EC310: Econometrics Part I
The Nature of Econometrics and Economic Data

Junhui Qian

September 14, 2014

Outline

- Basics of Econometrics
- Economic Data
- How to Obtain Data?
- Data Pre-processing

What is econometrics?

- Econometrics is a statistical approach to economic problems.
- Statistics provides the basic tools and economics provides the motivation.
- Basic objectives of econometrics:
  - Estimation of economic models
  - Inference (or hypothesis testing) on models and causal relationships
  - Forecasting based on economic or econometric models

Econometric Models

- Applied researchers use econometrics in estimating models, testing hypotheses on the models, and making forecasts based on the models. The models econometricians work on are called econometric models.
  - An econometric model describes the data generating processes (DGP) of relevant economic variables in an economic model.
  - An economic model may be deterministic, eg, Keynes' consumption function $C = C(Y)$, but an econometric model is always stochastic.
  - Parametric v.s. nonparametric v.s. semiparametric

From Economic Model to Econometric Model

- Formal economic modeling is sometimes the starting point for empirical studies, but it is more common to arrive at econometric models directly by economic "intuition".
- Econometric studies verify or reject economic models. Economic modeling accommodates empirical evidence and drives further econometric studies.
- Theoretical econometrics develops tools (estimation, hypothesis testing, forecasting) for applied economists.

Example: Job Training and Labor Productivity

A labor economist would like to examine the "treatment effects" of job training on worker productivity. A first problem he faces is that productivity is not observable. But the labor economist reasons that the productivity of a worker may be measured by his wage, which is paid commensurate with productivity at least on average. Also, economic intuition tells him that factors such as education and experience may also affect a worker's productivity. So he comes up with the following econometric model,

$$wage_i = \beta_0 + \beta_1 T_i + \beta_2 edu_i + \beta_3 expr_i + u_i,$$

where T_i denotes time of training for i-th individual and the constant β_1 is supposed to measure the effect of training. The error term u_i contains all other factors that may affect worker's productivity/wage.

Example: Estimation, Hypothesis Testing, and Prediction

- We use data to estimate the parameters.
- Then we conduct hypothesis testing. For example, we may conjecture that the job training is a waste of time. In other words, our null hypothesis is

$$H_0 : \beta_1 = 0.$$

  If the estimate of β_1 is close "enough" to 0, then we can not reject our hypothesis.
- How close is close enough? We need to know the distribution of $\hat{\beta}_1$, the estimator of β_1.
- With estimated parameters available, we can predict how much wage increase a person would obtain after he complete a certain level of job training.

What is economic data?

- Economic data are observed values of economic variables such as wages, prices, interest rates, and exchange rates, GDP, CPI, etc.
- More generally, economic data include all data that can be used in economic analysis. For example, in a well known study conducted by Steven D. Levitt (Freakonomics), the crime rate of the United States is economic data.
- Economic data are generally "nonexperimental".
- The structure of economic data
  - Cross-section data
  - Time series
  - Panel data

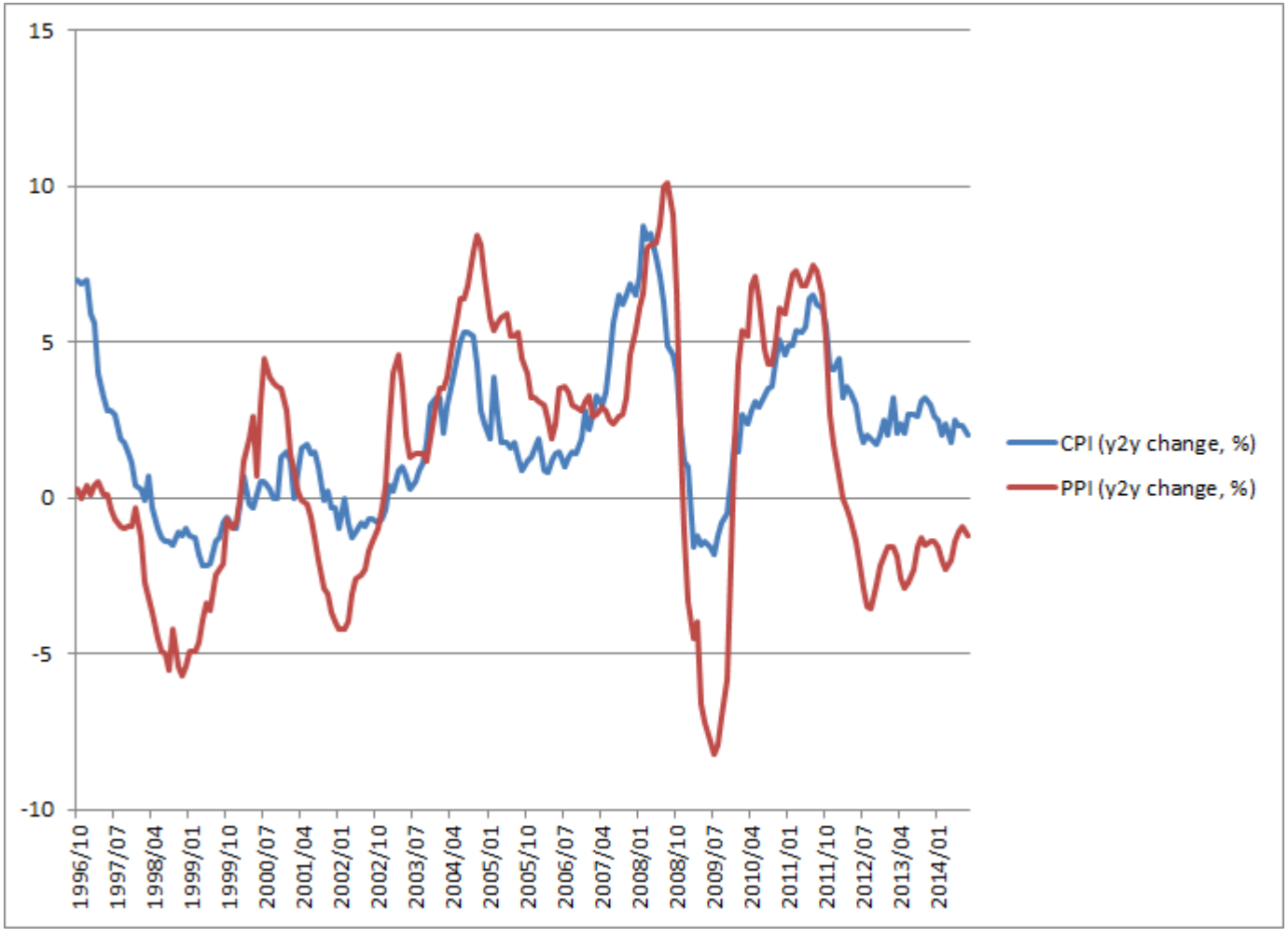
Cross-section data

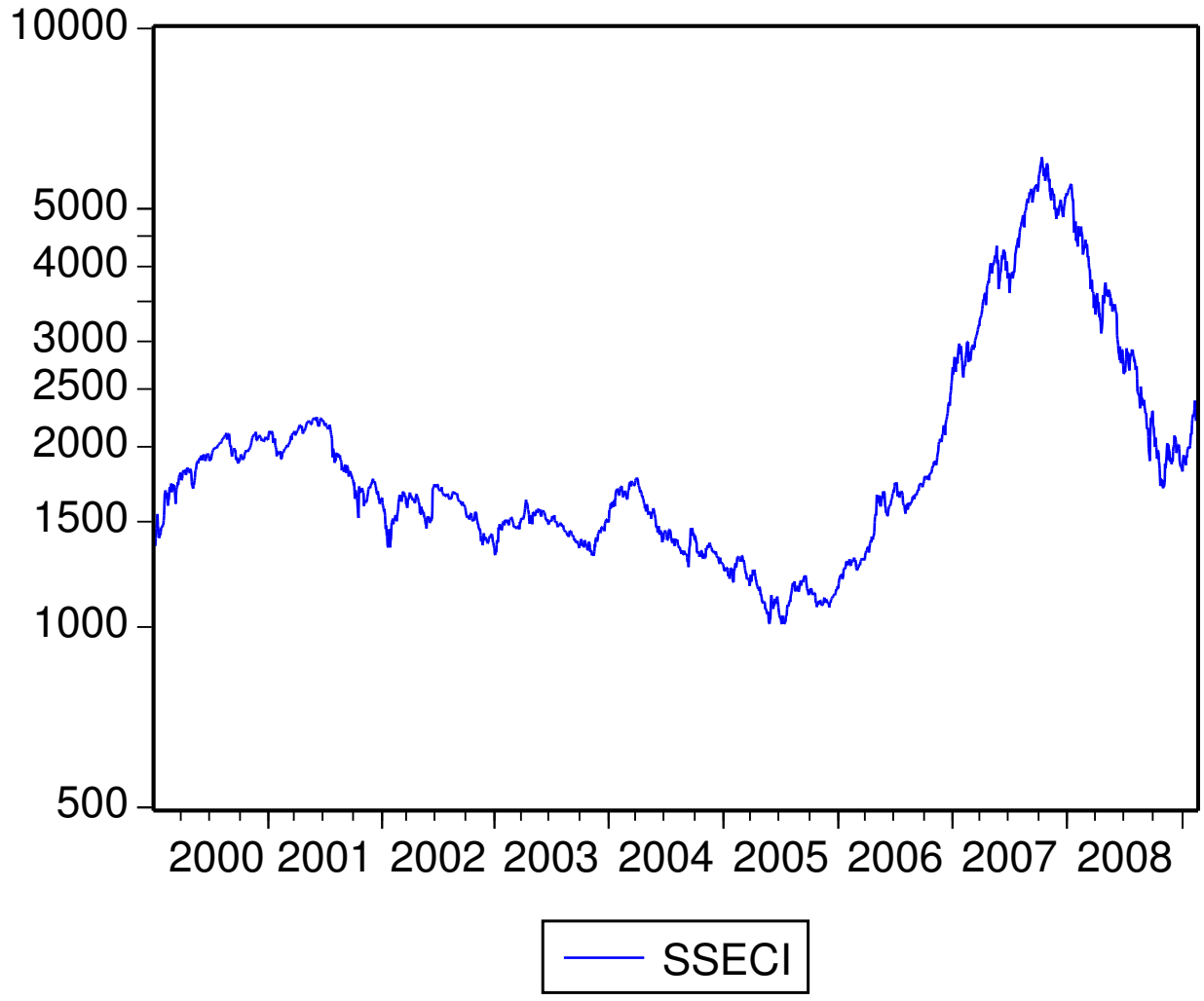
- Cross-section data consists of data vectors observed from a sample of individuals, households, firms, cities, provinces, countries, or other units.
- Timing differences in collecting data are considered miner and are ignored.
- The essential feature of cross-section data is that they can be treated as a random sample from a population.
- This implies that, observations obtained from each unit (individual, household, etc.) can be assumed to iid (identically independently distributed).
- This also implies that the sequential order of observations does not matter.
- Two common violations:
  - Sample selection problem
  - Nonindependent units

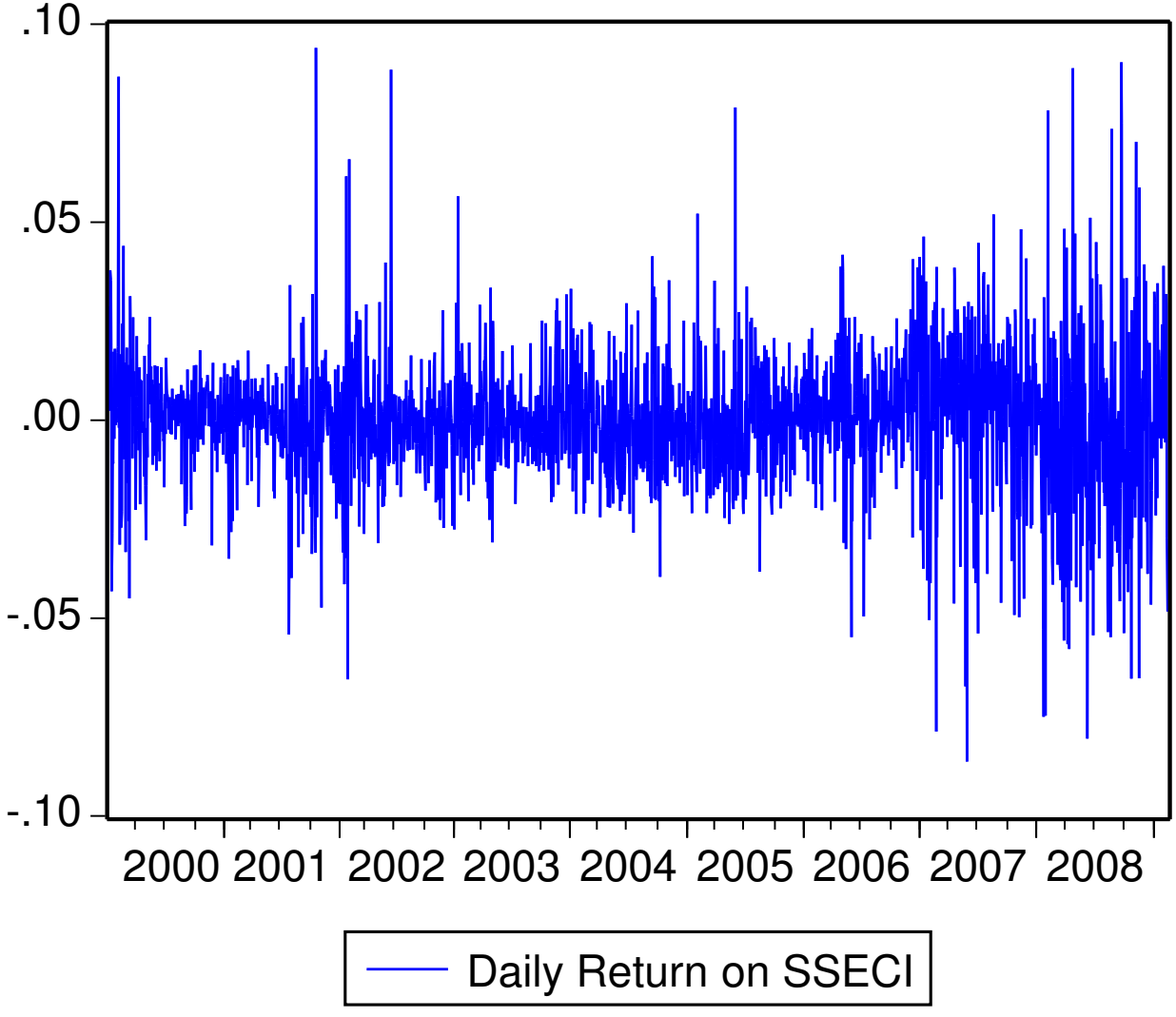
Time series data

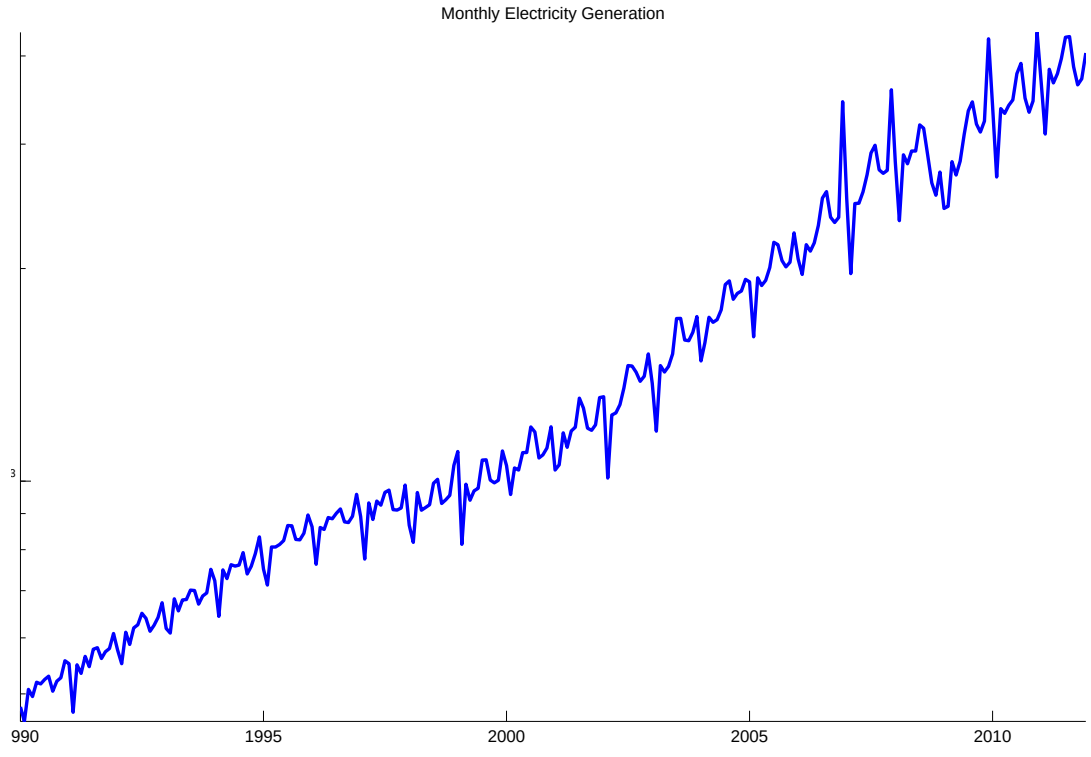
- A time series consists of scalar or vectors observed over time. For example, stock prices, dividends of a listed company, GDP, CPI, sales of TV's, etc.
- A time series is usually nonindependent. Hence the order of observations is crucial.
- Important features of time series data:
  - Frequency
  - Seasonality
  - Trends
- See examples of time series, daily stock indices, monthly CPI, quarterly GDP.


Chinese GDP Growth (%)
16
14
12
10
8
6
4
2
0
1992/03
1992/11
1993/07
1994/03
1994/11
1995/07
1996/03
1996/11
1997/07
1998/03
1998/11
1999/07
2000/03
2000/11
2001/07
2002/03
2002/11
2003/07
2004/03
2004/11
2005/07
2006/03
2006/11
2007/07
2008/03
2008/11
2009/07
2010/03
2010/11
2011/07
2012/03
2012/11
2013/07
2014/03


15
10
5
0
-5
-10
1996/10
1997/07
1998/04
1999/01
1999/10
2000/07
2001/04
2002/01
2002/10
2003/07
2004/04
2005/01
2005/10
2006/07
2007/04
2008/01
2008/10
2009/07
2010/04
2011/01
2011/10
2012/07
2013/04
2014/01
CPI (y2y change, %)
PPI (y2y change, %)


10000
5000
4000
3000
2500
2000
1500
1000
500
2000 2001 2002 2003 2004 2005 2006 2007 2008
SSECI


.10
.05
.00
-.05
-.10
2000
2001
2002
2003
2004
2005
2006
2007
2008
Daily Return on SSECI


Monthly Electricity Generation
3
990
1995
2000
2005
2010

Panel data

- A panel data (also called longitudinal data) collects data from a group of individual (households, firms, cities, provinces, countries, or other units) over a period of time.
- See an example of panel data.
- For each unit, a time series data set is collected. At a particular time point, a cross-section data set is available.
- Data collection is difficult.

Browse, Borrow or Buy

- Many macro and industrial data are available online or from 统计年鉴.
- Financial data are available from commercial database such as WIND.
- Data on individuals, households, and enterprises are difficult to obtain.
  - Very limited government disclosure.
  - The academia under-invests in data collection and database development.
- A few survey databases openly accessible:
  - 中国健康与营养调查
    http://www.cpc.unc.edu/projects/china
  - 中国社会调查开放数据库
    http://www.cssod.org/news01.php
  - China Family Panel Studies (CFPS)
    http://www.isss.edu.cn/cfps/EN/

Conduct a Survey

- Three steps:
  - Questionnaire construction
  - Sample Selection
  - Data collection
- Each step calls for thorough thought and professional training.

Feel the Data

- Browse the data, looking for abnormal numbers.
- Calculate summary statistics
- Distributions of variables
- Relationship between variables

Summary Statistics

- Number of observations
- Minimum
- Maximum
- Mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- Variance (or standard deviation)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Summary Statistics

- Skewness, measure of symmetry, or lack of symmetry. If the skewness is negative (positive), it is called "skewed left" ("skewed right"). Skewing left means that the left tail is long relative to the right tail.

$$\text{Skewness}(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{(n-1)s^3}$$

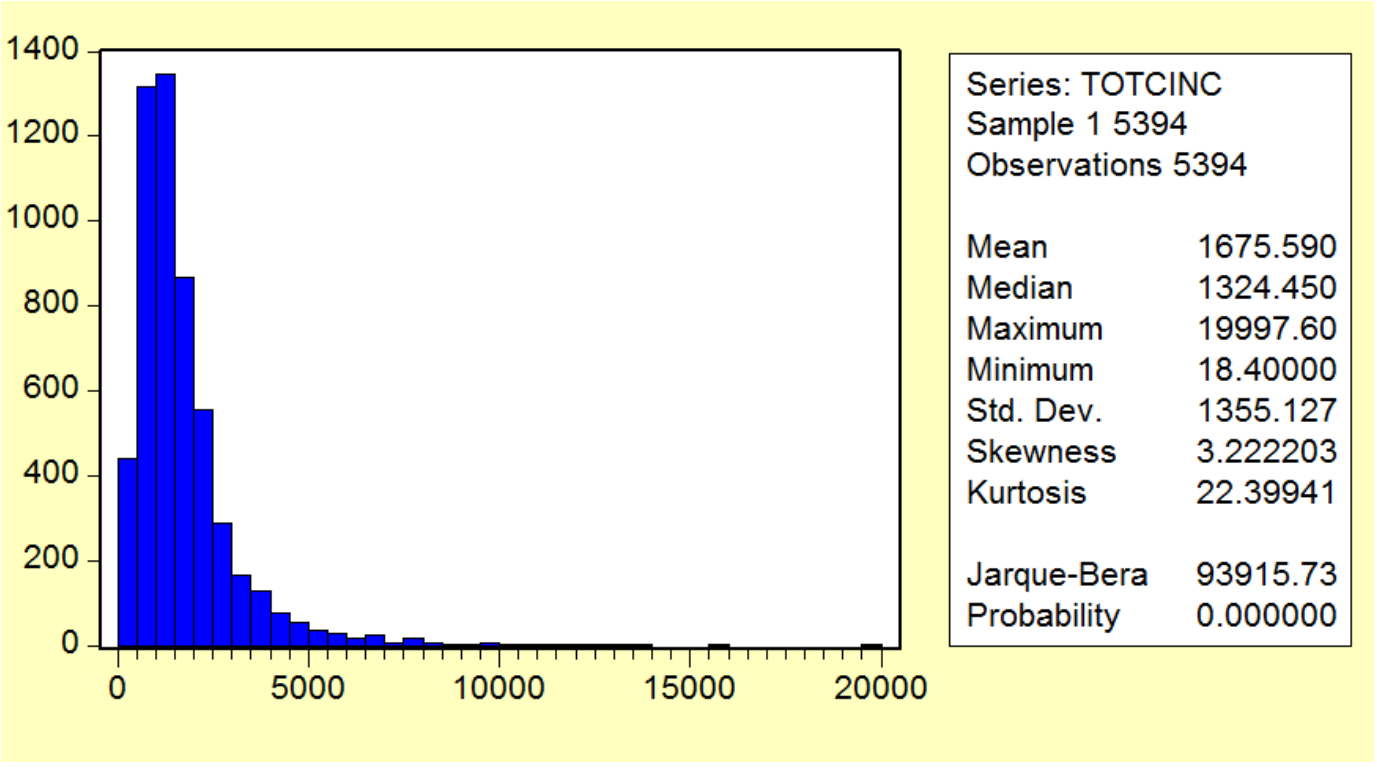
- Kurtosis, measure of peakedness relative to normal distribution. A kurtosis of higher than 0 indicates a "peaked" and heavy-tail distribution.

$$\text{Kurtosis}(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{(n-1)s^4} - 3$$

- α-Quantiles ($\alpha = 25\%, 50\%, 75\%$). 50%-quantile is called median.

$$P(X < Q(\alpha)) = \alpha$$

Summary Statistics: An Example



Distribution Plots

- Frequency distribution (or frequency table)
- Histogram
- Probability density function
  - It is obtained by kernel smoothing.
- Q-Q plot
  - A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate).
  - If the points are on a straight line, then the two distributions are linearly related.
  - If the points are on the 45° line, then the two distributions are identical.

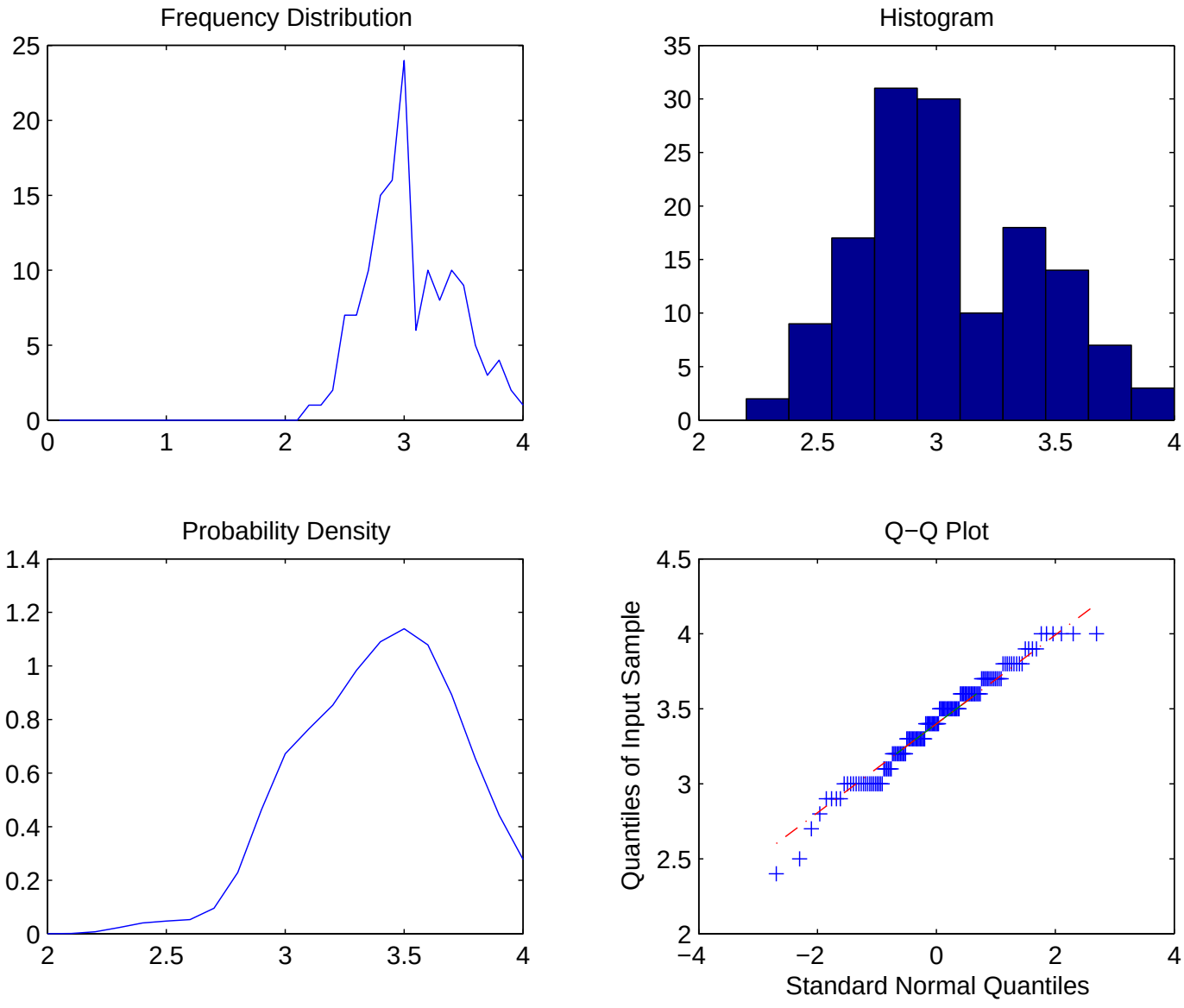


Figure : Representing Distributions

Data Exploration: Bivariate Relationship

- X-Y diagram
- Covariance

$$\mathrm{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

- Correlation coefficient

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{s(X)s(Y)}$$

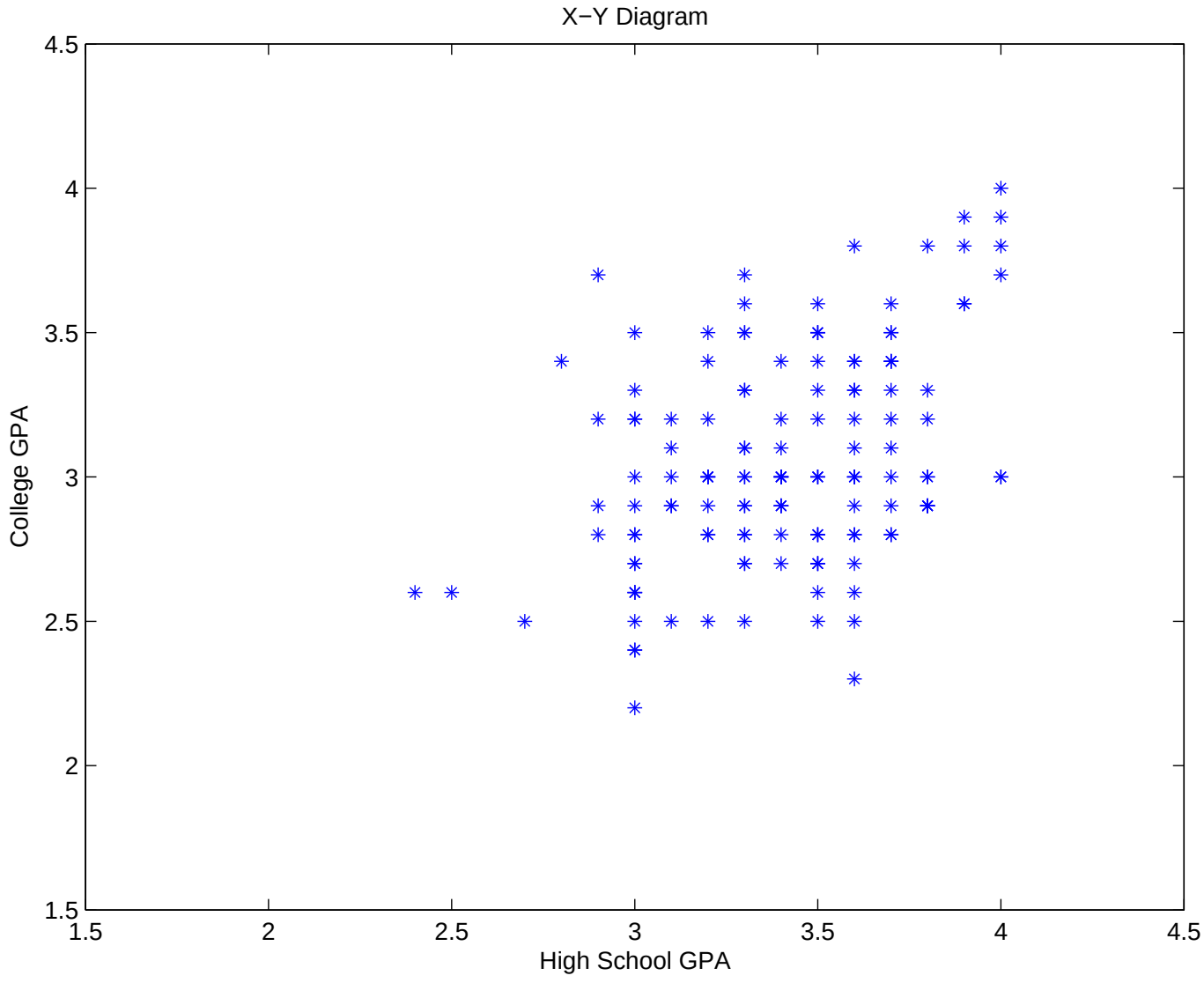


Figure : X-Y Diagram

Data Exploration: Time Series

- For time series data, time plot reveals much information, and
- Auto-covariance of order k for X_t,

$$\mathrm{cov}(X_t, X_{t-k}) = \frac{1}{n-k} \sum_{i=k+1}^{n} (X_i - \bar{X})(X_{i-k} - \bar{X})$$

- Correlogram: Autocorrelation of order k

$$\rho(k) = \frac{\mathrm{cov}(X_t, X_{t-k})}{s^2(X)}$$

- A plot of $\rho(k)$ reveals how a time series depends on its history.

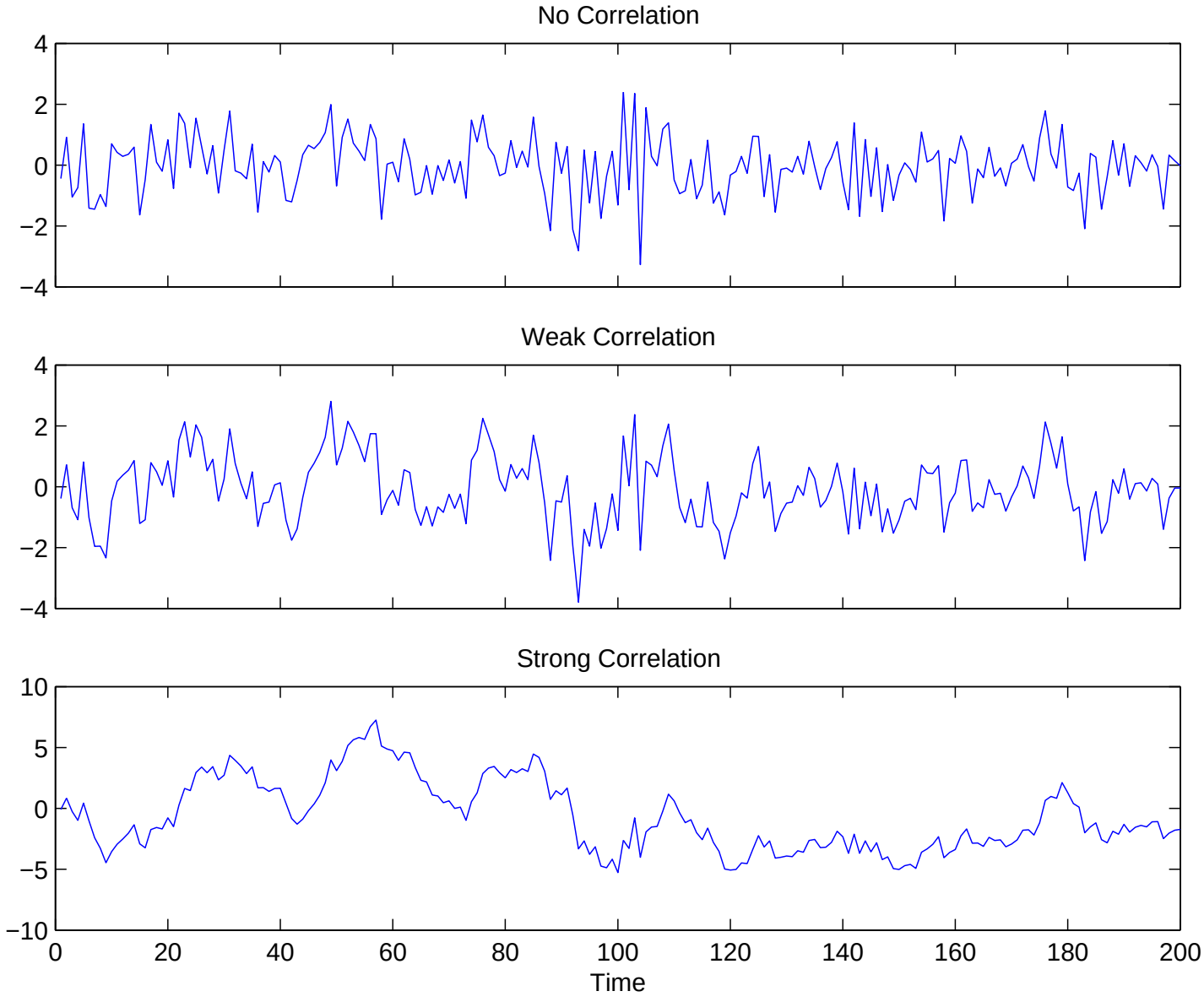


Figure : Residuals over time: $u_t = \alpha u_{t-1} + \varepsilon_t$, $\alpha = 0, 0.5, 0.95$, from top to bottom.

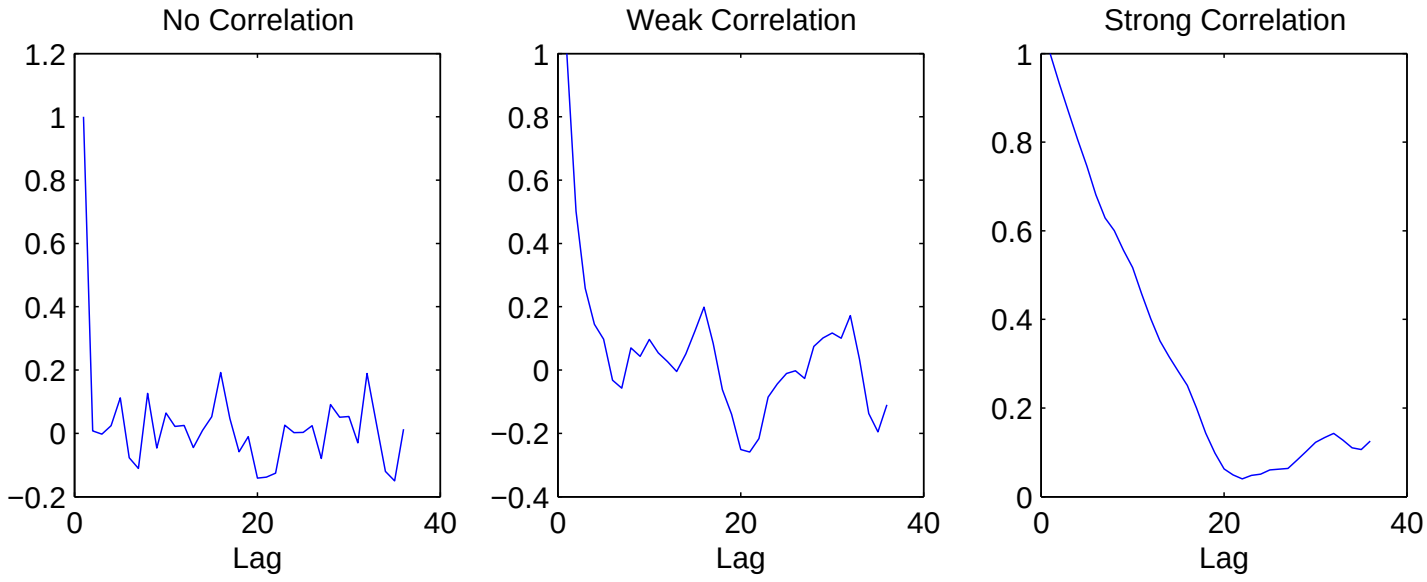


Figure : Correlograms: $u_t = \alpha u_{t-1} + \varepsilon_t$, $\alpha = 0, 0.5, 0.95$, from left to right.

Summary

- Econometrics is something dealing with economic models using nonexperimental economic data.
- This "something" includes theorizing, intuition, trial and error, theoretical proofs, etc.
- The farm of econometrics is data, which include cross-section data, time series data, panel data, etc.
- The first step of econometric investigation is to prepare data and explore the data.