Lecture 1:
Data and Models

Prof. Sharyn O'Halloran

Sustainable Development U9611

Econometrics II

Why Statistics Matters: Harpers Index

- Example 1: Logical Inconsistencies!?
  - Number of U.S. terrorism trials brought before a jury since September 11, 2001 : 1
  - Number of terrorism convictions resulting : 2
  - Number of them dismissed in June due to a "pattern of mistakes" by the prosecution : 2
- Example 2: Comparing Different Units
  - Average number of clothing items an adult American acquired in 2002 : 52
  - Estimated average amount of textiles thrown out by each U.S. household in 2001, in pounds : 66
- Lesson: Numbers don't speak for themselves.

General Approach: Data Visualization

- Standard econometric approach emphasizes:
  - Calculating the variance-covariance matrix
  - Applying “fixes” to get the right answer
- Ease of statistical programs and increased computing power emphasis has shifted to a multifaceted view of data analysis
  - Graphical presentation
  - Question driven estimation
  - Interpretation and inference

General Approach: Question Driven

- Modeling complex social science phenomena become a series of choices:
  - What is the process by which the data will be generated?
    - Random
    - Experimental
    - Observational
  - What is the appropriate estimation techniques?
    - Linear
    - Probabilistic
  - What is the scope of inference?
    - How general are the findings?
    - Has a lot to do with the research design.

Game Plan

- Examine variables individually
  - Transform variables as needed
- Examine key relations
- Identify appropriate estimation techniques
  - OLS, Probit, Logit, etc…
- Define model specification
  - Which variables to include in the analysis
  - Derived both from inspection of the data and theory
- Then run analysis
- Perform post-regression diagnostics
  - Tests for significance, graphical analysis, simulations
- Repeat!

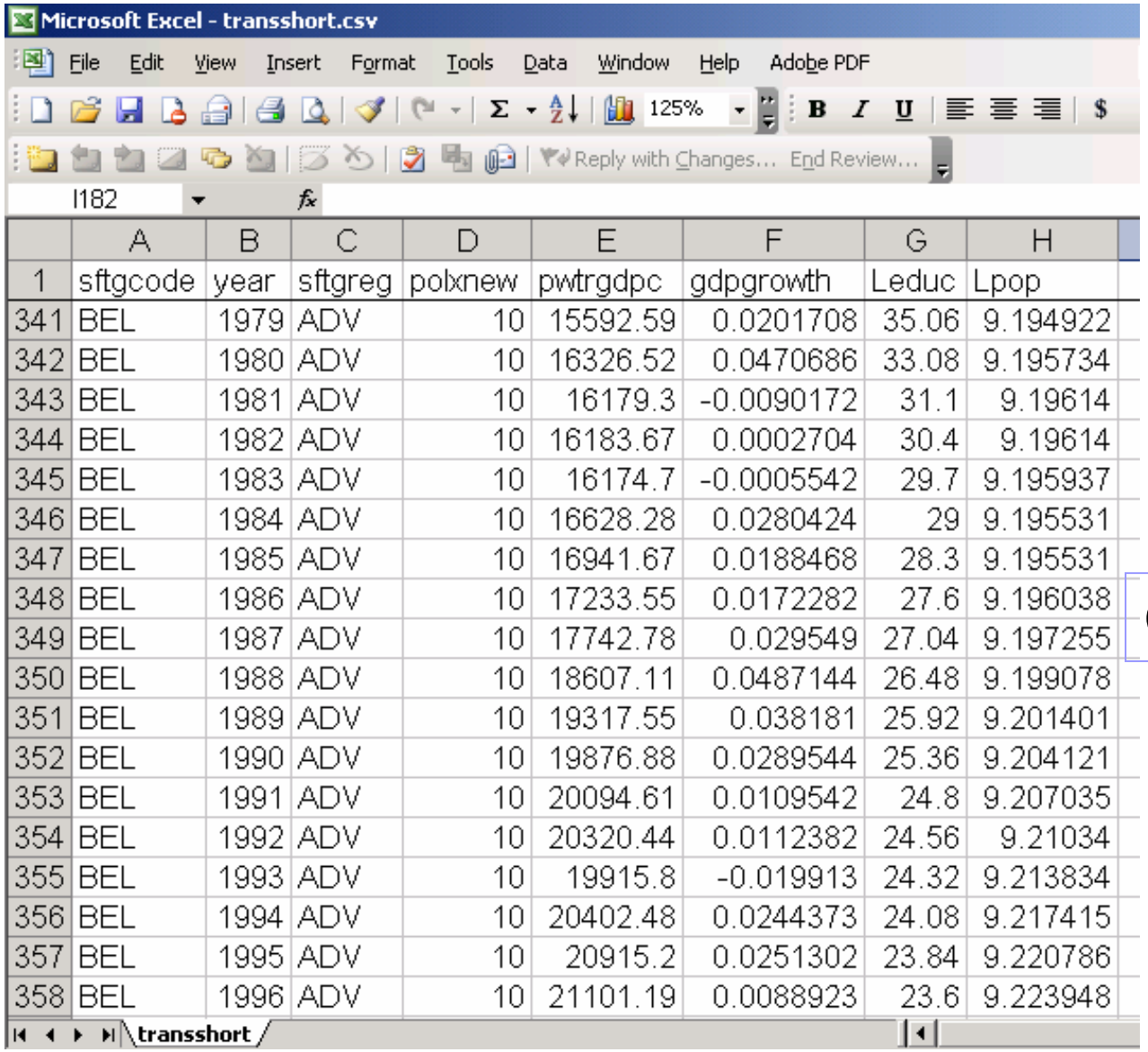
Example: Growth and Democracy

- Question: Does economic growth promote transitions to democracy?
  - Traditional answer had been **Yes**.
    - Democracy is like a luxury good
    - This is one of the classic findings in political economy
  - Recent rejoinder (PACL) says money does not predict transitions to democracy, but can help you stay there once you're rich.
    - The focus here quickly turns to political institutions such as property rights and the rule of law. (Rodrik, Shleifer et. al.)
- The importance of getting this right is more than academic.
  - A fundamental policy question is whether to promote economic or political reform first.

Example: Growth and Democracy

- How to address this debate?
  - Collect data yearly, across all countries
    - Measure of economic growth
      - GDP per capita
    - Political institution types:
      - Democracy
      - Autocracy
    - Other covariates
      - Education
      - Total Population
- This is an example of the type of analysis that we will be doing in the course.
  - And for your final paper!

Example: Growth and Democracy



	A	B	C	D	E	F	G	H
1	sftgcode	year	sftgreg	polxnew	pwtrgdpc	gdpgrowth	Leduc	Lpop
341	BEL	1979	ADV	10	15592.59	0.0201708	35.06	9.194922
342	BEL	1980	ADV	10	16326.52	0.0470686	33.08	9.195734
343	BEL	1981	ADV	10	16179.3	-0.0090172	31.1	9.19614
344	BEL	1982	ADV	10	16183.67	0.0002704	30.4	9.19614
345	BEL	1983	ADV	10	16174.7	-0.0005542	29.7	9.195937
346	BEL	1984	ADV	10	16628.28	0.0280424	29	9.195531
347	BEL	1985	ADV	10	16941.67	0.0188468	28.3	9.195531
348	BEL	1986	ADV	10	17233.55	0.0172282	27.6	9.196038
349	BEL	1987	ADV	10	17742.78	0.029549	27.04	9.197255
350	BEL	1988	ADV	10	18607.11	0.0487144	26.48	9.199078
351	BEL	1989	ADV	10	19317.55	0.038181	25.92	9.201401
352	BEL	1990	ADV	10	19876.88	0.0289544	25.36	9.204121
353	BEL	1991	ADV	10	20094.61	0.0109542	24.8	9.207035
354	BEL	1992	ADV	10	20320.44	0.0112382	24.56	9.21034
355	BEL	1993	ADV	10	19915.8	-0.019913	24.32	9.213834
356	BEL	1994	ADV	10	20402.48	0.0244373	24.08	9.217415
357	BEL	1995	ADV	10	20915.2	0.0251302	23.84	9.220786
358	BEL	1996	ADV	10	21101.19	0.0088923	23.6	9.223948

Excel sheet of data

Observational Data

Example: Growth and Democracy

transsshort.csv - Notepad

File Edit Format Help

```
sftgcode,year,sftgreg,polxnew,pwtrgdpc,gdpgrowth,Leduc,Lpop
"AFG",1960,"AS",-10,,,,
"AFG",1961,"AS",-10,,,.3,9.215427
"AFG",1962,"AS",-10,,,.28,9.237761
"AFG",1963,"AS",-10,,,.26,9.260178
"AFG",1964,"AS",-7,,,.24,9.28294
"AFG",1965,"AS",-7,,,.22,9.305923
"AFG",1966,"AS",-7,,,.2,9.329367
"AFG",1967,"AS",-7,,,.2,9.352968
"AFG",1968,"AS",-7,,,.2,9.376702
"AFG",1969,"AS",-7,,,.2,9.400713
"AFG",1970,"AS",-7,,,.2,9.425451
"AFG",1971,"AS",-7,,,.2,9.45101
"AFG",1972,"AS",-7,,,.2,9.477157
"AFG",1973,"AS",-7,,,.2,9.503234
"AFG",1974,"AS",-7,,,.2,9.528503
"AFG",1975,"AS",-7,,,.2,9.552084
"AFG",1976,"AS",-7,,,.2,9.573107
"AFG",1977,"AS",-7,,,.18,9.592332
"AFG",1978,"AS",0,,,.16,9.609385
"AFG",1979,"AS",-7,,,.14,9.622052
"AFG",1980,"AS",-7,,,.12,9.627338
"AFG",1981,"AS",-7,,,.1,9.623575
"AFG",1982,"AS",-7,,,.18,9.610123
"AFG",1983,"AS",-7,,,.26,9.588434
"AFG",1984,"AS",-7,,,.34,9.561771
"AFG",1985,"AS",-7,,,.42,9.534957
"AFG",1986,"AS",-7,,,.5,9.512443
"AFG",1987,"AS",-7,,,.52,9.494466
"AFG",1988,"AS",-7,,,.54,9.481512
```

Comma separated text, as viewed in Notepad

Insheet data file into Stata
Insheet using “fileName”

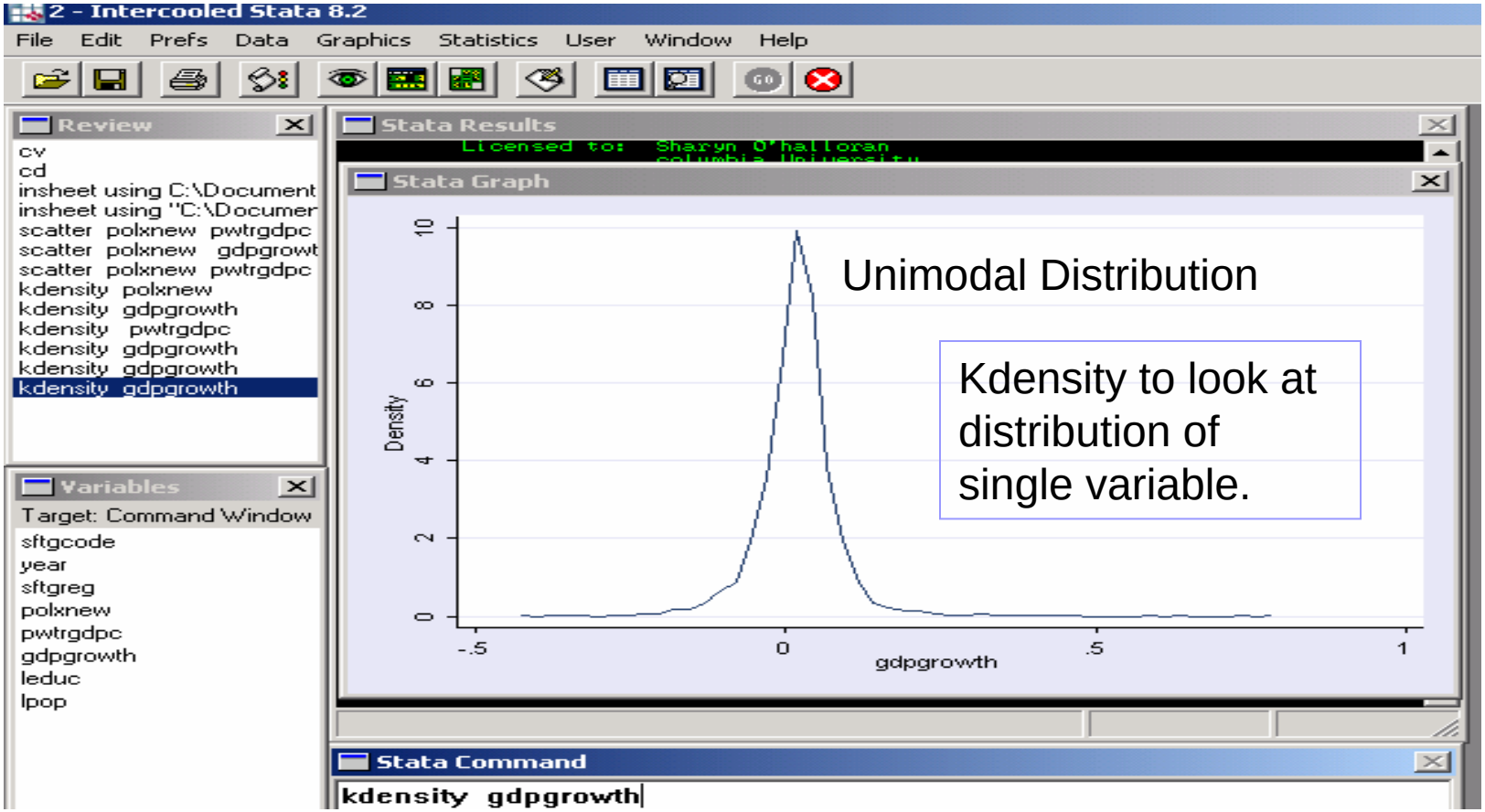
The Importance of Being Normal

- Before running any analysis, check the distributions of all key variables.
- Easiest to work with if they are normal:
  - Comparing normal distributions involves only comparing means and standard deviations
  - Some statistical procedures assume variables are normally distributed
  - Other procedures work better with normality
- Other authors used straight GDP/capita…

Example: Growth and Democracy

GDP Growth = $(GDP-GDP_{t-1})/GDP_{t-1}$

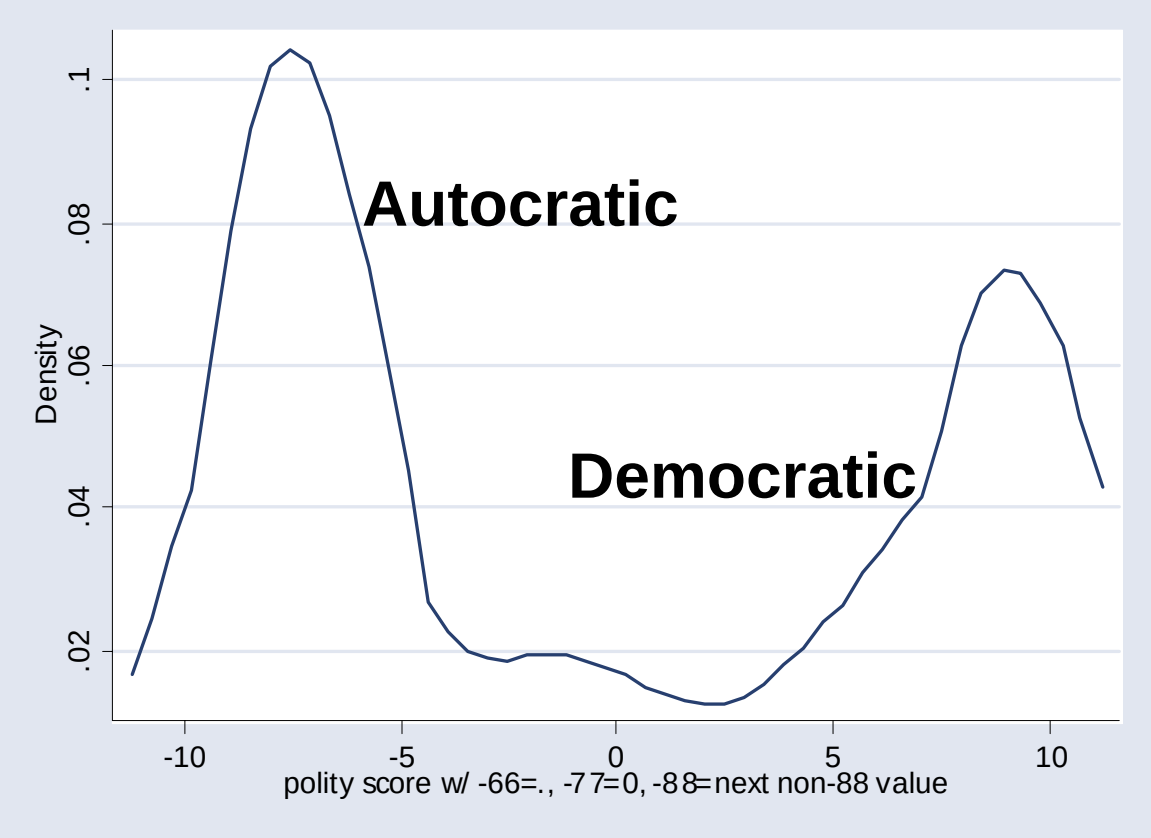
Variable	Obs	Mean	Std. Dev.	Min	Max
gdpgrowth	4269	0.0188657	0.0655765	-0.4189689	0.7769101



Example: Growth and Democracy

Polity Score = Ordinal ranking of how democratic a country is, on a -10 to 10 scale

Variable	Obs	Mean	Std. Dev.	Min	Max
polxnew	5671	-0.4456004	7.579176	-10	10

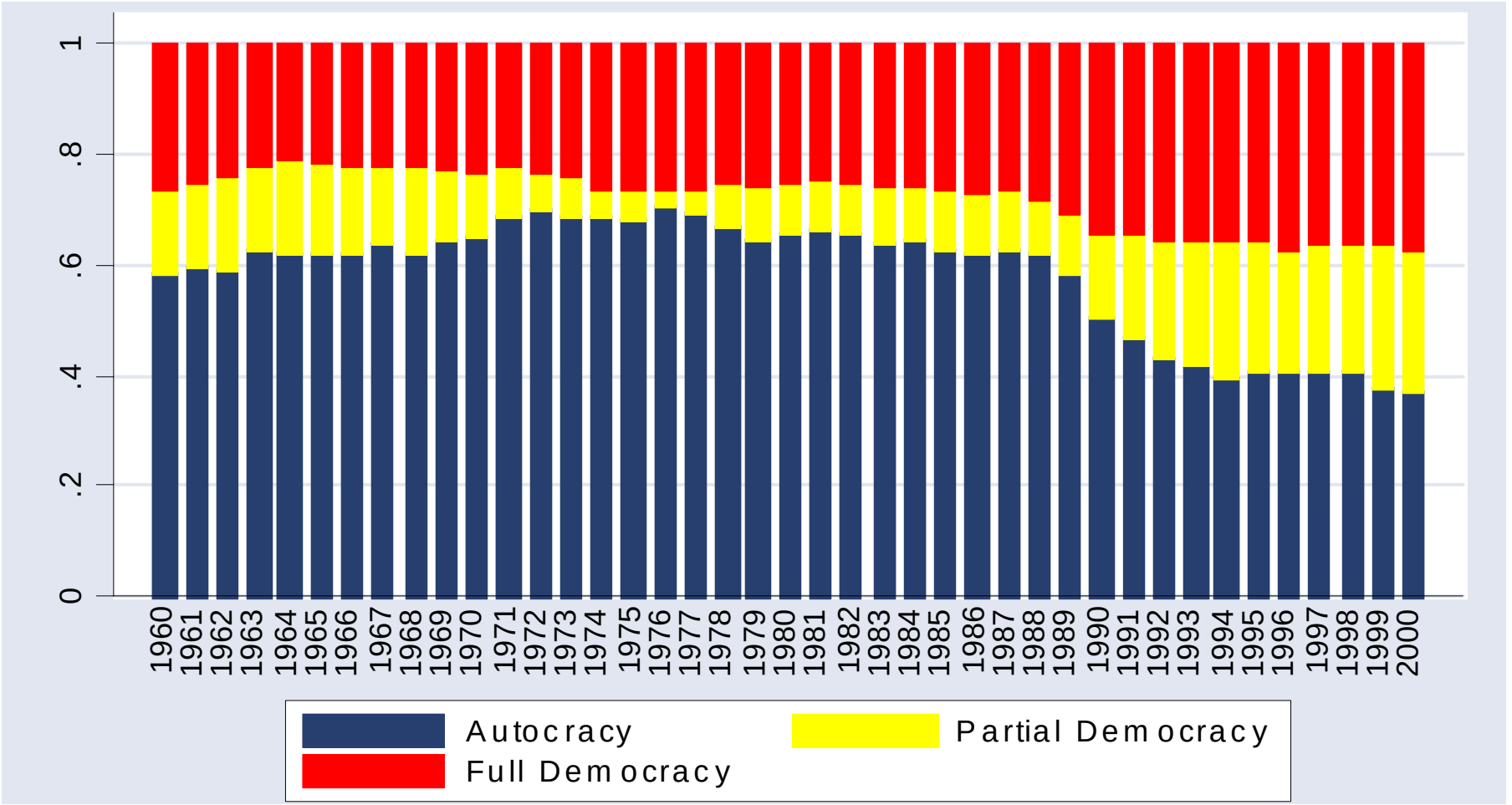


Distribution of Polity scores is bimodal.

This means that we might want to separate the data into two distinct categories.

Plus, number of partial democracies is growing…

Example: Growth and Democracy

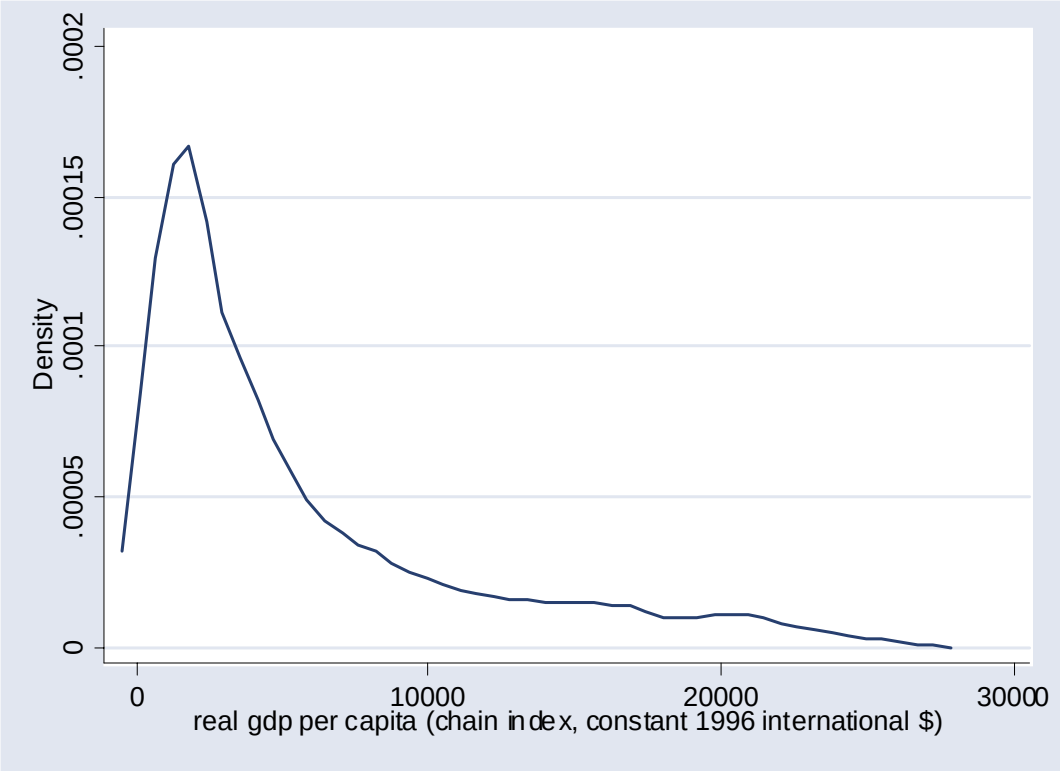


Over the last 50 years, move to democracy.

The Importance of Being Normal

Per Capita GDP = Total GDP/ Population

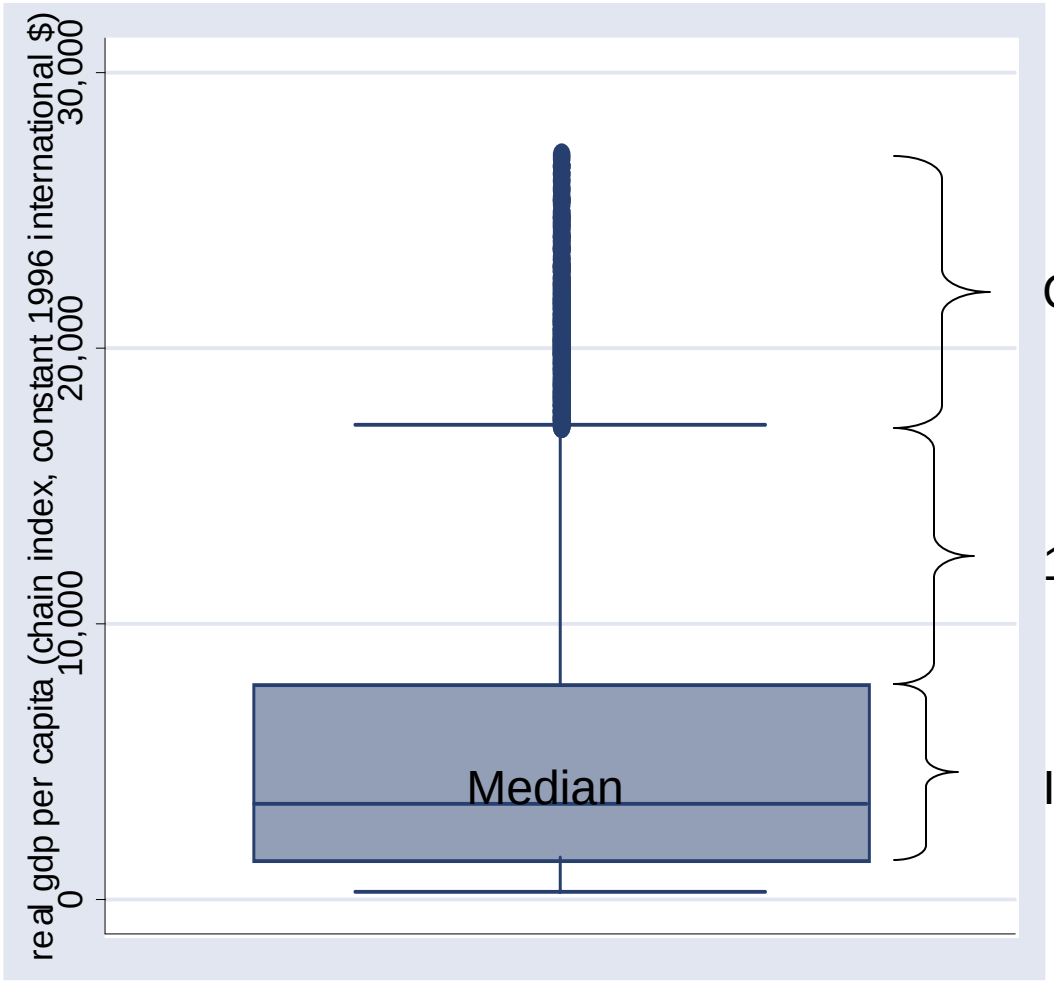
Variable	Obs	Mean	Std. Dev.	Min	Max
pwtrgdpc	4417	5713.802	5838.832	281.2581	27060.44



Per capita GDP is a unimodal but skewed distribution

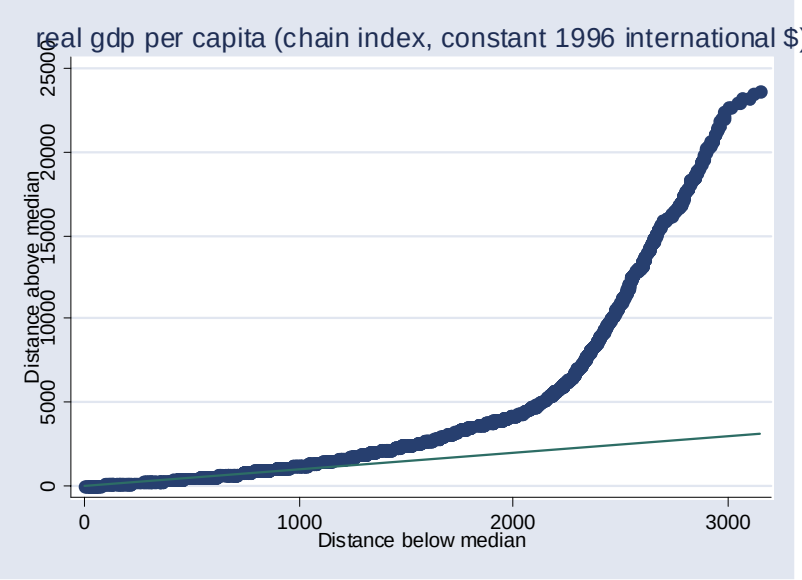
The Importance of Being Normal

Box Plot



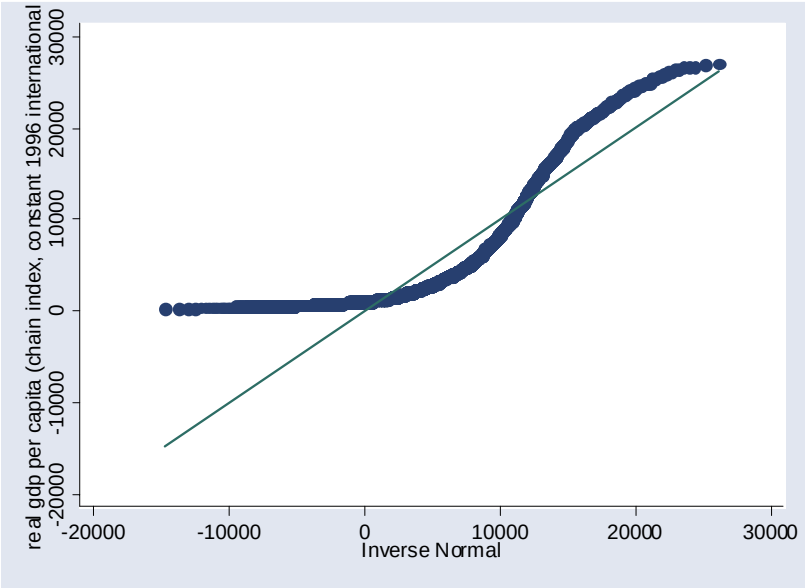
The Importance of Being Normal

Symmetry Plot



Shows that the values at the high end of the distribution are farther from the median than those at the low end.

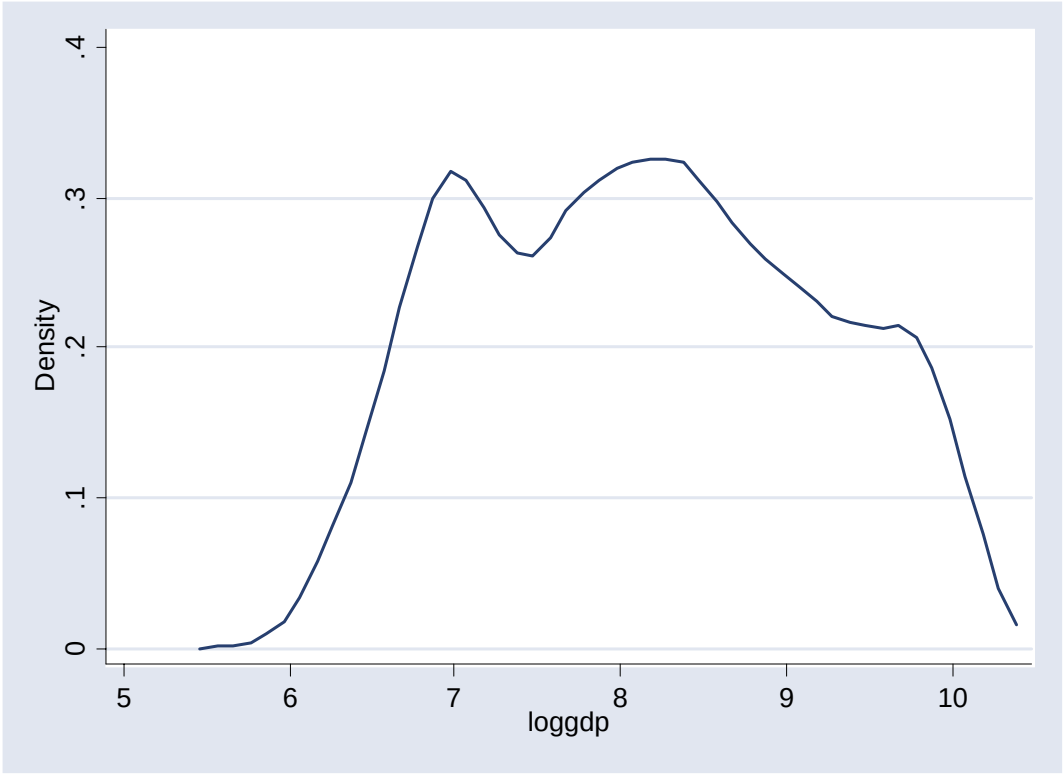
Quantile-Normal Plot



Relative to a normal distribution, there is more weight at the left tail and less in the middle.

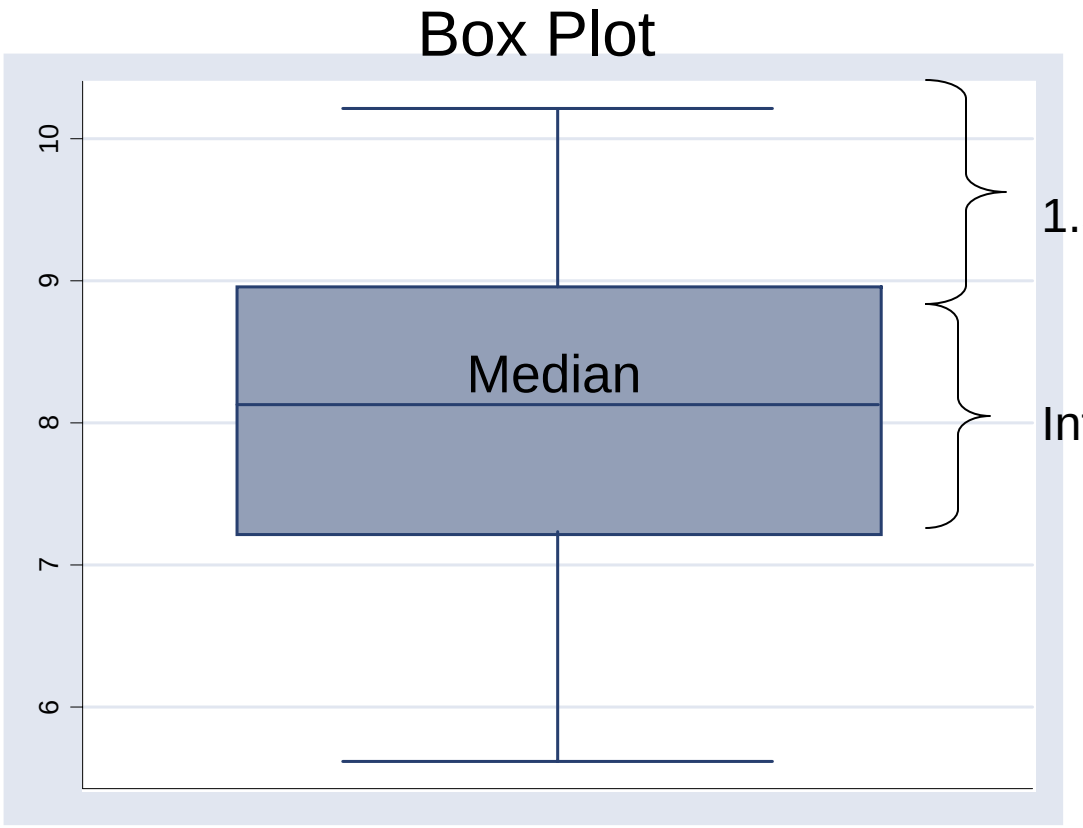
The Importance of Being Normal

Log of Per capita GDP = Log (Total GDP/ Population)



Distribution is more symmetric now, even if the tails are a little thin.

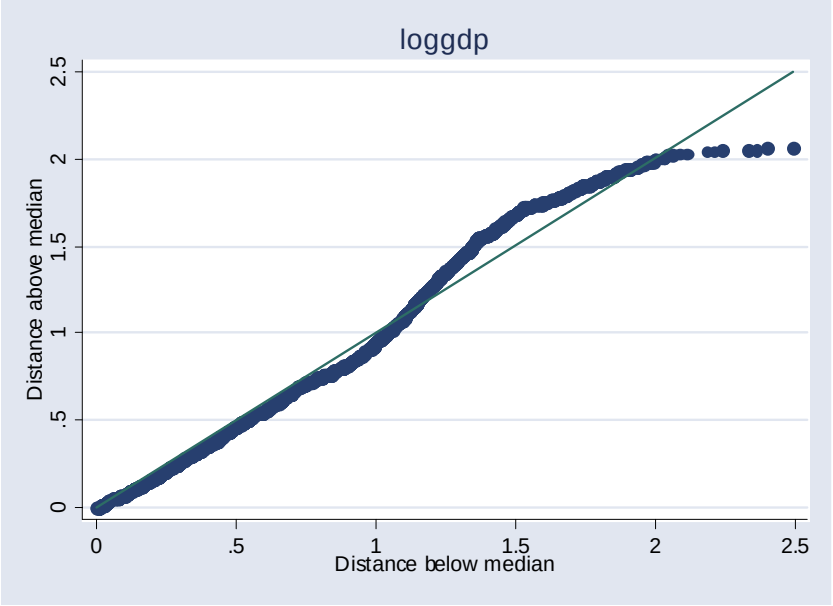
The Importance of Being Normal



Note: Much more symmetric, with no outliers.

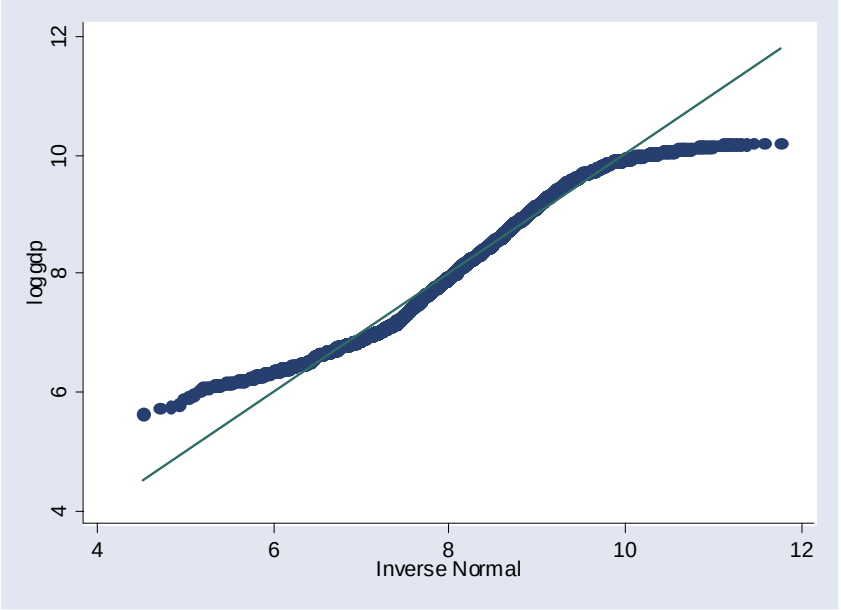
The Importance of Being Normal

Symmetry Plot



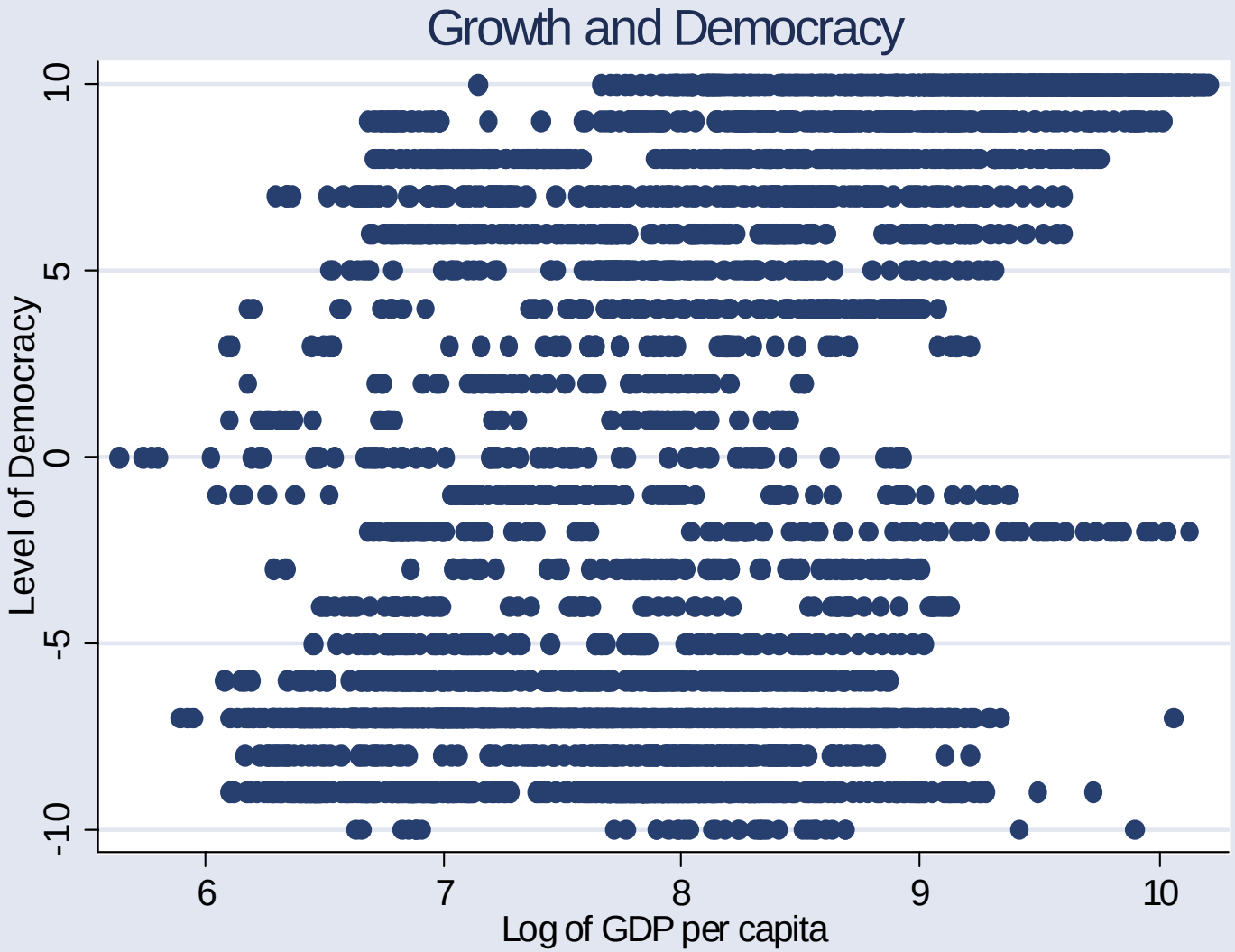
Now this follows the 45 degree line almost exactly.

Quantile-Normal Plot



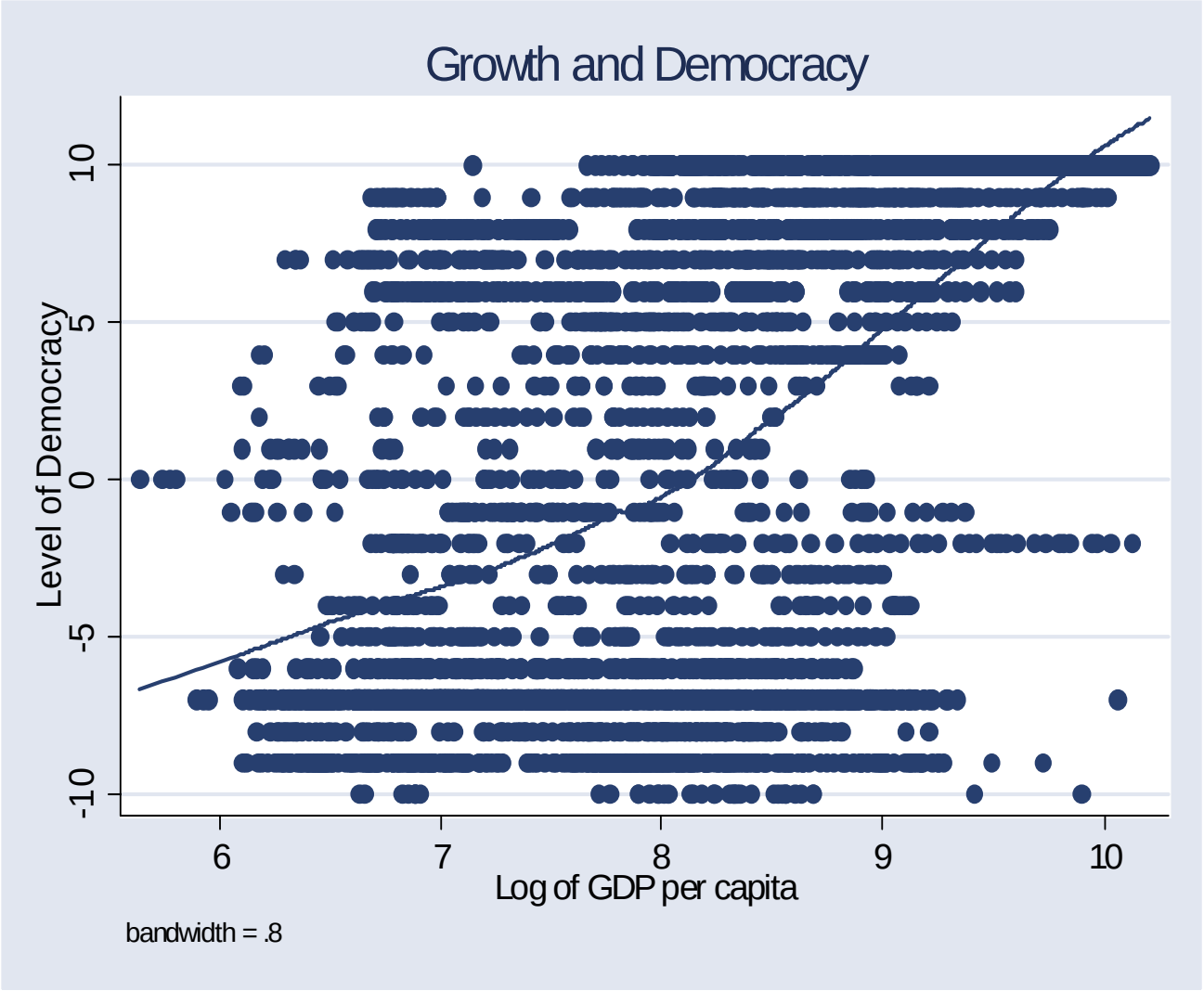
Much more similar to a normal distribution, except for the thin tails.

Inspecting Key Relationships



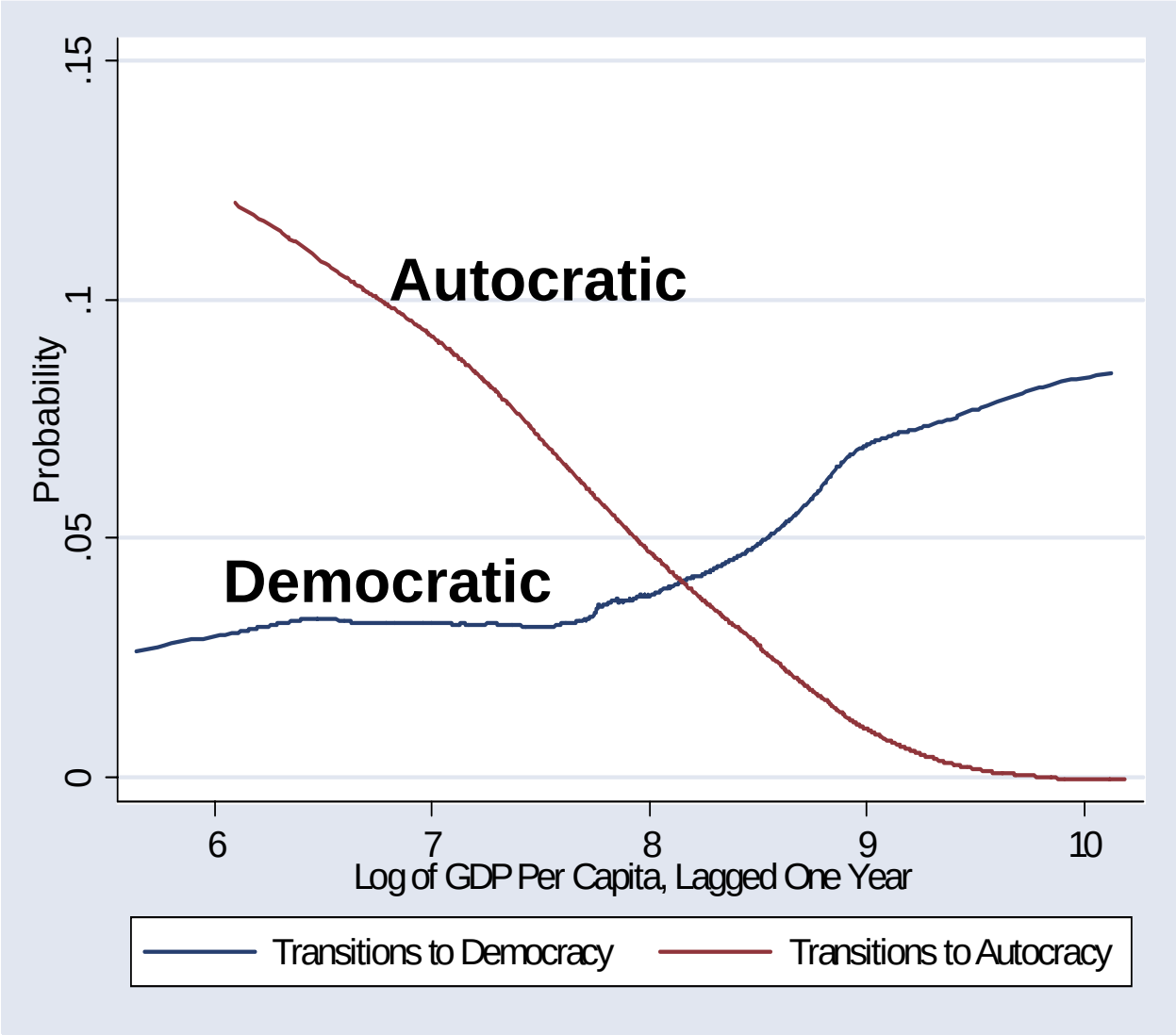
Hard to tell what's going on here.

Inspecting Key Relationships



- Add a “lowess”, or local regression line.
- A data summary technique.
- Shows a clear positive relation between the variables.

Inspecting Key Relationships



- But we should check to see if democratic and autocratic transitions act differently.
- Both show a clear impact of GDP on transition probabilities.

Estimation Techniques

- Say we decide to look at transitions:
  - Autocracy → Democracy
  - Democracy → Autocracy
- Then the dependent variable has only two values: Transition or No Transition
  - This type of "qualitative" dependent variable occurs often in social science:
    - Voting for a Republican or Democrat
    - Supreme Court Decisions overrule or uphold
    - Yea and Nay votes when passing legislation, etc…
- Appropriate estimation technique is "Probit".
  - Estimates nonlinear probabilities

Comparing Model Specifications

- PACL regress transitions on:
  - GDP per capita
  - GDP per capita squared
  - GDP growth
- Adjusting for previous regime type

anydem	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Lpdem	7.993817	7.57117	1.06	0.291	-6.845404	22.83304
Lgdp	-1.284701	1.053302	-1.22	0.223	-3.349136	.7797335
Lpdemgdp	-1.575778	1.951307	-0.81	0.419	-5.400269	2.248714
Lgdp2	.0909221	.0676323	1.34	0.179	-.0416348	.2234789
Lpdemgdp2	.1274521	.1248902	1.02	0.307	-.1173281	.3722323
Lgrowth	-.4754997	.6829151	-0.70	0.486	-1.813989	.8629892
Lpdemgrowth	-1.48283	1.18237	-1.25	0.210	-3.800234	.8345726
_cons	2.670554	4.066588	0.66	0.511	-5.299812	10.64092

Comparing Model Specifications

- PACL regress transitions on:
  - GDP per capita
  - GDP per capita squared
  - GDP growth
- Adjusting for previous regime type

anydem	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Lpdem	7.993817	7.57117	1.06	0.291	-6.845404	22.83304
Lgdp	-1.284701	1.053302	-1.22	0.223	-3.349136	.7797335
Lpdemgdp	-1.575778	1.951307	-0.81	0.419	-5.400269	2.248714
Lgdp2	.0909221	.0676323	1.34	0.179	-.0416348	.2234789
Lpdemgdp2	.1274521	.1248902	1.02	0.307	-.1173281	.3722323
Lgrowth	-.4754997	.6829151	-0.70	0.486	-1.813989	.8629892
Lpdemgrowth	-1.48283	1.18237	-1.25	0.210	-3.800234	.8345726
_cons	2.670554	4.066588	0.66	0.511	-5.299812	10.64092

Comparing Model Specifications

- Why include GDP per capita and its square as independent variables?
  - You would do this to check if a variable has a curvilinear effect.
    - For example, higher levels of incomes have a negative impact on transitions.
  - But if the impact is not significant and there is no good theoretical reason to include it, it should be dropped from the regression.

Alternative Model Specification

- So let us try the same analysis without the square term.

anydem	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Lpdem	-.0334039	.7746361	-0.04	0.966	-1.551663	1.484855
Lgdp	.1283458	.0646331	1.99	0.047	.0016672	.2550243
Lpdemgdp	.4573869	.0991983	4.61	0.000	.2629618	.651812
Lgrowth	-.4904828	.6819738	-0.72	0.472	-1.827127	.8461612
Lpdemgrowth	-1.515618	1.176133	-1.29	0.198	-3.820797	.789561
_cons	-2.756896	.4980252	-5.54	0.000	-3.733007	-1.780784

Alternative Model Specification

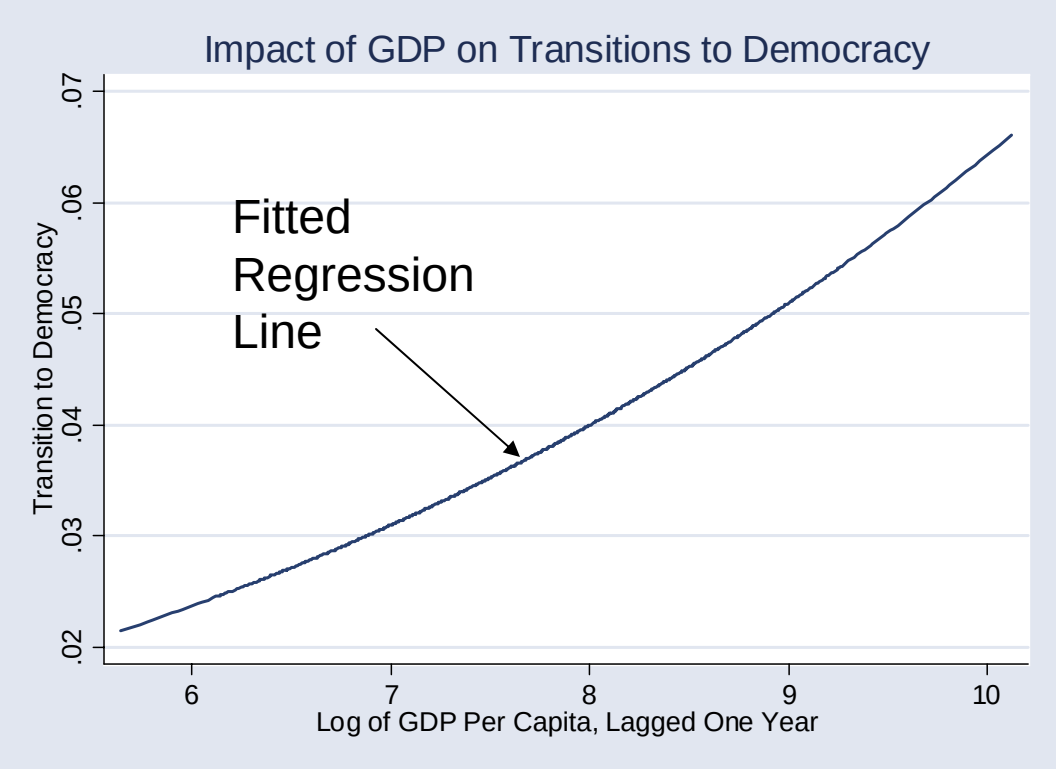
- So let us try the same analysis without the square term.

anydem	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Lpdem	-.0334039	.7746361	-0.04	0.966	-1.551663	1.484855
Lgdp	.1283458	.0646331	1.99	0.047	.0016672	.2550243
Lpdemgdp	.4573869	.0991983	4.61	0.000	.2629618	.651812
Lgrowth	-.4904828	.6819738	-0.72	0.472	-1.827127	.8461612
Lpdemgrowth	-1.515618	1.176133	-1.29	0.198	-3.820797	.789561
_cons	-2.756896	.4980252	-5.54	0.000	-3.733007	-1.780784

- It works!
- Significant results with other covariates added, as well (population, education, etc.)

Regression Diagnostics

- So GDP predicts democracy
  - But does this really have a substantive impact?

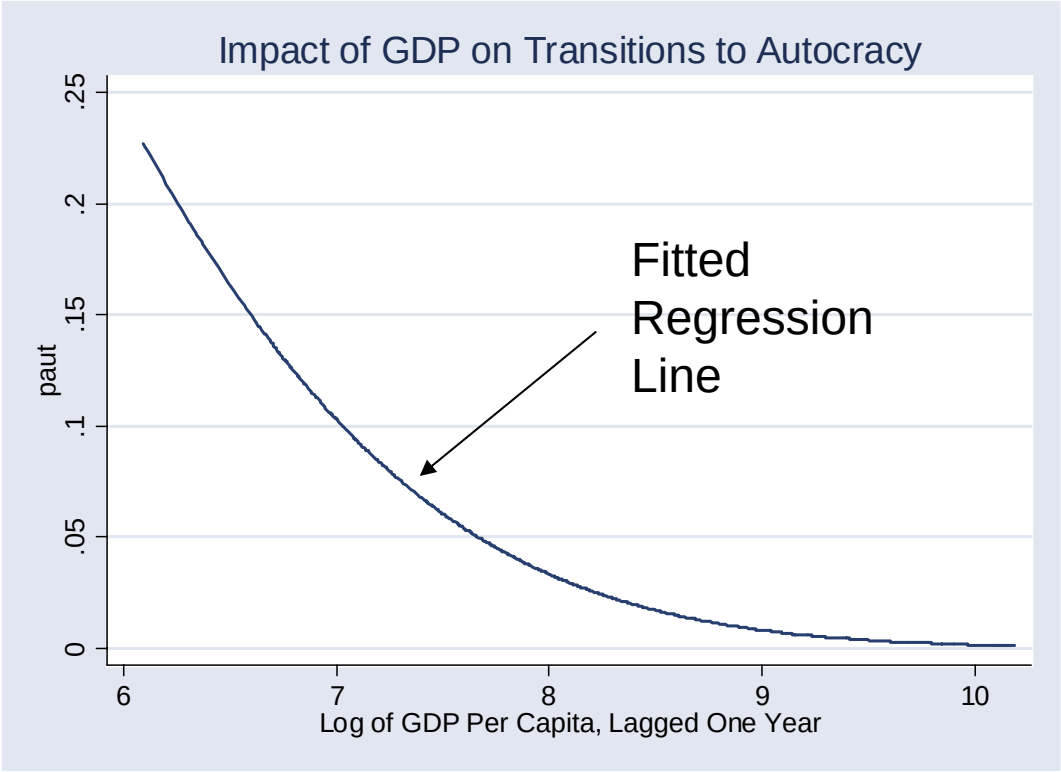


Raising GDP from its minimum to maximum value increases the probability of a democratic transition from 2% to 6.5%.

This decreases the expected life of an autocratic regime from 34 years to just 10 years.

Regression Diagnostics

- So GDP predicts Democracy,
  - But does this really have a substantive impact?



Raising GDP from its minimum to maximum value decreases the probability of an autocratic transition from 23% to (essentially) 0%.

The End (Or the Beginning)

- What else could you do with this analysis?
  - Add more covariates
    - Education
    - Population
    - Resource Curse
  - Treat data differently
    - Use entire democracy-autocracy scale, rather than dividing it into discrete categories
    - Treat this as a survival problem
  - Others?

Class Organization

- Text: *Statistical Sleuth*
- Website: CourseWorks
- Grades:
  - Homework: 35%
  - Midterm: 25%
  - Final Paper and Presentation: 35%
  - Participation: 5%