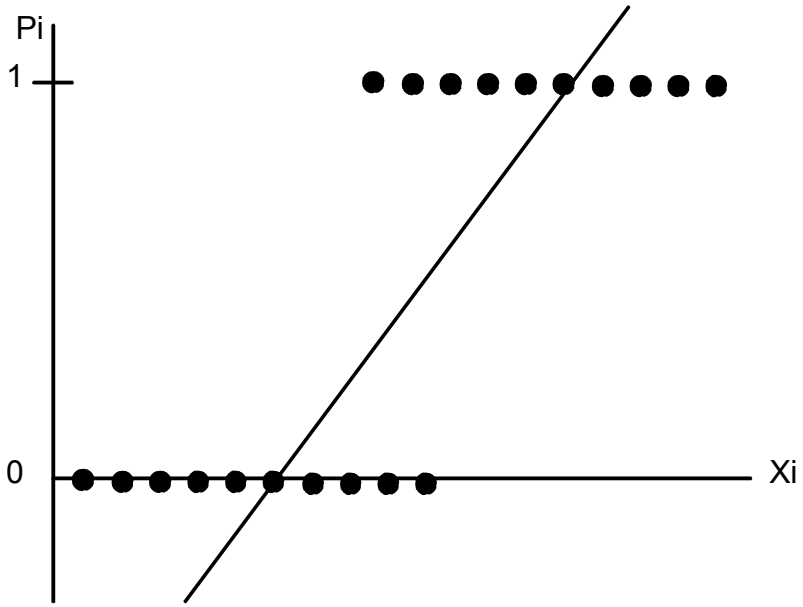
CHAPTER 5: DUMMY DEPENDENT VARIABLES AND NON-LINEAR REGRESSION

1. The Problem of Dummy Dependent Variables

- You already learned about dummies as independent variables. But what do you do if the dependent variable is a dummy?
- One answer is: Logistic regression
- Of course, you could also run OLS, which, however, has obvious limitations.

Figure 1: OLS in Dummy Dependent Estimation



- Problems with OLS when the dependent variable is a binominal dummy are:
- The error term is obviously not normally distributed. The error term is heteroskedastic.
- R-squared becomes a useless measure
- Most importantly, the model is problematic for forecasting purposes. One would like to forecast the probability of a certain set of independent variables to create a certain binominal outcome. OLS could create probabilities of greater than one or smaller than zero.
- Logistic regression is a non-linear estimation technique, which solves the problem of unboundedness of OLS.

2. The Logit - Model

- The Logit-Model is defined as:

$$LN\left(\frac{P_i}{(1-P_i}\right) = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_2 + \varepsilon \qquad (1)$$

- It is based on the cumulative logistic distribution

- (1) an be rearranged to:

$$P_i = \frac{1}{1 + e^{-[\beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_2 + \varepsilon]}} \qquad (2)$$

- How?

- Define

$$LN\left(\frac{P}{1-P}\right) = Z \qquad \text{|Take Antilog}$$

$$\frac{P}{1-P} = e^{Z} \qquad \text{| Times (1-P)}$$

$$P = e^{Z} - Pe^{Z} \qquad \text{| Divide by P}$$

$$1 = \frac{e^{Z}}{P} - e^{Z} \qquad \text{| Plus } e^{Z}$$

$$1 + e^{Z} = \frac{e^{Z}}{P} \qquad \text{| Take inverse}$$

$$\frac{1}{1+e^{Z}} = \frac{P}{e^{Z}} \qquad \text{| Multiply by } e^{Z}$$

$$P = \frac{e^{Z}}{1+e^{Z}} \qquad \text{| Expand right side by } \frac{e^{-Z}}{e^{-Z}}$$

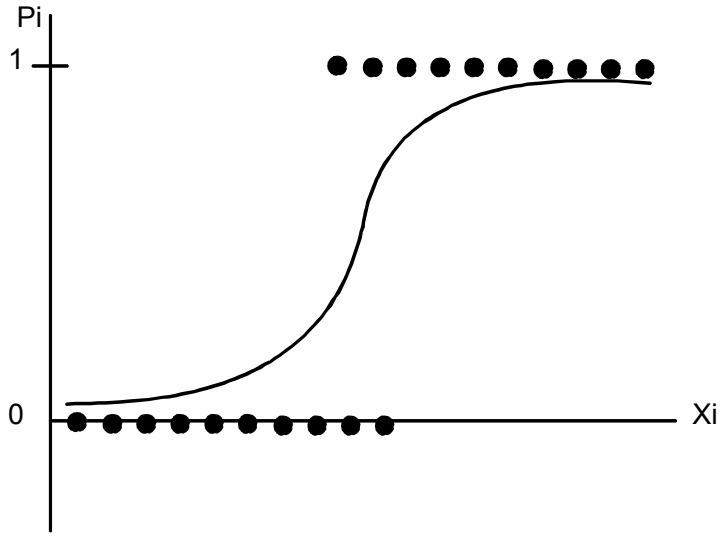
$$P = \frac{1}{1+e^{-Z}}$$

- Thus,

$$LN\left(\frac{P_i}{(1-P_i)}\right) = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_2 + \varepsilon = P_i = \frac{1}{1 + e^{-[\beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_2 + \varepsilon]}}$$

- If $Z \rightarrow 0$ ➔ $P_i \rightarrow 0$ and if $Z \rightarrow \infty$ ➔ $P_i \rightarrow 1$
- Therefore, logistic regression solves the unboundedness problem of OLS.

Figure 2: Logistic Regression



- You will also encounter Probit models. Their idea is similar to the logit. The only difference is that the probit estimates are derived out of the cumulative normal distribution.
- In practice, either method yields pretty identical results.

3. Interpretation

- A positive coefficient in estimating

$$LN\left(\frac{P_i}{(1-P_i}\right) = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_2 + \varepsilon$$

tells you that as X_i increases, the likelihood that the DV takes the value 1 increases.

- Most statistical software packages actually calculate the probability P_i
- Statistical software packages also report the so-called odds ratio. If it is positive, $P_i > 0.5$; if it is negative $P_i<0.5$. An odds ratio of 2:1 (2), for example, tells you that $P_i = 0.66$; and an odds ratio of 1:2 (0.5), that $P_i = 0.33$.

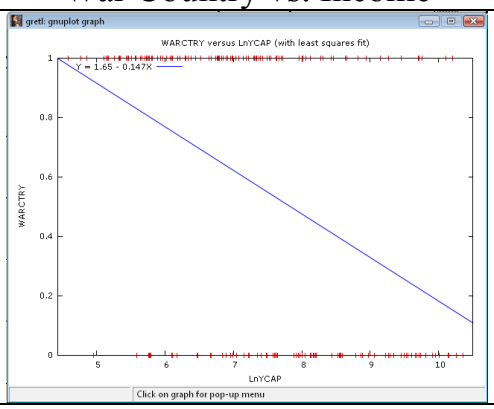
4. Example: War Risk

- Upload the dataset "War.xls," which classifies countries as a war country if they had at least one year of armed conflict between 1960 and 2005.
- The other variables are per capita income in (2000 USD, ln), Polity (a measure of democracy), income inequality, manufacturing export share (% of GDP), and Muslim Christian Polarization (the likelihood of obtaining a Muslim and a Christian in a random drawing from the population).
- The dataset also contains neighborhood effects for the regions DivMENA (diversified economies of the Middle East), OilMENA (oil economies of the Middle East), Sub Saharan Africa (SSA), Latin America and the Caribbean (LAC), South Asia (SA), East Asia and the Pacific (EAP), East Asian Tigers (EAT), North America (NAM), Western Europe (WE), and Eastern and Central Europe (ECE).
- The neighborhood effects are population weighted regional polity and regional oil (fuel exports as a percentage of GDP).
- It also contains the number of refugees per 100,000 (RegRef).

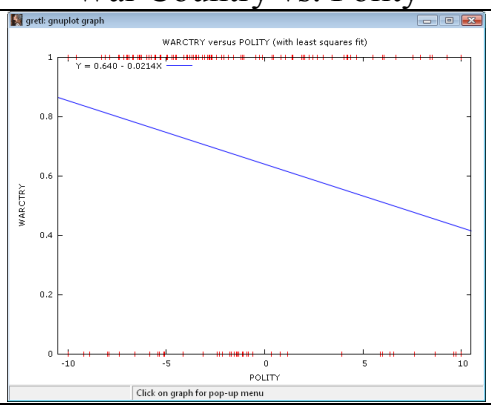
- A look at some scatter plots is useful to see why OLS is problematic with dummy dependent variables.

- The dependent variable is always "War Country" while the independent variables are per capita income (ln), polity, inequality, and manufacturing export shares respectively.

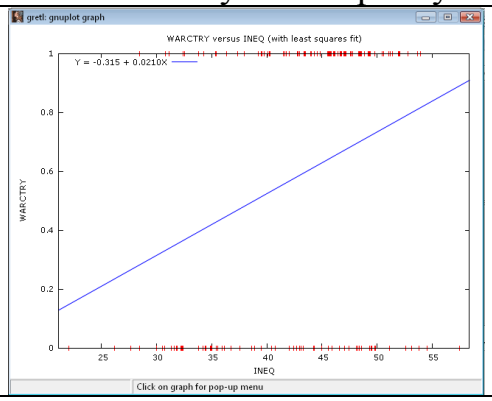
War Country vs. Income



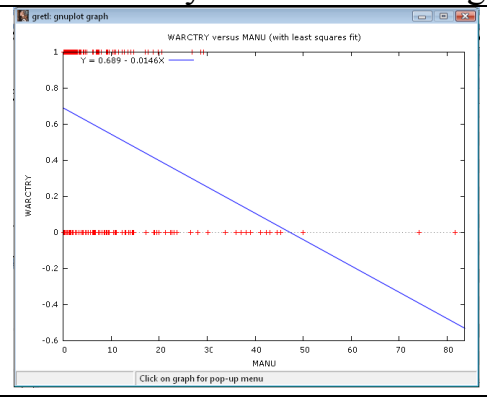
War Country vs. Polity



War Country vs. Inequality

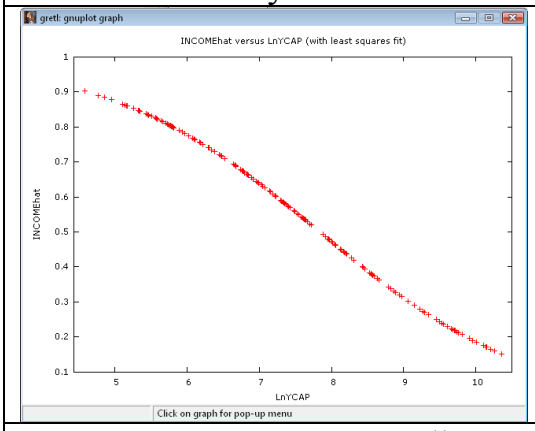
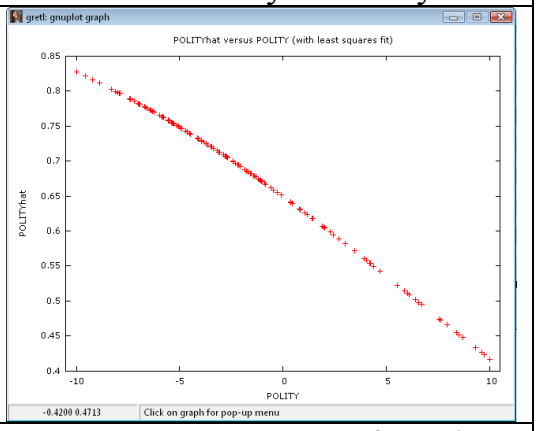
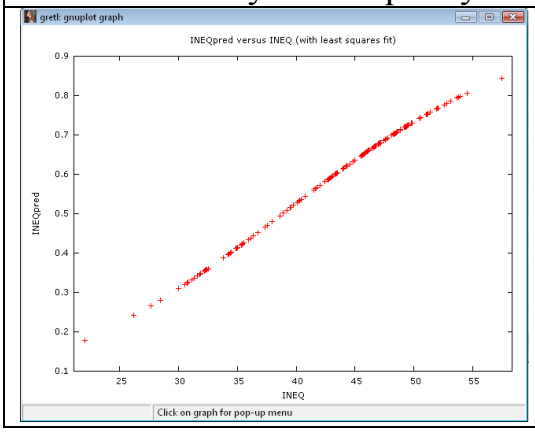
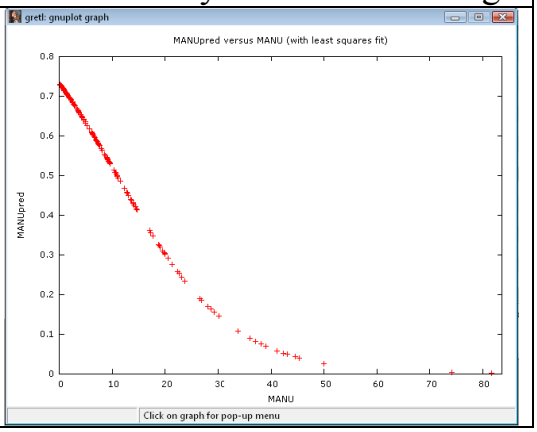


War Country vs. Manufacturing



- In all cases the predicted value are difficult to interpret, especially in the case of "War Country vs. Manufacturing," which illustrates the problem of unboundedness.

- Logistic regression is more meaningful.

War Country vs. Income	War Country vs. Polity
	
War Country vs. Inequality	**War Country vs. Manufacturing**
	

- The above scatter plots are generated as follows. Go to Model → Nonlinear models → Logit → Binary. The dependent variable is "War Country." The independent variable is, for example, per capita income. Run the model. Save the fitted values. Create scatter plot of "Fitted value vs. Income."

- Reporting logistic regression results.
- Model

$$Ln\left(\frac{WarCtry}{NoWarCtry}\right) = \beta_0 + \beta_1 Inc_i + \beta_2 Polity_i + \beta_3 Ineq_i + \beta_4 Manu_i + \varepsilon_i$$

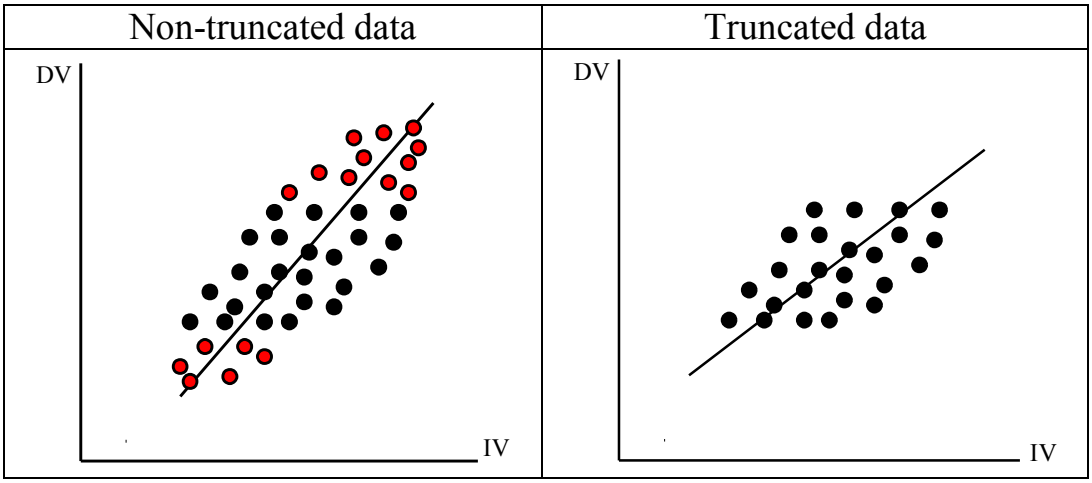
- Regression results (Example)

	I	II	III	IV	V
Const	5.35 (<0.01)	0.61 (<0.01)	-3.52 (<0.01)	1.00 (<0.01)	2.19 (0.29)
LnYCAP	-0.68 (<0.01)				-0.53 (<0.01)
POLITY		-0.09 (<0.01)			
INEQ			0.09 (<0.01)		0.05 (0.09)
MANU				-0.09 (<0.01)	
N	182	160	152	176	145
% Class.	70.9%	70.6%	69.7%	67.0%	73.1%

- There is obviously a multicollinearity problem between inequality, polity, and manufacturing.
- Policy simulation: Assume Lebanon has today a per capita income of $5,000 and a Gini coefficient of 60. What is Lebanon's war country likelihood? By how much would Lebanon's level of inequality be reduced in order to drive the war country likelihood below 50%?

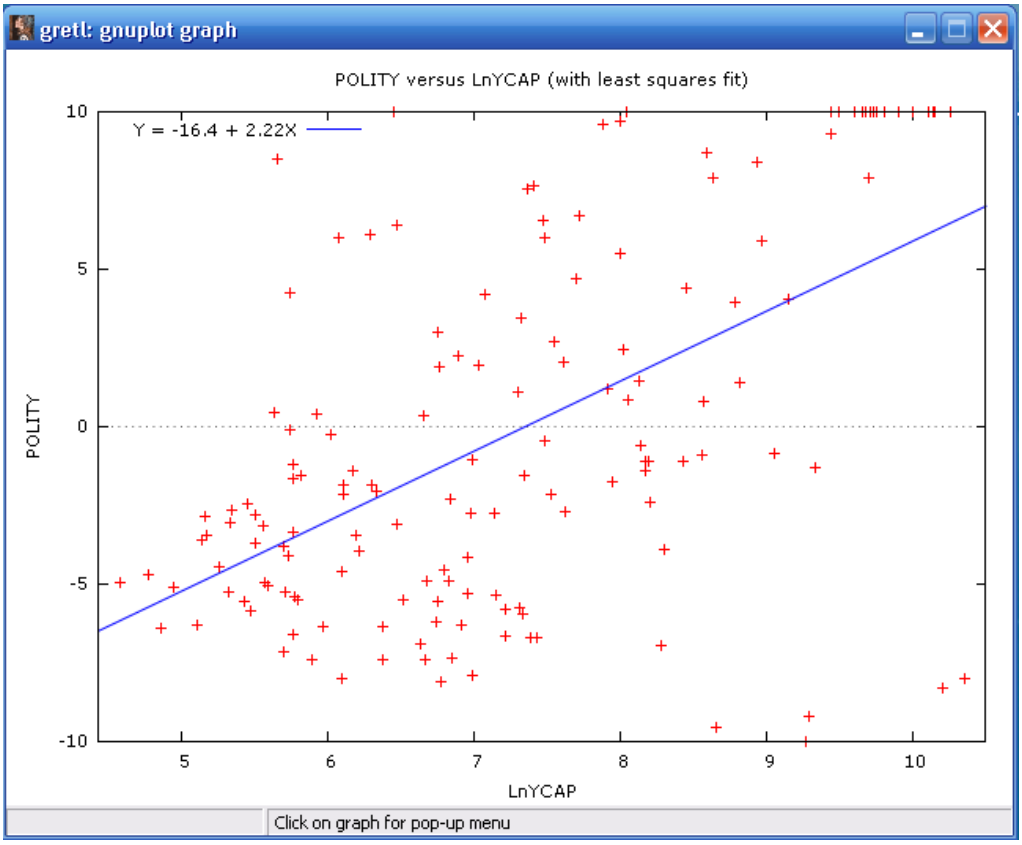
5. Truncated Data and Logistic Regression

- Logistic regression, which underlies the logit model, can also be applied to data which is "somehow cut off."
- Such cut off data is called truncated or censored.
- Truncated data causes a truncation bias, which makes the trend line "flatter" than it would be if the data was normally distributed.

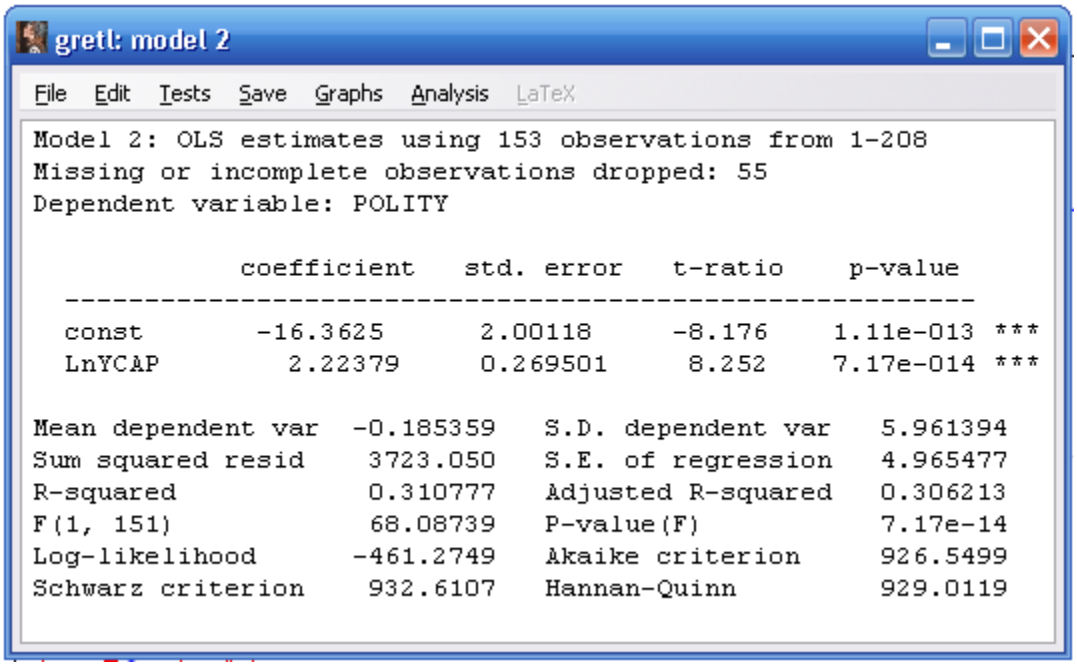


- A striking empirical regularity is that the maximum
- Dividing the OLS estimates by the proportion of nonlimit observations in the sample is a common "practitioner's solution" to correct for this bias (Greene (2003): Econometric Analysis, p. 768).
- However, an even better model would be again a non-linear fit, similar to the logit model.

- An example for truncated data is the “Polity” dataset which takes values between -10 (Autocracy) and +10 (Democracy).
- Running from the “War.xls” dataset Polity on per capita income (LnYCAP) yields the following scatter plot.
- Scatter Plot of “Polity” (truncated) on “Per Capita Income”



- Running from the "War.xls" dataset Polity on per capita income (LnYCAP) yields the following scatter plot.



- A closer look at the data reveals that out of the 153 observations, 20 observations are limit observations (reproduce this result in excel), which gives a percentage of non-limit observations of (133/153=87%).
- The truncation-adjusted coefficient would therefore be 2.22/0.87=2.55.
- Instead of running a linear regression, truncated data is always a natural candidate for logistic regression.

- When running a logistic regression on truncated data, it is necessary to transform the data first such that that all values of the dependent variable are positive.
- It is recommended to add to the dependent variable the minimum plus one, which is eleven in the case of Polity.
- In running a logistic regression with truncated data, "gretl" also will ask you to specify the asymptotic maximum, which in the case of Polity is now 21.
- For computational reasons, the asymptotic must be slightly above the maximum value, for example, 21.001
- In gretl you open the logistic regression module in Model → Nonlinear models → Logistic
- The regression results are summarized below.
- A comparison of the adjusted R^2 shows that the logistic regression is a much better fit, increasing the R^2 by almost 7 percentage points.

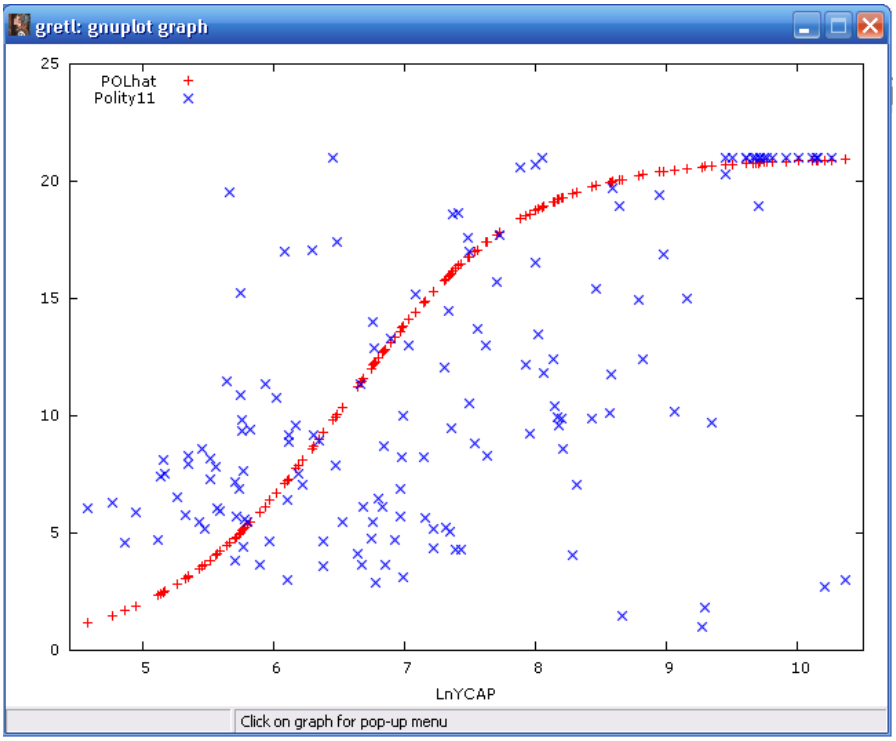
- Logistic regression results

gretl: model 3

File Edit Tests Save Graphs Analysis LaTeX

```
Model 3: Logistic estimates using 153 observations from 1-208
Missing or incomplete observations dropped: 55
Dependent variable: Polity11
yhat = 21.001 / (1 + exp(-X*b))

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       -9.52819      1.13518     -8.394    3.15e-014 ***
  LnYCAP       1.45661      0.152875     9.528    3.83e-017 ***

Statistics based on the transformed data:

Sum squared resid    1197.987   S.E. of regression   2.816681
R-squared            0.375477   Adjusted R-squared   0.371341
F(1, 151)            90.78452   P-value(F)           3.83e-17
Log-likelihood      -374.5316   Akaike criterion     753.0631
Schwarz criterion    759.1240   Hannan-Quinn         755.5251

Statistics based on the original data:

Mean dependent var   10.51256   S.D. dependent var   6.017572
Sum squared resid    6854.025   S.E. of regression   6.737277
```

- Scatter plot "Actual vs. Predicted" using logistic regression



6. Further Readings

Dummy dependent estimation techniques are nicely explained in the following texts:

Cameron, S. (2005), Econometrics, McGraw Hill, Boston, Chapter 8.

Schmidt, S., Econometrics, McGraw Hill, Boston, 2005, chapter 19.

Studenmund, A., Using Econometrics, A Practical Guide, 4th Edition, Pearson Education, 2001, Chapter 13.