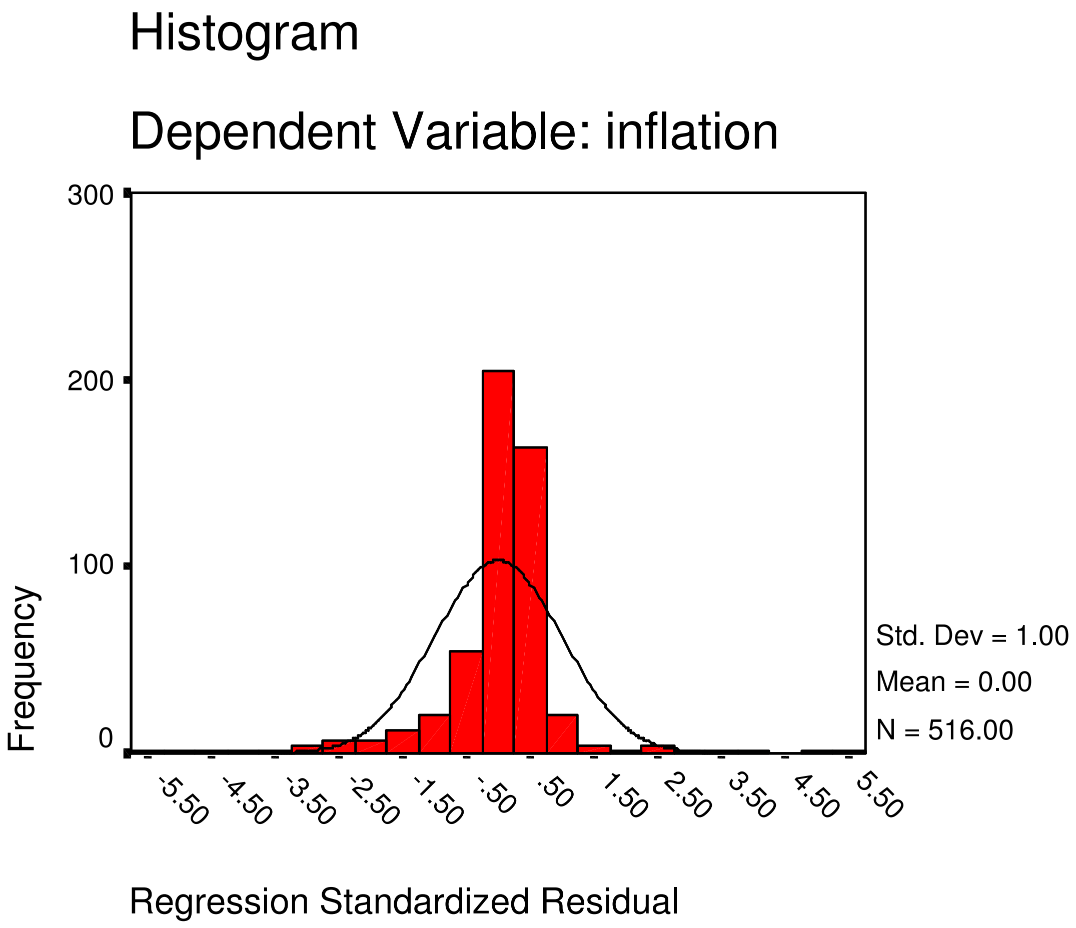
Normally Distributed Errors:



Normal P-P Plot of Regression Standardized Residual

Dependent Variable: Imports per capita

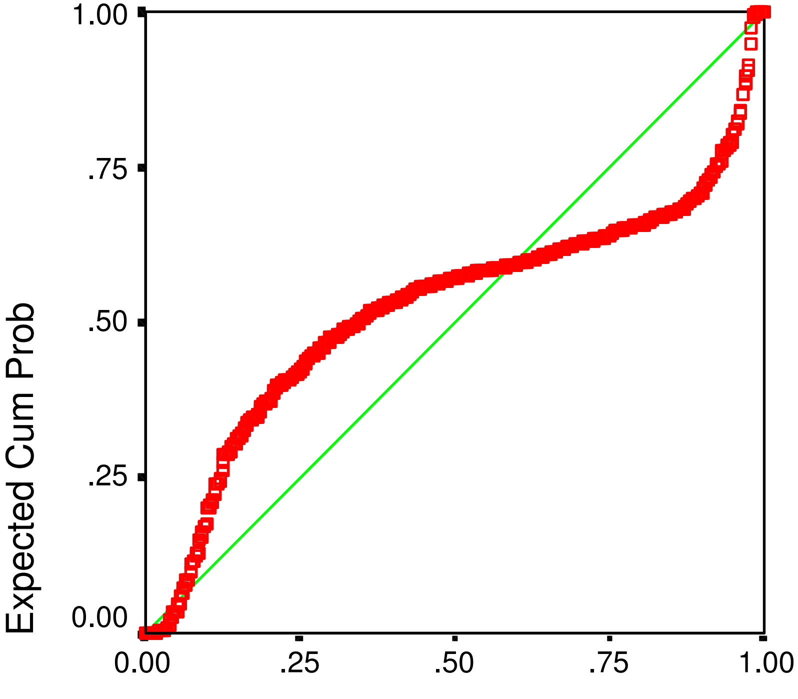
Non-Normal Errors:



Non-Normal Errors:

Normal P-P Plot of Regression Standardized Residual

Dependent Variable: inflation



Observed Cum Prob

Summary

- (1) Regression Assumptions
- (2) Omitted variables [I(b)]
- (3) Inclusion of Irrelevant Variables [1(c)]
- (4) Errors in variables [1(d)]
- (5) Error term with non zero mean [2]

- Reading:
  - Kennedy (1998) “A Guide to Econometrics”, Chapters 5,6,7 and 9
  - Maddala, G.S. (1992) “Introduction to Econometrics” chapter 12
  - Field, A. (2000) chapter 4, particularly pages 141-162.