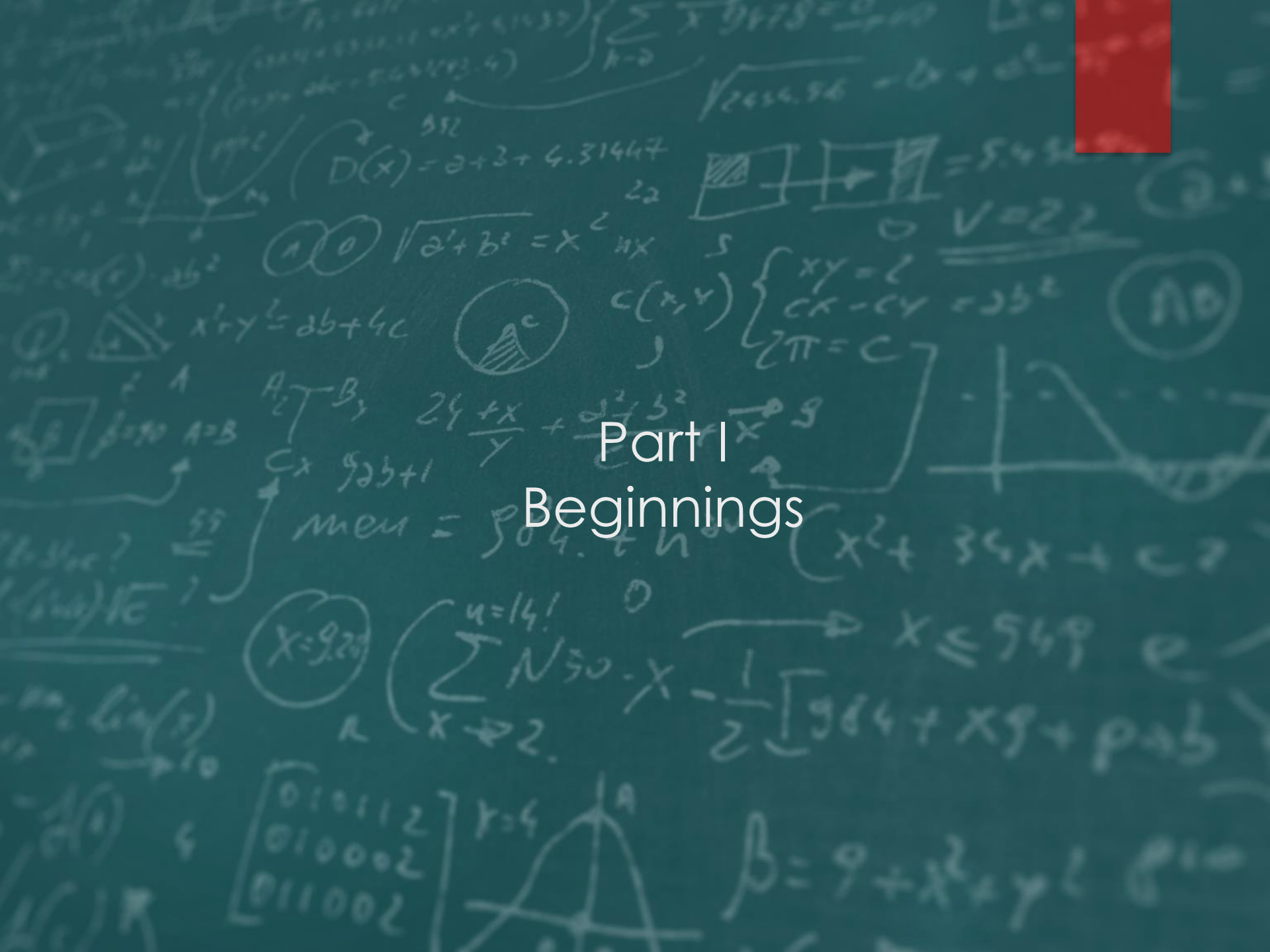
Econometric Data Science



Part I
Beginnings

Numerous Communities Use Econometrics

Economists, statisticians, analysts, ”data scientists” in:

- Finance (Commercial banking, retail banking, investment banking, insurance, asset management, real estate, ...)
- Traditional Industry (manufacturing, services, advertising, brick-and-mortar retailing, ...)
- e-Industry (Google, Amazon, eBay, Uber, Microsoft, ...)
- Consulting (financial services, litigation support, ...)
- Government (treasury, agriculture, environment, commerce,...)
- Central Banks and International Organizations (FED, IMF, World Bank, OECD, BIS, ECB, ...)

Econometrics is Special

Economists, statisticians, analysts, ”data scientists” in:

Econometrics is not just “statistics using economic data”. Many properties and nuances of economic data require knowledge of economics for successful analysis.

- Econometrics has special focus on prediction. In many respects the goal of econometrics is to help agents (consumers, firms, investors, policy makers, ...) make better decisions, and good forecasts are key inputs to good decisions.
- Econometrics must confront the special issues and features that arise routinely in economic data, such as heteroskedasticity and serial correlation.
- Econometrics must confront the special problems arising due to its largely non-experimental nature: Model mis-specification, structural change, etc.

Types of Recorded Economic Data

1. continuous or binary.

Continuous data take values on a continuum, as for example with GDP growth, which in principle can take any value in the real numbers.

Binary data take just two values, as with a 0-1 indicator for whether or not someone purchased a particular product during the last month.

2. Over time, over space, or some combination of the two.

Time series data are recorded over time, as for example with U.S. GDP, which is measured once per quarter. A GDP dataset might contain quarterly data for, say, 1960 to the present.

- Cross section: Standard cross-section notation: $i = 1, \ldots, N$
- Time series: Standard time-series notation: $t = 1, \ldots, T$

A Few Leading Econometrics Web Data Resources (Clickable)

Indispensable:

- Resources for Economists (AEA)
- FRED (Federal Reserve Economic Data)

More specialized:

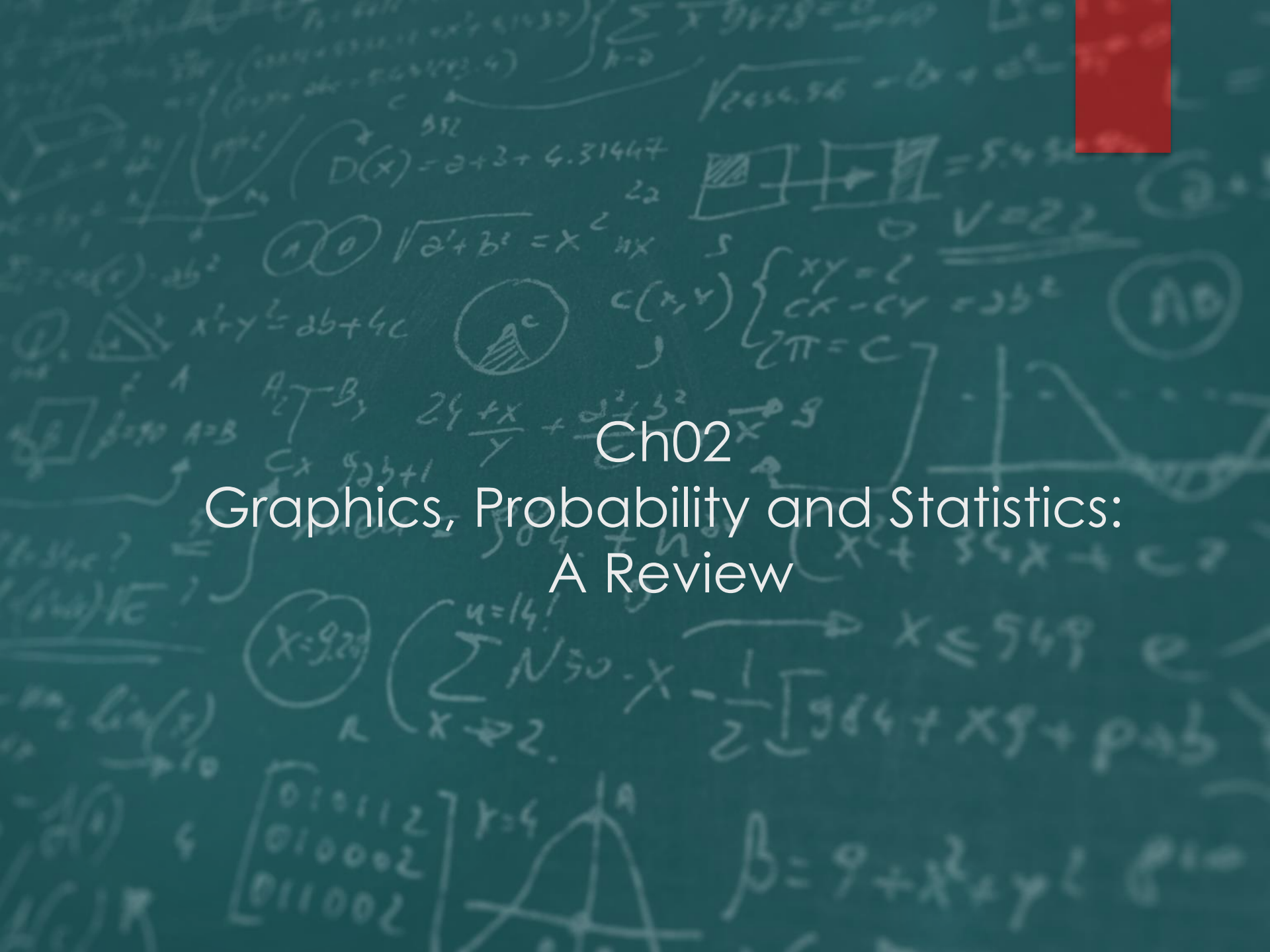
- National Bureau of Economic Research
- FRB Phila Real-Time Data Research Center
- Many more

Software

Econometric software tools are widely available. Two good and time-honored high-level environments with extensive capabilities are Stata and EViews.

Stata has particular strength in cross sections and panels, and Eviews has particular strength in time series. Both reflect a balance of generality and specialization well-suited to the sorts of tasks that will concern us.

There are also many flexible and more open-ended "mid-level" environments in which you can quickly program, evaluate, and apply new tools and techniques. R is one popular such environment, with special strengths in modern statistical methods and graphical data analysis. Other notable and increasingly-popular environments include Python and Julia.



Ch02
Graphics, Probability and Statistics: A Review

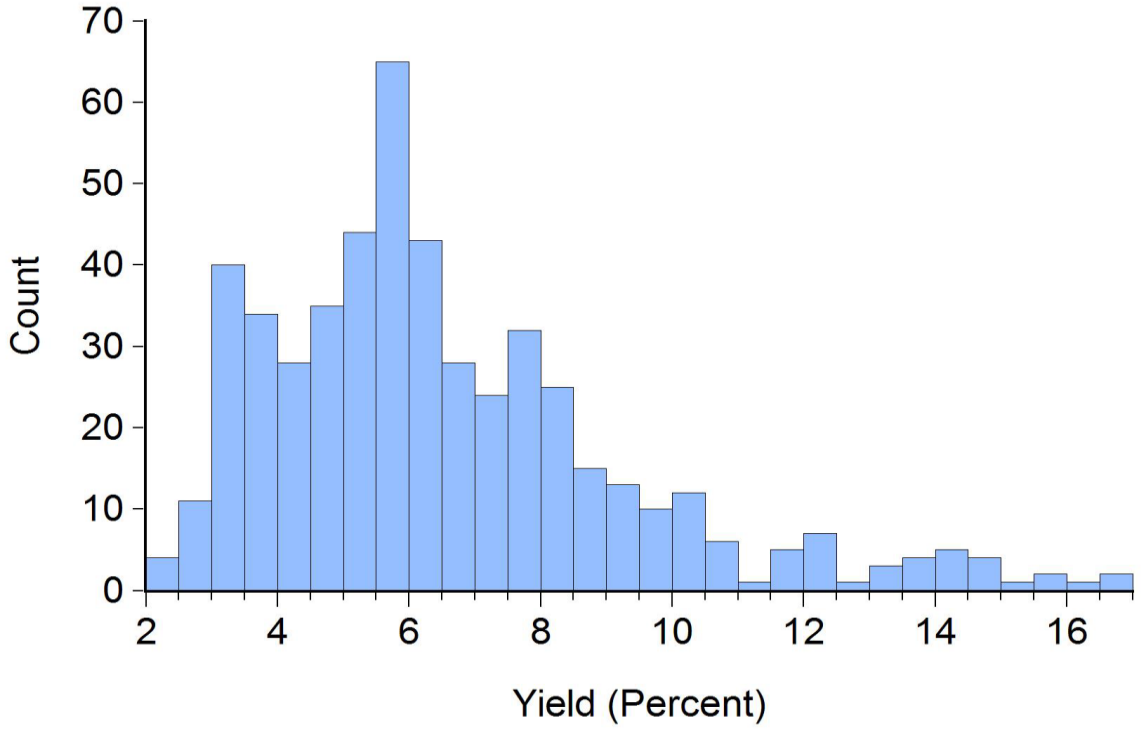
Graphics Review

It's almost always a good idea to begin an econometric analysis with graphical data analysis.

Graphics Help us to:

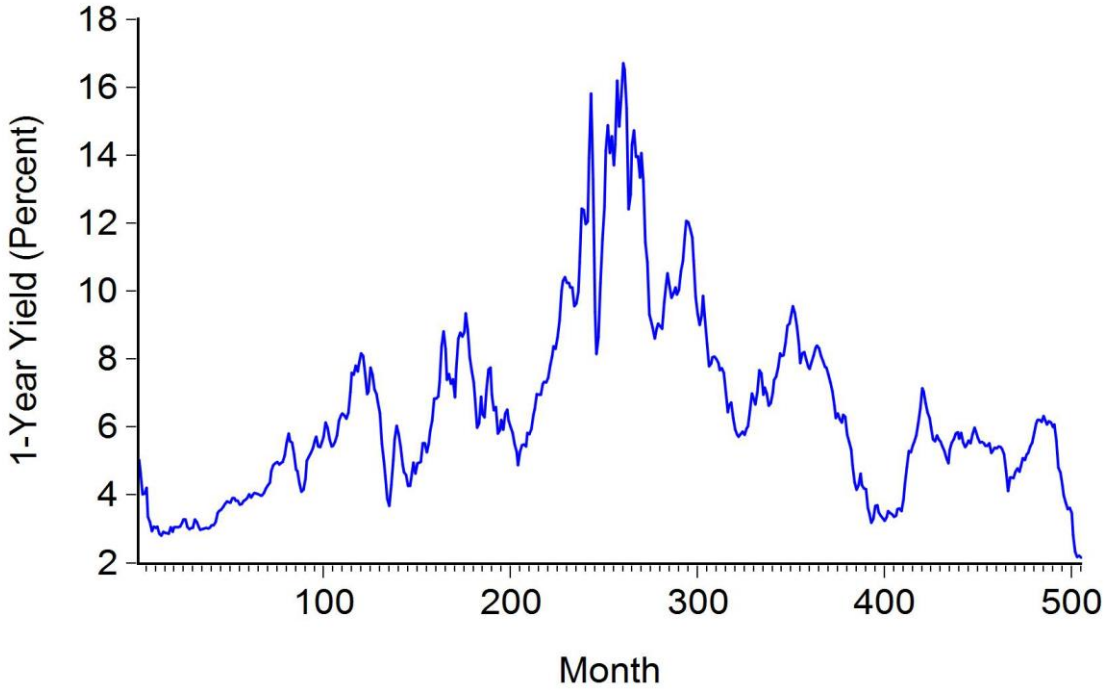
- Summarize and reveal patterns in univariate cross-section data. Histograms and density estimates are helpful for learning about distributional shape. Symmetric, skewed, fat-tailed, ...
- Summarize and reveal patterns in univariate time-series data. Time Series plots are useful for learning about dynamics. Trend, seasonal, cycle, outliers, ...
- Summarize and reveal patterns in multivariate data (cross-section or time-series). Scatterplots are useful for learning about relationships. Does a relationship exist? Is it linear or nonlinear? Are there outliers?
- Graphics facilitates and encourages comparison of different pieces of data via multiple comparisons. The scatterplot matrix is a classic example of a multiple comparison graphic.

Graphics Review



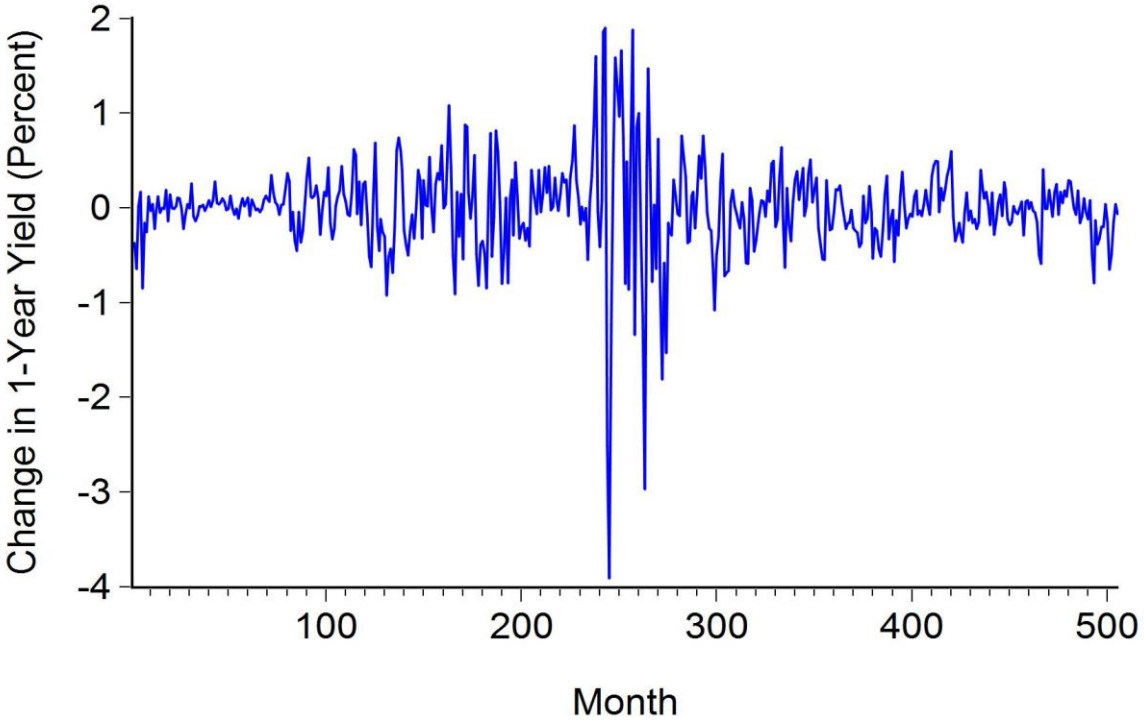
A histogram for the 1-year bond yield. This provides a simple estimate of the probability density of a random variable. The observed range of variation of the series is split into a number of segments of equal length, and the height of the bar placed at a segment is the percentage of observations falling in that segment.

Graphics Review



A time series plot of a 1-year Government bond yield over approximately 500 months. A number of important features of the series are apparent. Among other things, its movements appear sluggish and persistent, it appears to trend gently upward until roughly the middle of the sample, and it appears to trend gently downward thereafter.

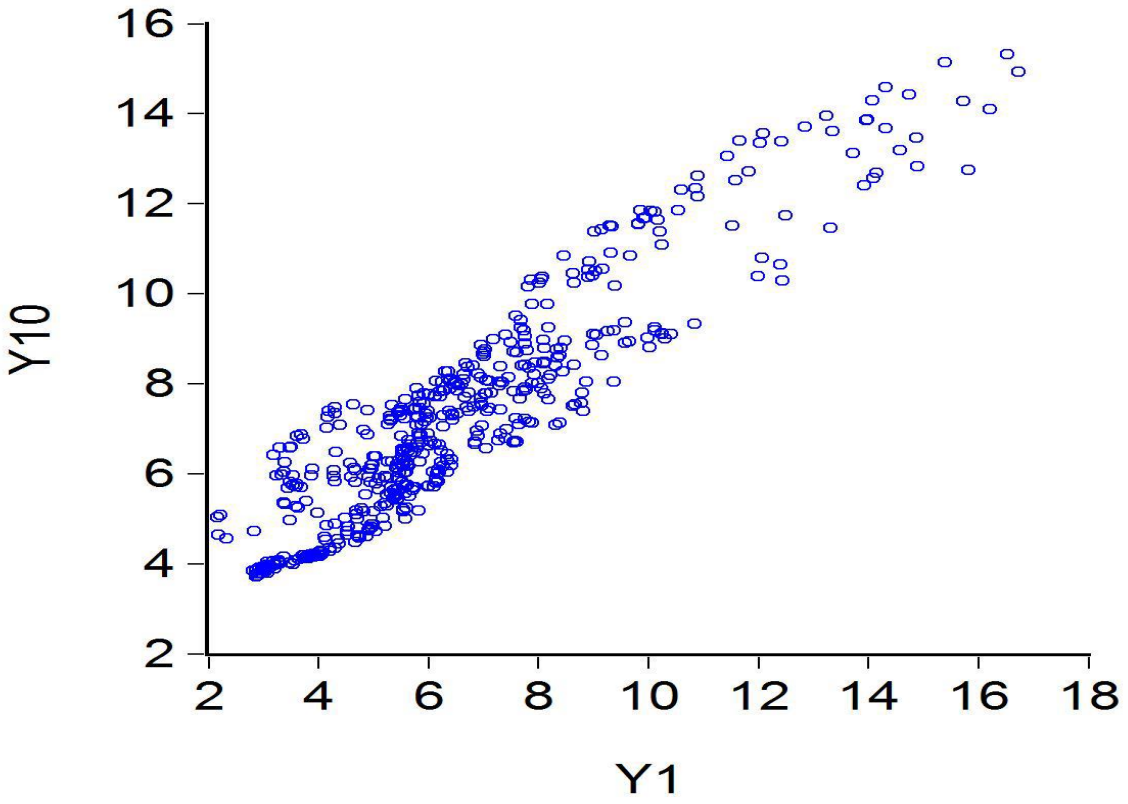
Graphics Review



A time series plot of the change in the 1-year bond yield, which highlights volatility fluctuations.

Interest rate volatility is very high in mid-sample.

Graphics Review



A bivariate scatterplot of the 1-year U.S. Treasury bond rate vs. the 10-year U.S. Treasury bond rate, from 1960.01 to 2005.03.

The scatterplot indicates that the two move closely together; in particular, they are positively correlated.

Probability and Statistics Review

Here we review a few aspects of probability and statistics that we will rely upon at various times.

Populations: Random Variables, Distributions and Moments

Discrete random variables, that is, random variables with discrete probability distributions, can assume only a countable number of values $y_i, ; i = 1, 2, \ldots,$ each with positive probability p_i such that $\sum p_i = 1$.

The probability distribution $f(y)$ assigns a probability p_i to each such value y_i. As an example flipping a coin twice, Y to be the number of heads observed in the two flips is a discrete random variable.

Continuous random variables can assume a continuous range of values, and the probability density function $f(y)$ is a non-negative continuous function such that the area under $f(y)$ between any points a and b is the probability that Y assumes a value between a and b.

Moments provide important summaries of various aspects of distributions. Roughly speaking, moments are simply expectations of powers of random variables, and expectations of different powers convey different sorts of information.

**Use a bit of review before plunging into regression, so begin by studying the provided supplementary material.

Empirical warm-up

** Let's dive into EViews and unlock its power for econometric analysis!

(a) Obtain time series of quarterly real GDP and quarterly real consumption for a country of your choice. Provide details.

(b) Display time-series plots and a scatterplot (put consumption on the vertical axis).

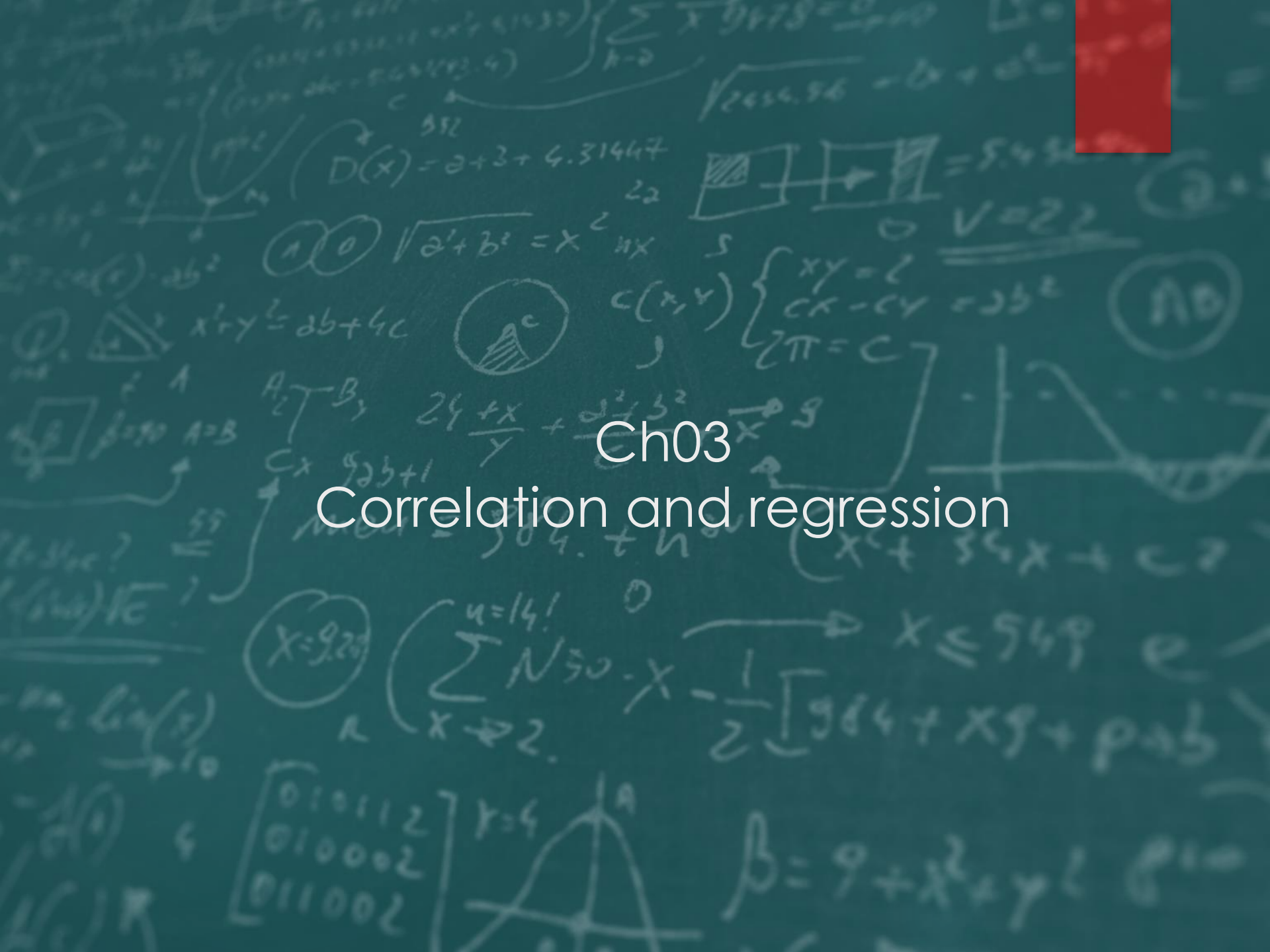
(c) Convert your series to growth rates in percent, and again display time series plots.

(d) From now on use the growth rate series only.

(e) For each series, provide summary statistics (e.g., mean, standard deviation, range, skewness, kurtosis, ...).

(f) For each series, perform t-tests of the null hypothesis that the population mean growth rate is 2 percent.

(g) For each series, calculate 90 and 95 percent confidence intervals for the population mean growth rate. For each series, which interval is wider, and why?



Ch03
Correlation and regression

Introduction

- Correlation and regression are techniques for investigating the statistical relationship between two, or more, variables.
- Although graphical methods (visually) helpful, this did not provide any precise measurement of the strength of the relationship.
- The χ^2 test did provide a test of the significance of the association between two category-based variables, but this test cannot be applied to variables measured on a ratio scale.
- Correlation and regression fill in these gaps: the strength of the relationship between two (or more) ratio scale variables can be measured and the significance tested.
- Correlation and regression are the techniques most often used by economists and forecasters.
- They can be used to answer such questions as
- ● Is there a link between the money supply and the price level?
- ● Do bigger firms produce at lower cost than smaller firms?
- ● Does instability in a country's export performance hinder its growth?

Correlation

- The correlation coefficient is a number which summarizes the relationship between two variables.
- The different types of possible relationship between any two variables, X and Y, may be summarized as follows:
- High values of X tend to be associated with low values of Y and vice versa. This is termed negative correlation.
- High (low) values of X tend to be associated with high (low) values of Y. This is positive correlation.
- No relationship between X and Y exists. High (low) values of X are associated about equally with high and low values of Y. This is zero, or the absence of, correlation.

Correlation

- The sample correlation coefficient, r, is a numerical statistic which has the following properties:
- It always lies between −1 and +1. This makes it relatively easy to judge the strength of an association.
- A positive value of r indicates positive correlation, a higher value indicating a stronger correlation between X and Y (i.e. the observations lie closer to a straight line).
- $r = 1$ indicates perfect positive correlation and means that all the observations lie precisely on a straight line with positive slope.
- A negative value of r indicates negative correlation. Similar to the above, a larger negative value indicates stronger negative correlation.
- $r = -1$ signifies perfect negative correlation.
- A value of r = 0 (or close to it) indicates a lack of correlation between X and Y.

Correlation

- The relationship is symmetric, i.e. the correlation between X and Y is the same as between Y and X. It does not matter which variable is labelled Y and which is labelled X.
- The formula for calculating the correlation coefficient is given in equation

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[(n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2)]}}$$

Example: In the provided Table, calculate for the relationship between birth rate (Y) and growth (X)

Country	Birth rate Y	GNP growth X	Y^2	X^2	XY
Brazil	30	5.1	900	26.01	153.0
Colombia	29	3.2	841	10.24	92.8
Costa Rica	30	3.0	900	9.00	90.0
India	35	1.4	1225	1.96	49.0
Mexico	36	3.8	1296	14.44	136.8
Peru	36	1.0	1296	1.00	36.0
Philippines	34	2.8	1156	7.84	95.2
Senegal	48	−0.3	2304	0.09	−14.4
South Korea	24	6.9	576	47.61	165.6
Sri Lanka	27	2.5	729	6.25	67.5
Taiwan	21	6.2	441	38.44	130.2
Thailand	30	4.6	900	21.16	138.0
Totals	380	40.2	12 564	184.04	1139.7

Correlation

- **Are the results significant?**
- These results come from a (small) sample, one of many that could have been collected. Once again we can ask the question, what can we infer about the population (of all developing countries) from the sample?
- Assuming the sample was drawn at random (which may not be justified) we can use the principles of hypothesis testing.
- As usual, there are two possibilities.
- (1) The truth is that there is no correlation (in the population) and that our sample exhibits such a large (absolute) value by chance.
- (2) There really is a correlation between the birth rate and the growth rate and the sample correctly reflects this.

Correlation

- Denoting the true but unknown population correlation coefficient by $\boldsymbol{\rho}$ the possibilities can be expressed in terms of a hypothesis test
- $H_0: \rho = 0$ and $H_1: \rho \neq 0$
- The test statistic in this case is not r itself but a transformation of it:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- This has a t distribution with $n - 2$ degrees of freedom.

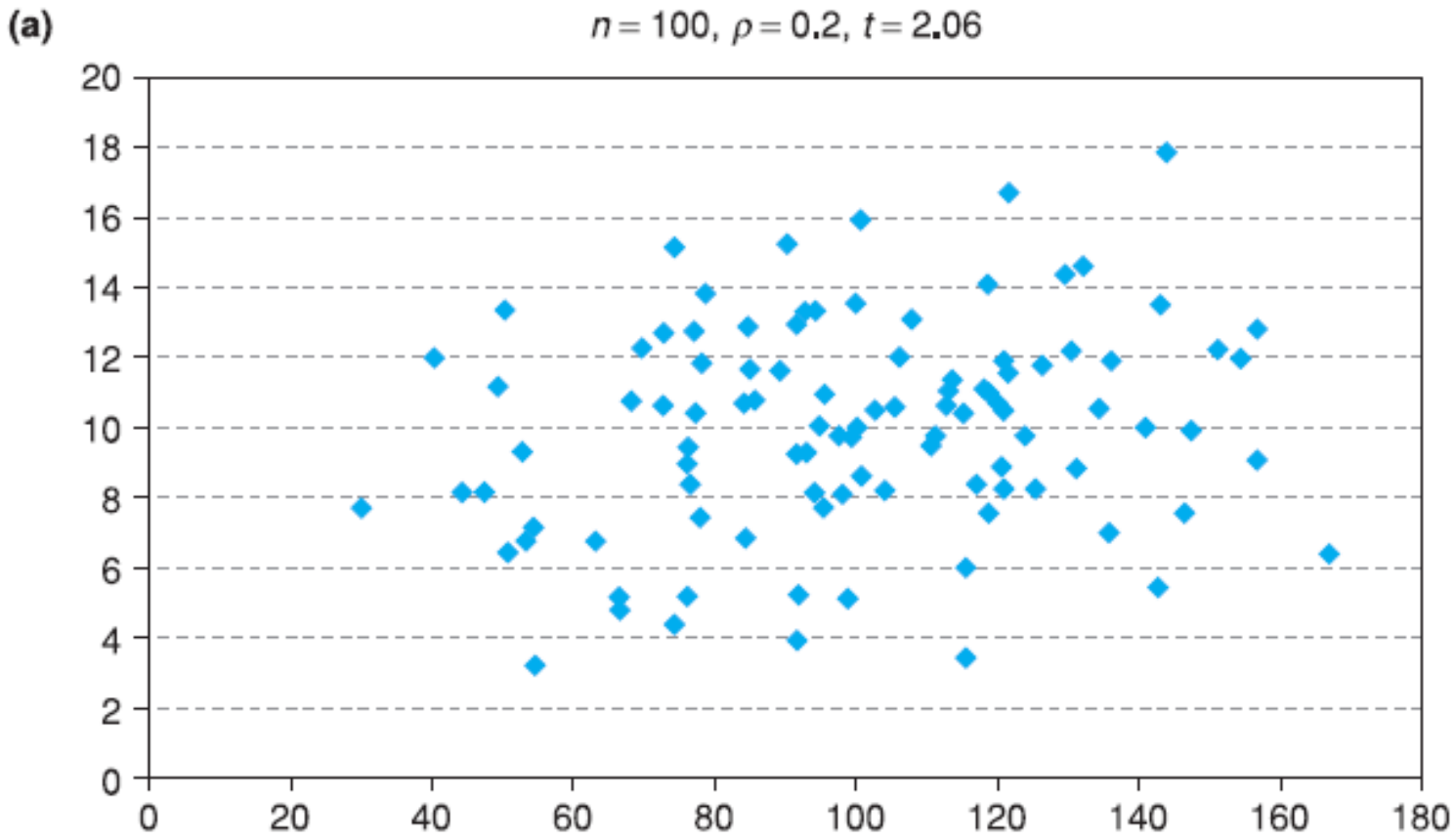
Correlation

- The five steps of the test procedure are therefore:
- (1) Write down the null and alternative hypotheses (shown above).
- (2) Choose the significance level of the test: 5% by convention.
- (3) Look up the critical value of the test for n − 2 = 10 degrees of freedom:
- (4) Calculate the test statistic using equation.
- (5) Compare the test statistic with the critical value.
- In the current case $t < -t^*$ so H_0 is rejected.
- There is a less than 5% chance of the sample evidence occurring if the null hypothesis were true, so the latter is rejected.

Correlation

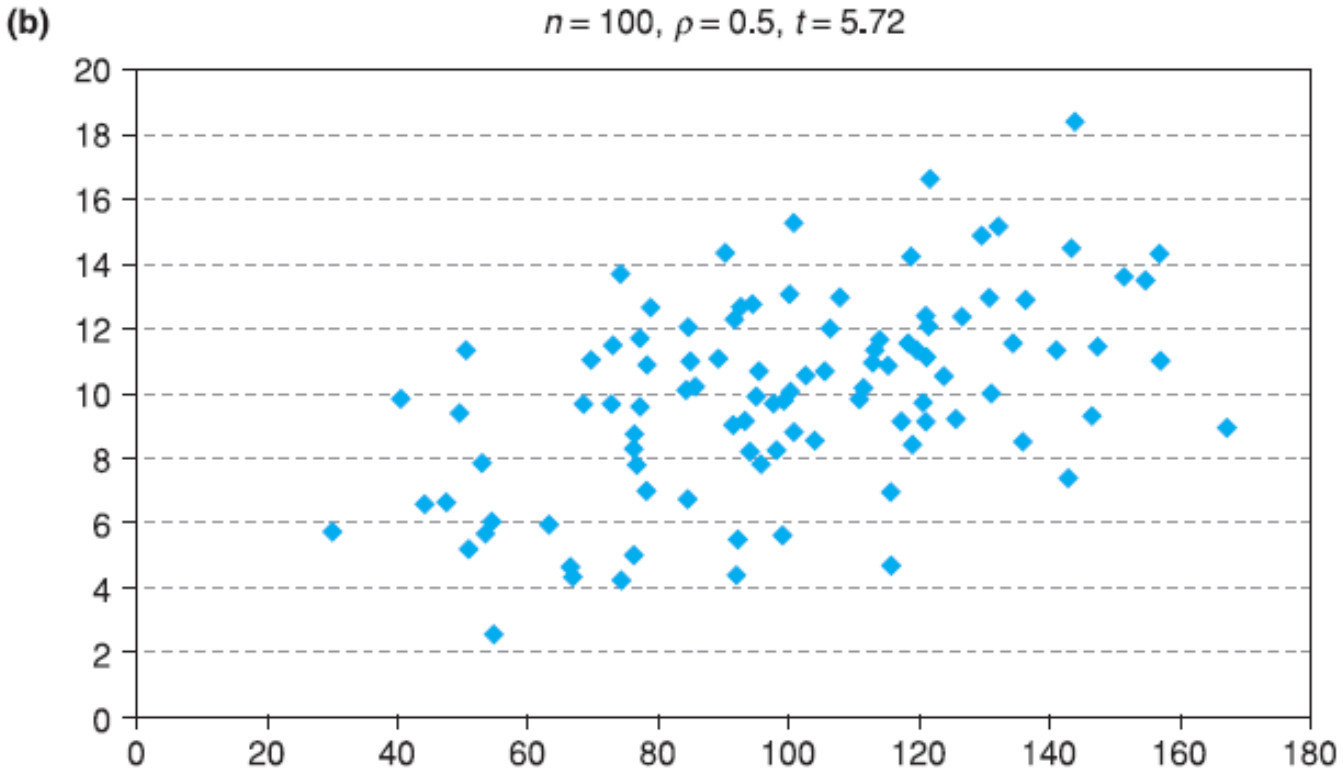
- **Are significant results important?**
- we might ask if a certain value of the correlation coefficient is economically important as well as being significant.
- We saw earlier that ‘significant’ results need not be important. The difficulty in this case is that we have little intuitive understanding of the correlation coefficient.
- Is $\rho = 0.5$ important, for example? Would it make much difference if it were only 0.4?
- Our understanding may be helped if we look at some graphs of variables with different correlation coefficients.
- Three are shown in Figure below.

Correlation



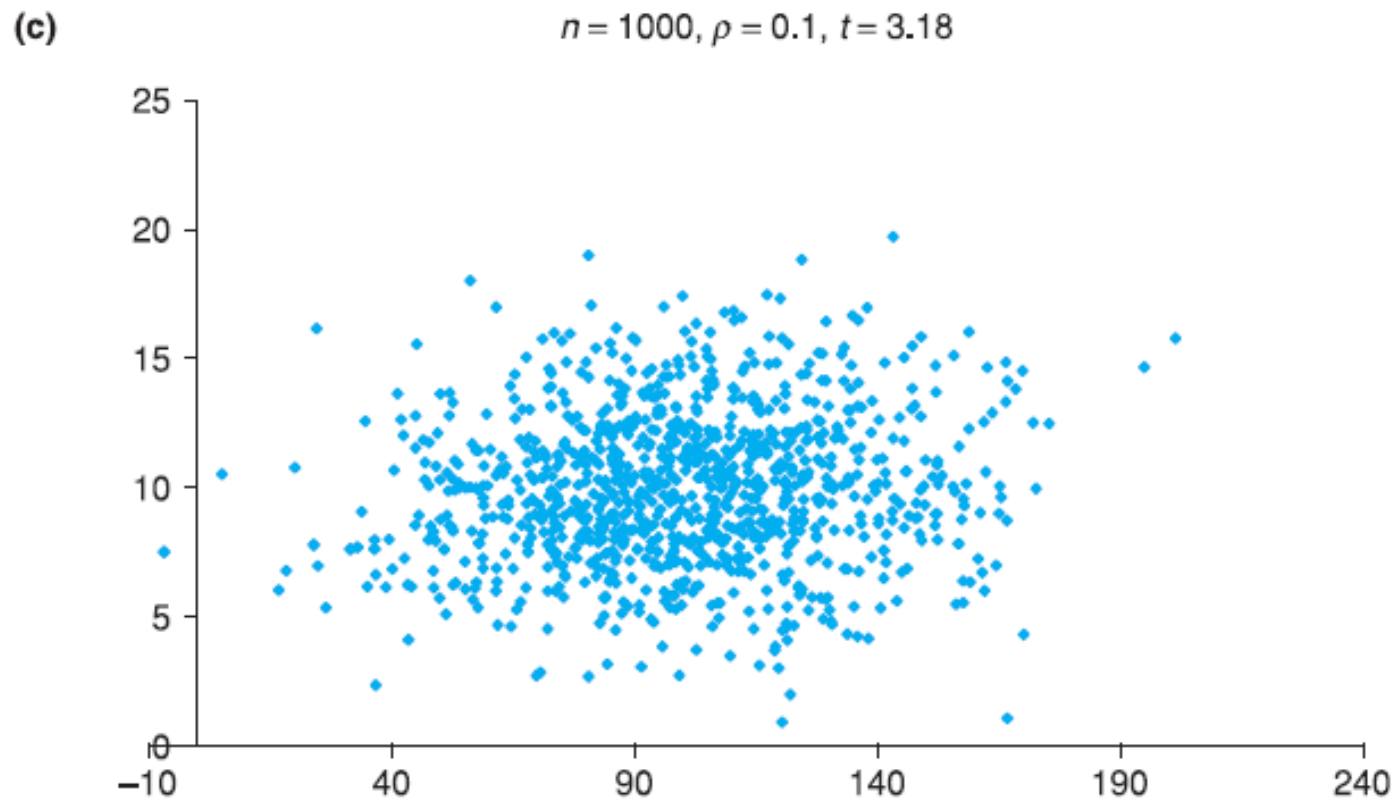
Panel (a) of the figure graphs two variables with a correlation coefficient of 0.2. Visually there seems little association between the variables, yet the correlation coefficient is (just) significant: t = 2.06 (n = 100 and the Prob-value is 0.046). This is a significant result which does not impress much.

Correlation



In panel (b) the correlation coefficient is 0.5 and the association seems a little stronger visually, though there is still a substantial scatter of the observations around a straight line. Yet the t statistic in this case is 5.72, highly significant (Prob-value 0.000).

Correlation



panel (c) shows an example where n = 1000. To the eye this looks much like a random scatter, with no discernable pattern. Yet the correlation coefficient is 0.1 and the t statistic is 3.18, again highly significant (Prob-value = 0.002).

Correlation

- The lessons from this seem fairly clear. What looks like a random scatter on a chart may in fact reveal a relationship between variables which is statistically significant, especially if there are a large number of observations.
- On the other hand, a high t-statistic and correlation coefficient can still mean there is a lot of variation in the data, revealed by the chart.
- Panel (b) suggests, for example, that we are unlikely to get a very reliable prediction of the value of y, even if we know the value of x.

Correlation and causality

- It is important to test the significance of any result because almost every pair of variables will have a non-zero correlation coefficient, even if they are totally unconnected (the chance of the sample correlation coefficient being exactly zero is very, very small).
- Therefore it is important to distinguish between correlation coefficients which are significant and those which are not, using the t test just outlined. But even when the result is significant one should beware of the danger of ‘spurious’ correlation.
- Many variables which clearly cannot be related turn out to be ‘significantly’ correlated with each other. One now famous example is between the price level and cumulative rainfall.
- Because they both rise year after year, it is easy to see why they are correlated, yet it is hard to think of a plausible reason why they should be causally related to each other.

Correlation and causality

- Apart from spurious correlation there are four possible reasons for a non-zero value of r.
- (1) X influences Y.
- (2) Y influences X.
- (3) X and Y jointly influence each other.
- (4) Another variable, Z, influences both X and Y.

Correlation and causality

- Correlation alone does not allow us to distinguish between these alternatives.
- For example, wages (X) and prices (Y) are highly correlated. Some people believe this is due to cost–push inflation, i.e. that wage rises lead to price rises. This is case (1) above.
- Others believe that wages rise to keep up with the cost of living (i.e. rising prices), which is (2).
- Perhaps a more convincing explanation is (3), a wage–price spiral where each feeds upon the other. Others would suggest that it is the growth of the money supply, Z, which allows both wages and prices to rise.
- To distinguish between these alternatives is important for the control of inflation, but correlation alone does not allow that distinction to be made.
- Correlation is best used therefore as a suggestive and descriptive piece of analysis, rather than a technique which gives definitive answers. It is often a preparatory piece of analysis, which gives some clues to what the data might yield, to be followed by more sophisticated techniques such as regression.

The coefficient of rank correlation

- On occasion it is inappropriate or impossible to calculate the correlation coefficient as described above and an alternative approach is required.
- Sometimes the original data are unavailable but the ranks are. For example, schools may be ranked in terms of their exam results, but the actual pass rates are not available.
- Similarly, they may be ranked in terms of spending per pupil, with actual spending levels unavailable.
- Although the original data are missing, one can still test for an association between spending and exam success by calculating the correlation between the ranks. If extra spending improves exam performance, schools ranked higher on spending should also be ranked higher on exam success, leading to a positive correlation.

The coefficient of rank correlation

- Second, even if the raw data are available, they may be highly skewed and hence the correlation coefficient may be influenced heavily by a few outliers.
- In this case, the hypothesis test for correlation may be misleading as it is based on the assumption of underlying Normal distributions for the data. In this case we could transform the values to ranks, and calculate the correlation of the ranks.
- In a similar manner to the median, described in Chapter 1, this can effectively deal with heavily skewed distributions.
- In these cases, it is Spearman's coefficient of rank correlation that is calculated.
- (The 'standard' correlation coefficient described above is more fully known as Pearson's product-moment correlation coefficient, to distinguish it.)
- The formula to be applied is the same as before, though there are a few tricks to be learned about constructing the ranks and also the hypothesis test is conducted in a different manner.

Regression analysis

Regression analysis

- Regression analysis is a more sophisticated way of examining the relationship between two (or more) variables than is correlation.
- The major differences between correlation and regression are the following:
- Regression can investigate the relationships between two or more variables.
- A direction of causality is asserted, from the explanatory variable (or variables) to the dependent variable.
- The influence of each explanatory variable upon the dependent variable is measured.
- The significance of each explanatory variable can be ascertained.

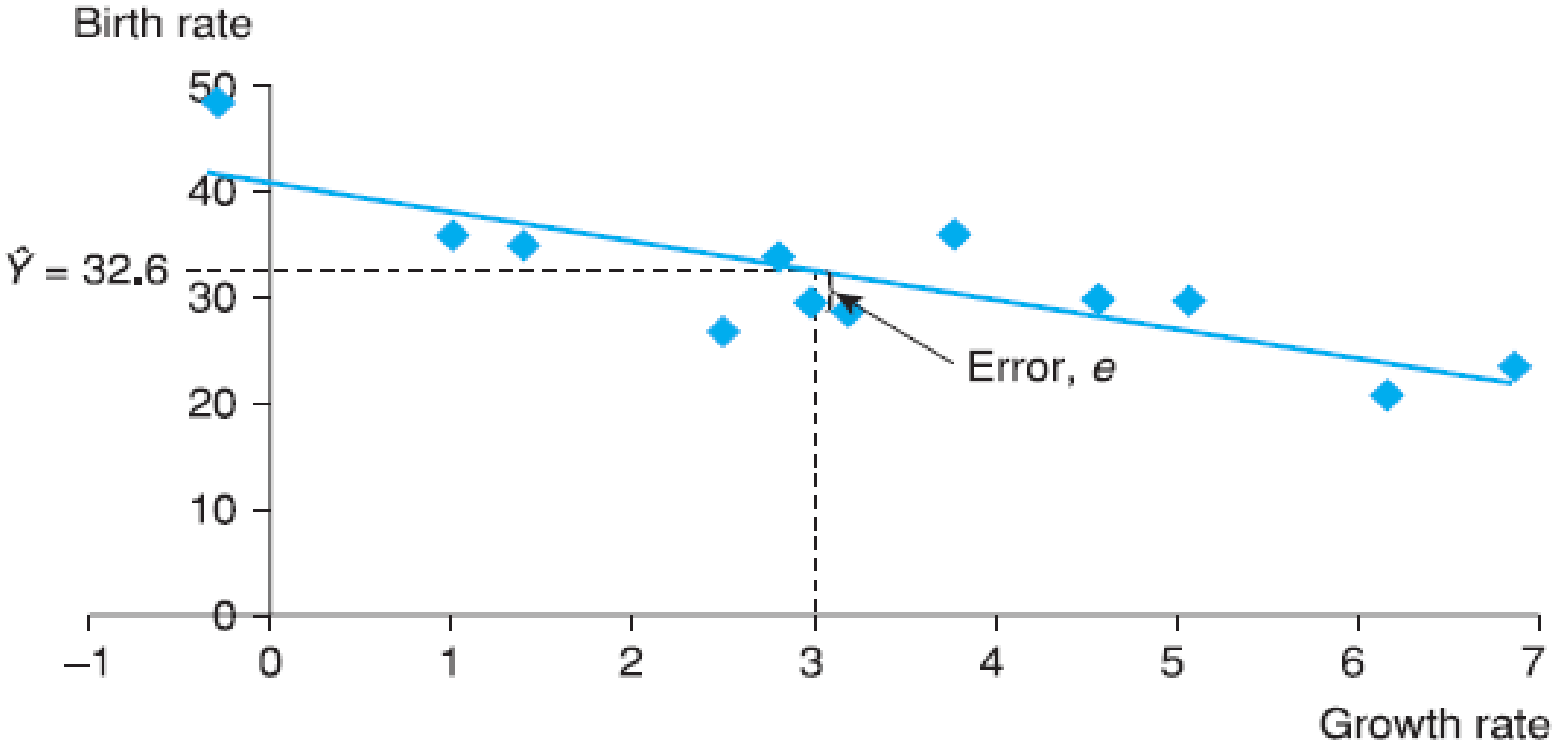
Regression analysis

- Follows the analysis in Michael Todaro's book, Economic Development in the Third World (3rd edn, pp. 197–200) where he tries to establish which of three variables (gross national product (GNP) per capita, the growth rate per capita or income inequality) is most important in determining a country's birth rate.
- The regression permits answers to such questions as:
- ● Does the growth rate influence a country's birth rate?
- ● If the growth rate increases, by how much might a country's birth rate be
- expected to fall?
- ● Are other variables important in determining the birth rate?

Regression analysis

- In this example we assert that the direction of causality is from the growth rate (X) to the birth rate (Y) and not vice versa. The growth rate is therefore the explanatory variable (also referred to as the independent or exogenous variable) and the birth rate is the dependent variable (also called the explained or endogenous variable).
- Regression analysis describes this causal relationship by fitting a straight line drawn through the data, which best summarises them. It is sometimes called ‘the line of best fit’ for this reason.
- This is illustrated in Figure below for the birth rate and growth rate data.

Regression analysis



- This regression line is downward sloping (its derivation will be explained shortly) for the same reason that the correlation coefficient is negative, i.e. high values of Y are generally associated with low values of X and vice versa.
- Note that (by convention) the explanatory variable is placed on the horizontal axis, the explained on the vertical.

Regression analysis

- Since the regression line summarizes knowledge of the relationship between X and Y, it can be used to predict the value of Y given any particular value of X.
- In Figure above the value of X = 3 (the observation for Costa Rica) is related via the regression line to a value of Y (denoted by Z) of 32.6.
- This predicted value is close (but not identical) to the actual birth rate of 30. The difference reflects the absence of perfect correlation between the two variables.

Regression analysis

- The difference between the actual value, Y, and the predicted value, Z, is called the error or residual. It is labelled *e* in Figure above. (Note: The italic *e* denoting the error term should not be confused with the roman letter*e*, used as the base for natural logarithms.
- Why should such errors occur?
- The relationship is never going to be an exact one for a variety of reasons. There are bound to be other factors besides growth which affect the birth rate (e.g. the education of women) and these effects are all subsumed into the error term.
- There might additionally be simple measurement error (of Y) and, of course, people do act in a somewhat random fashion rather than follow rigid rules of behaviour.
- All of these factors fall into the error term and this means that the observations lie around the regression line rather than on it. If there are many of these factors, none of which is predominant, and they are independent of each other, then these errors may be assumed to be Normally distributed about the regression line.

Regression analysis

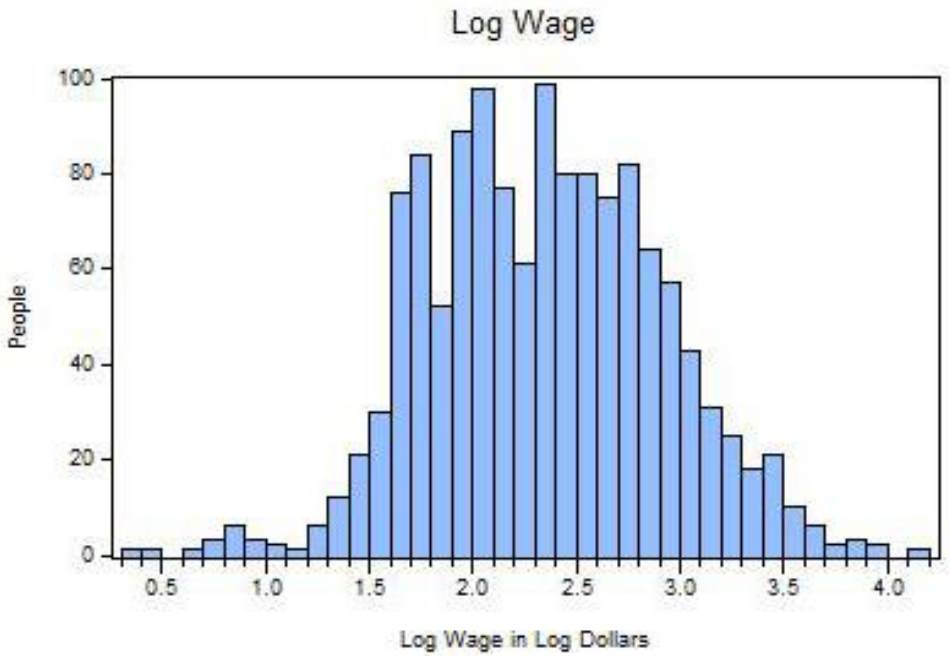
- Why not include these factors explicitly?
- On the face of it this would seem to be an improvement, making the model more realistic. However, the costs of doing this are that the model becomes more complex, calculation becomes more difficult (not so important now with computers) and it is generally more difficult for the reader (or researcher) to interpret what is going on.
- If the main interest is the relationship between the birth rate and growth, why complicate the model unduly?
- There is a virtue in simplicity, as long as the simplified model still gives an undistorted view of the relationship.

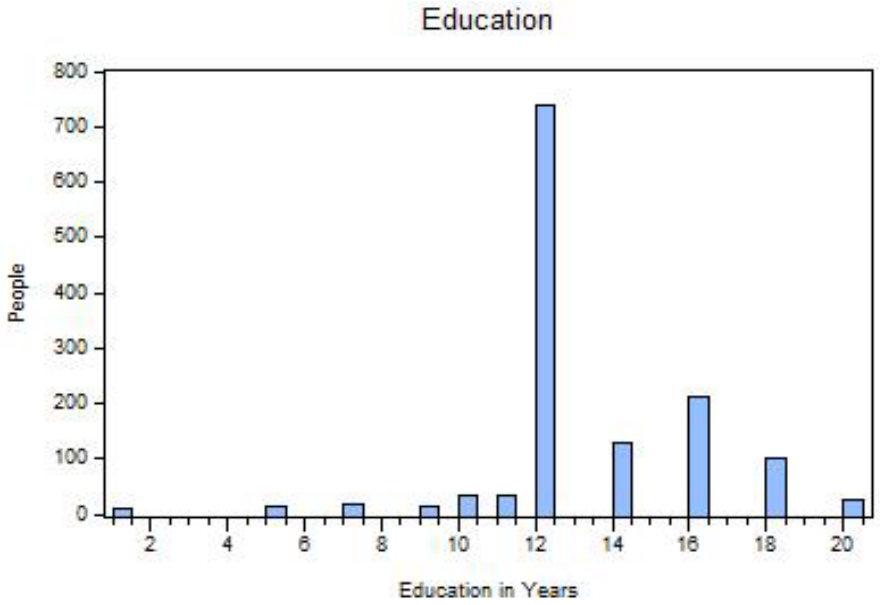
Regression Under Ideal Conditions

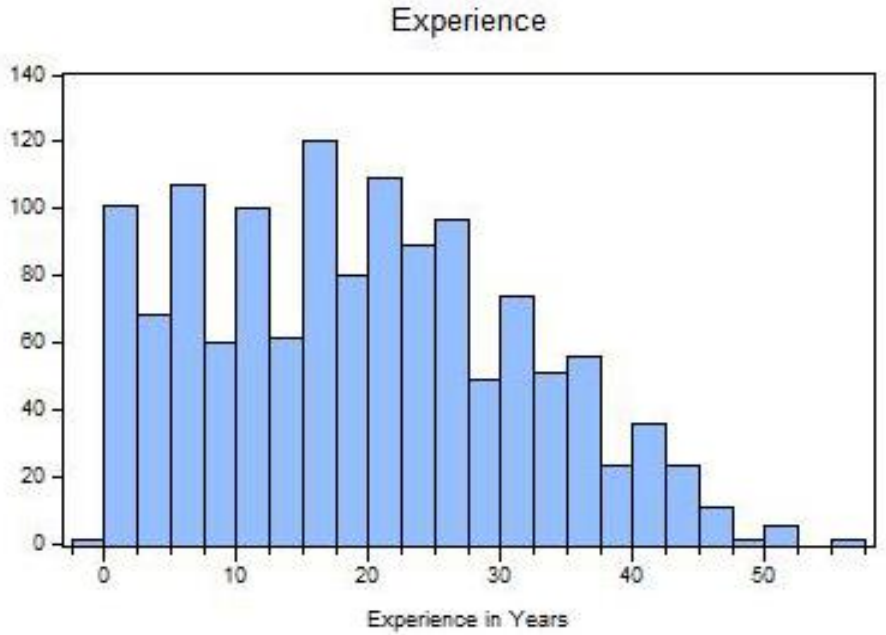
Regression Under Ideal Conditions

- Preliminary Graphics
- In this chapter we'll be working with cross-sectional data on log wages, education and experience.
- For convenience lets reproduce it is Figure, together with the distributions of the new data on education and experience.

Regression Under Ideal Conditions







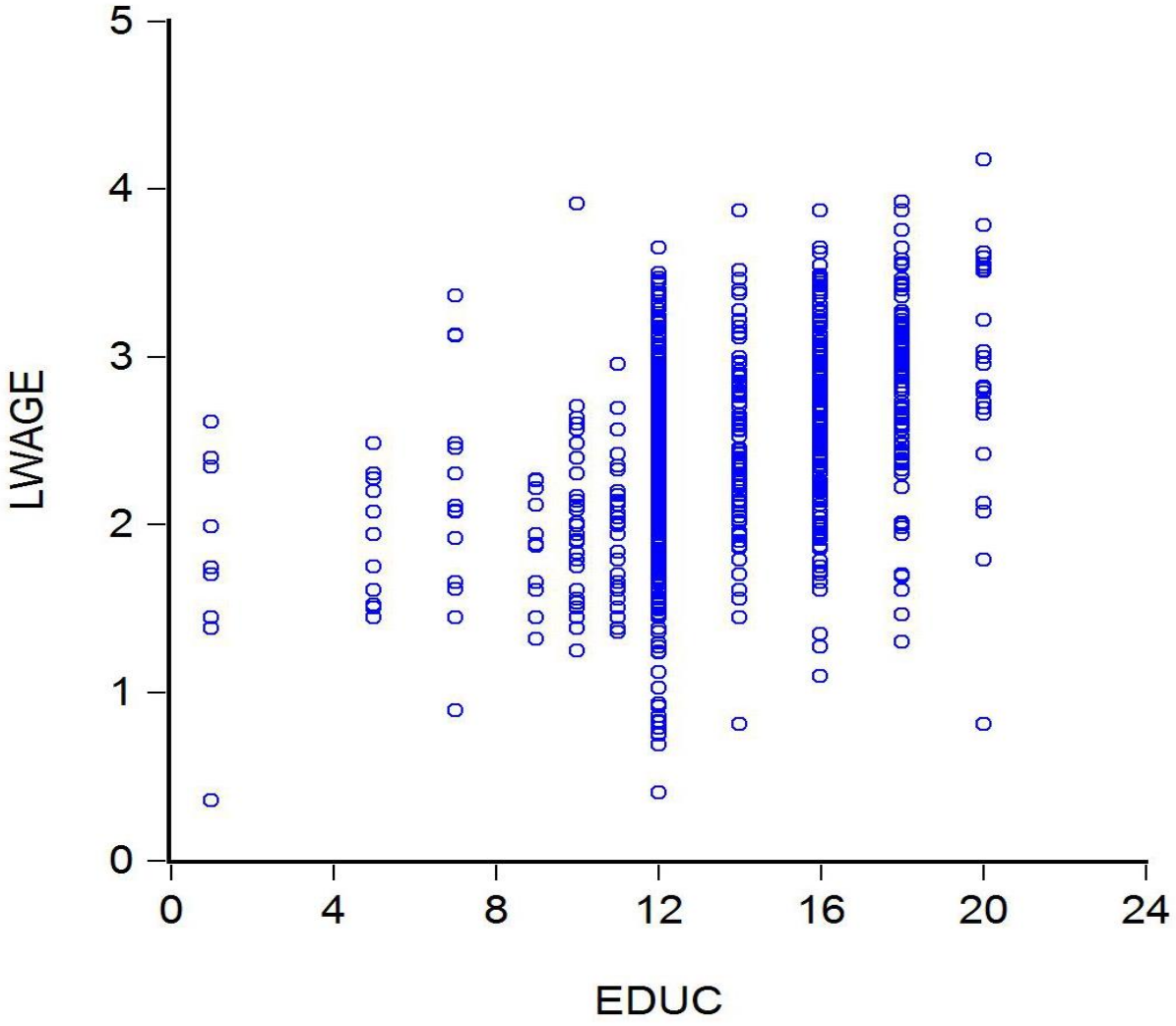
Regression as Curve Fitting

- Bivariate, or Simple, Linear Regression
- Suppose that we have data on two variables, y and x, as in Figure below, and suppose that we want to find the linear function of x that best fits y.
- This regression line is generally represented as:

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

- "best fits" means that the sum of squared (vertical) deviations of the data points from the fitted line is as small as possible.
- "run a regression," or "fit a regression line," that's what we do.
- The estimation strategy is called least squares, or sometimes "ordinary least squares" to distinguish it from fancier versions that we'll introduce later.
- The specific data that we show in Figure below are log wages (LWAGE, y) and education (EDUC, x) for a random sample of nearly 1500 people.

Regression as Curve Fitting



Regression as Curve Fitting

- When we run the regression, we use a computer to fit the line by solving the problem:

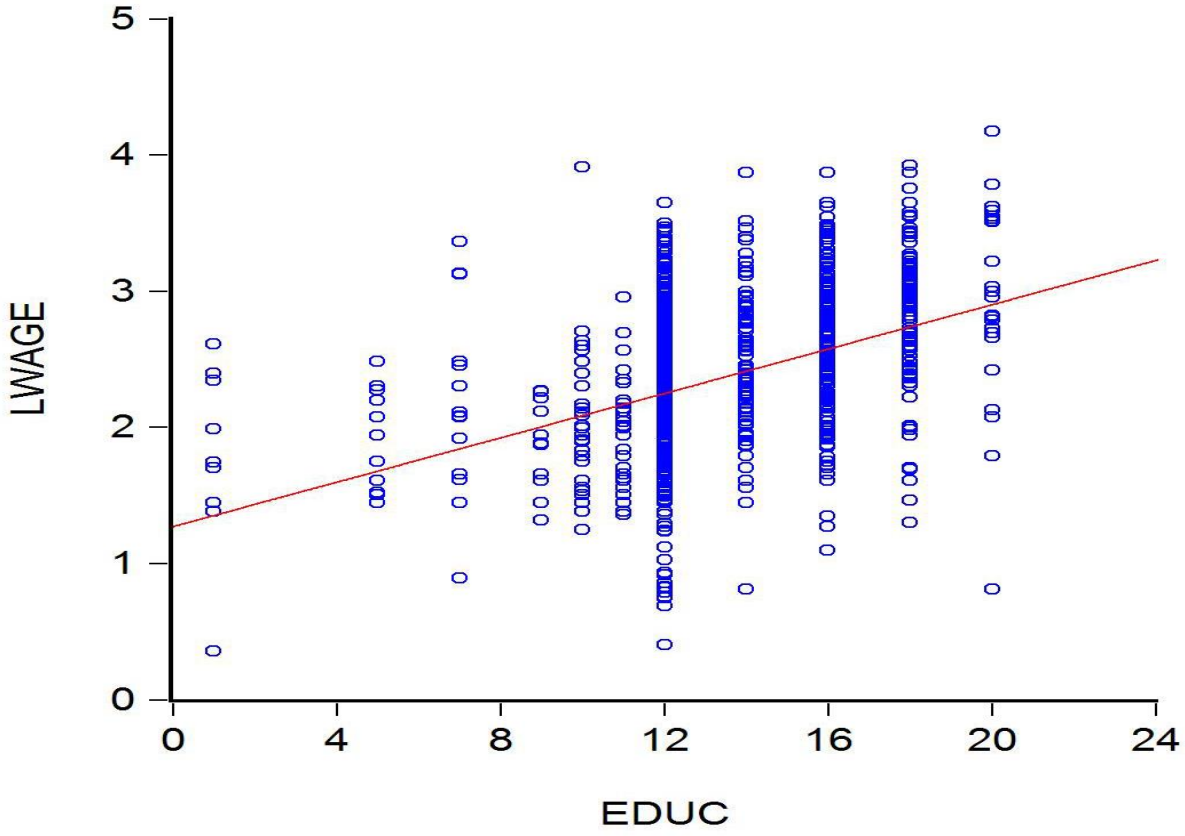
$$\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_i)^2$$

- where β is shorthand notation for the set of two parameters, β_1 and β_2. We denote the set of fitted parameters by $\hat{\beta}$, and its elements by $\hat{\beta}_1$ and $\hat{\beta}_2$.
- It turns out that the β_1 and β_2 values that solve the least squares problem have well-known mathematical formulas.
- We can use a computer to evaluate the formulas, simply, stably and instantaneously.
- The fitted values are

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i; \quad \forall\, i = 1,2,\dots,N$$

- The residuals are the difference between actual and fitted values, $e_i = y_i - \hat{y}_i$,

Regression as Curve Fitting



- We illustrate graphically the results of regressing LWAGE on EDUC.
- The best-fitting line slopes upward, reflecting the positive correlation between LWAGE and EDUC.
- Note that the data points don’t satisfy the fitted linear relationship exactly; rather, they satisfy it on average.
- To predict LWAGE for any given value of EDUC, we use the fitted line to find the value of LWAGE that corresponds to the given value of EDUC.

Regression as Curve Fitting

Multiple Linear Regression

- Everything generalizes to allow for more than one RHS variable. This is called multiple linear regression.
- Suppose, for example, that we have two RHS variables, x_2 and x_3. Before we fit a least-squares line to a two-dimensional data cloud; now we fit a least squares plane to a three-dimensional data cloud. We use the computer to find the values of β_1, β_2, and β_3 that solve the problem

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_{2i} - \beta_3 x_{3i})^2$$

- where β denotes the set of three model parameters. We denote the set of estimated parameters by $\hat{\beta}$, with elements $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$. The fitted values Are

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} \; ; \quad \forall \, i = 1,2, \dots, N$$

Regression as Curve Fitting

- WORTH NOTING:
- **First**, we now have two ways to analyze data and reveal its patterns.
- One is the graphical scatterplot with which we started, which provides a visual view of the data. The other is the fitted regression line which summarizes the data through the lens of a linear fit.
- Each approach has its merit, and the two are complements, not substitutes, but note that linear regression generalizes more easily to high dimensions.
- **Second,** least squares as introduced thus far has little to do with statistics or econometrics. Rather, it is simply a way of instructing a computer to fit a line to a scatterplot in a way that's rigorous, replicable and arguably reasonable.
- We now turn to a probabilistic interpretation.
- We work with the full multiple regression model (simple regression is of course a special case). Collect the RHS variables into the vector X, where $X_i' = (1, x_{2i}, \ldots, x_{ki})$.

Regression as a Probability Model

A Population Model and a Sample Estimator

- Thus far we have not postulated a probabilistic model that relates y_i and x_i ; instead, we simply ran a mechanical regression of y_i on x_i to find the best fit to y_i formed as a linear function of x_i.
- It's easy to construct a probabilistic framework that lets us make statistical assessments about the properties of the fitted line.
- We assume that y_i is linearly related to an exogenously-determined x_i, and we add an independent and identically distributed zero-mean (iid) Gaussian disturbance:

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \; ; \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- The intercept of the line is β_1, the slope parameters are the other β_2's, and the variance of the disturbance is σ^2. Collectively, we call the β's (and σ) the model's parameters.
- We assume that the linear model sketched is true in population; that is, it is the data-generating process (DGP).
- But in practice, of course, we don't know the values of the model's parameters, β_1, β_2, ..., β_K and σ^2. Our job is to estimate them using a sample of data from the population.
- We estimate the $\beta's$ precisely as before, using the computer to solve $\min_{\beta} \sum_{i=1}^{N} \varepsilon_i^2$.

Regression as a Probability Model

Notation, Assumptions and Results: The Ideal Conditions

The discussion thus far was intentionally a bit loose, focusing on motivation and intuition. Let us now be more precise about what we assume and what results we obtain.

A Bit of Matrix Notation

It will be useful to arrange all RHS variables into a matrix X. X has K columns, one for each regressor.

Inclusion of a constant in a regression amounts to including a special RHS variable that is always 1.

We put that in the leftmost column of the X matrix, which is just ones.

The other columns contain the data on the other RHS variables, over the cross section in the cross-sectional case, or over time in the time-series case. Notationally, X is a $N \times K$ matrix.

$$X = \begin{pmatrix} 1 & x_{21} & x_{31} & \cdots & x_{k1} \\ 1 & x_{22} & x_{32} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2N} & x_{3N} & \cdots & x_{kN} \end{pmatrix}$$

Regression as a Probability Model

- One reason that the X matrix is useful is because the regression model can be written very compactly using it.

- stack $y_i, i = 1, \ldots, N$ into the vector Y, where $y' = (y_1, y_2, \ldots, y_N)$, and stack $\beta_j, j = 1, \ldots, K$ into the vector β, where $\beta' = (\beta_1, \beta_2, \ldots, \beta_K)$, and stack $\varepsilon_i, i = 1, \ldots, N$, into the vector ε, where $\varepsilon' = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)$. Then we can write the complete model over all observations as

$$Y = X\beta + \varepsilon$$

- In addition,

$$\varepsilon_i \underset{\sim}{\overset{iid}{\sim}} N(0, \sigma^2)$$

- becomes

$$\varepsilon \underset{\sim}{\overset{iid}{\sim}} N(0, \sigma^2 I)$$

- Indeed above representation is crucially important, not simply because it is concise, but also because key results for estimation and inference may be stated very simply withing it, and because the various assumptions that we need to make to get various statistical results are most naturally and simply stated on X and ε in above equation. We now proceed to discuss such assumptions.

Regression as a Probability Model

- Assumptions: The Ideal Conditions (IC)

- 1. The data-generating process (DGP) is:

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \text{ where } \varepsilon_i \sim \overset{iid}{\sim} N(0, \sigma^2)$$

- The fitted model matches it exactly.

- This assumption has many important sub-conditions embedded. For example:

  - 1. The fitted model is correctly specified

  - 2. The disturbances are Gaussian

  - 3. The coefficients (β's) are fixed (whether over space or time, depending on whether we're working in a time-series or cross-section environment)

  - 4. The relationship is linear.

  - 5. The ε_i's have constant variance σ2

  - 6. The ε_i's are uncorrelated (whether over space or time, depending on whether we're working in a time-series or cross-section environment)

Regression as a Probability Model

- Assumptions: The Ideal Conditions (IC)
- 2. ε_i is independent of $(x_{1i}, \ldots, x_{Ki})$, for all i
- IC2 also has many important sub-conditions embedded
  - 1. $E(\varepsilon_i x_{ik}) = 0$, for all i, k (ε_i is uncorrelated with the x_{ik}'s)
  - 2. $E(\varepsilon_i | x_{i1}, \ldots, x_{iK}) = 0$, for all i (ε_i is conditional mean independent of the x_{ik}'s)
  - 3. $var(\varepsilon_i | x_{i1}, \ldots, x_{iK}) = \sigma^2$, for all i (ε_i is conditional variance independent of the x_{ik}'s)
- IC2 is subtle, and it may seem obscure at the moment, but it is very important in the context of causal estimation.
- The IC's are surely heroic in many contexts, and much of econometrics is devoted to detecting and dealing with various IC failures. But before we worry about IC failures, it's invaluable first to understand what happens when they hold.

Regression as a Probability Model

- Results Under the IC
- The least squares estimator is

$$\hat{\beta}_{LS} = (X'X)^{-1}X'y$$

- and under the IC it is (among other things) consistent, asymptotically efficient, and asymptotically normally distributed.
- We write:

$$\hat{\beta}_{LS} \overset{a}{\sim} N(\beta, V)$$

- We consistently estimate the covariance matrix V using $\hat{V} = s^2(X'X)^{-1}$, where $s^2 = \sum_{i=1}^{N}(\frac{e_i^2}{N-K})$.

Regression: A Wage Equation

- A Wage Equation
- let's look in detail at the computer output for a regression of log wages (LWAGE) on an intercept, education (EDUC) and experience (EXPER). The output is in Eviews format.
- Before proceeding, note well that the IC may not be satisfied for this dataset, yet we will proceed assuming that they are satisfied. We will confront violations of the various assumptions –indeed that's what econometrics is largely about – and we'll return repeatedly to this dataset and others. But we must begin at the beginning.

Regression: A Wage Equation

Equation: UNTITLED Workfile: GRAPHS::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LWAGE
Method: Least Squares
Date:
Sample (adjusted): 1 1323
Included observations: 1323 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.867382	0.075331	11.51431	0.0000
EDUC	0.093229	0.005045	18.48002	0.0000
EXPER	0.013104	0.001164	11.26208	0.0000

R-squared	0.232224	Mean dependent var	2.341995
Adjusted R-squared	0.231061	S.D. dependent var	0.561435
S.E. of regression	0.492318	Akaike info criterion	1.422881
Sum squared resid	319.9376	Schwarz criterion	1.434644
Log likelihood	-938.2358	Hannan-Quinn criter.	1.427291
F-statistic	199.6260	Durbin-Watson stat	1.926045
Prob(F-statistic)	0.000000		

Regression: A Wage Equation

- The four statistics associated with each RHS variable are the estimated coefficient ("Coefficient"), its standard error ("Std. Error"), a t statistic, and a corresponding probability value ("Prob.").
- The standard errors of the estimated coefficients indicate their likely sampling variability, and hence their reliability.
- The estimated coefficient plus or minus one standard error is approximately a 68% confidence interval for the true but unknown population parameter, and the estimated coefficient plus or minus two standard errors is approximately a 95% confidence interval, assuming that the estimated coefficient is approximately normally distributed, which will be true if the regression disturbance is normally distributed or if the sample size is large.
- Thus large coefficient standard errors translate into wide confidence intervals.

Regression: A Wage Equation

- Each t statistic provides a test of the hypothesis of variable irrelevance: that the true but unknown population parameter is zero, so that the corresponding variable contributes nothing to the regression and can therefore be dropped.
- One way to test variable irrelevance, with, say, a 5% probability of incorrect rejection, is to check whether zero is outside the 95% confidence interval for the parameter. If so, we reject irrelevance.
- The t statistic is just the ratio of the estimated coefficient to its standard error, so if zero is outside the 95% confidence interval, then the t statistic must be bigger than two in absolute value. Thus we can quickly test irrelevance at the 5% level by checking whether the t statistic is greater than two in absolute value.

Regression: A Wage Equation

- Finally, associated with each t statistic is a probability value, which is the probability of getting a value of the t statistic at least as large in absolute value as the one actually obtained, assuming that the irrelevance hypothesis true. Hence if a t statistic were two, the corresponding probability value would be approximately .05.
- The smaller the probability value, the stronger the evidence against irrelevance. There's no magic cutoff, but typically probability values less than 0.1 are viewed as strong evidence against irrelevance, and probability values below 0.05 are viewed as very strong evidence against irrelevance.
- Probability values are useful because they eliminate the need for consulting tables of the t distribution.
- Effectively the computer does it for us and tells us the significance level at which the irrelevance hypothesis is just rejected.

Regression: A Wage Equation

- **Interpretation:**
- The estimated **intercept** is approximately .867, so that conditional on zero education and experience, our best forecast of the log wage would be 86.7 cents.
- Moreover, the intercept is very precisely estimated, as evidenced by the small standard error of .08 relative to the estimated coefficient. An approximate 95% confidence interval for the true but unknown population intercept is .867 ± 2(.08), or [.71, 1.03].
- Zero is far outside that interval, so the corresponding t statistic is huge, with a probability value that's zero to four decimal places.

Regression: A Wage Equation

- **Interpretation:**
- The estimated **coefficient on EDUC** is .093, and the standard error is again small in relation to the size of the estimated coefficient, so the t statistic is large and its probability value small.
- The coefficient is positive, so that LWAGE tends to rise when EDUC rises. In fact, the interpretation of the estimated coefficient of .09 is that, holding everything else constant, a one year increase in EDUC will produce a .093 increase in LWAGE.
- The estimated **coefficient on EXPER** is .013. Its standard error is also small, and hence its t statistic is large, with a very small probability value. Hence we reject the hypothesis that EXPER contributes nothing to the forecasting regression.
- A one-year increase in EXPER tends to produce a .013 increase in LWAGE.

Regression: A Wage Equation

- **Diagnostic Statistics:**
- A variety of statistics help us to evaluate the adequacy of the regression. Here we introduce them very briefly:
- **1. Mean dependent var 2.342**
- The sample mean of the dependent variable is $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$
- It measures the central tendency, or location, of y.
- **2. S.D. dependent var .561**
- The sample standard deviation of the dependent variable is $SD = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N-1}}$
- It measures the dispersion, or scale, of y.

Regression: A Wage Equation

- **3. Sum squared resid 319.938**
- Minimizing the sum of squared residuals is the objective of least squares estimation. It's natural, then, to record the minimized value of the sum of squared residuals.
- In isolation it's not of much value, but it serves as an input to other diagnostics that we'll discuss shortly.
- Moreover, it's useful for comparing models and testing hypotheses.

$$SSR = \sum_{i=1}^{N} (e_i)^2$$

Regression: A Wage Equation

- **4. Log likelihood -938.236**
- The likelihood function is the joint density function of the data, viewed as a function of the model parameters. Hence a natural estimation strategy, called maximum likelihood estimation, is to find (and use as estimates) the parameter values that maximize the likelihood function.
- After all, by construction, those parameter values maximize the likelihood of obtaining the data that were actually obtained. In the leading case of normally-distributed regression disturbances, maximizing the likelihood function (or equivalently, the log likelihood function, because the log is a monotonic transformation) turns out to be equivalent to minimizing the sum of squared residuals, hence the maximum-likelihood parameter estimates are identical to the least-squares parameter estimates.
- The number reported is the maximized value of the log of the likelihood function.

Regression: A Wage Equation

- Like the sum of squared residuals, it's not of direct use, but it's useful for comparing models and testing hypotheses.
- A natural estimation strategy with wonderful asymptotic properties, called maximum likelihood estimation, is to find (and use as estimates) the parameter values that maximize the likelihood function. After all, by construction, those parameter values maximize the likelihood of obtaining the data that were actually obtained.
- In the leading case of normally-distributed regression disturbances, maximizing the likelihood function turns out to be equivalent to minimizing the sum of squared residuals, hence the maximum-likelihood parameter estimates are identical to the least-squares parameter estimates.

Regression: A Wage Equation

- To see why maximizing the Gaussian log likelihood gives the same parameter estimate as minimizing the sum of squared residuals, let us derive the likelihood for the Gaussian linear regression model with non-stochastic regressors,

$$y_i = x_i' \beta + \varepsilon; \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- The model implies that

$$\mathrm{y_i} \overset{iid}{\sim} N(x_i' \beta, \sigma^2)$$

- So that the density function of $\mathrm{y_i}$ is

$$f(y_i) = (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i' \beta)^2}$$

- Hence $f(y_1, \ldots, y_N) = f(y_1) f(y_2) \cdots f(y_N)$ (by independence of the y_i's). In particular, the likelihood of the sample, denoted by L, is

$$L = \prod_{i=1}^{N} \left((2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i' \beta)^2} \right)$$

Regression: A Wage Equation

$$\ln(L) = \ln\left((2\pi\sigma^2)^{-\frac{N}{2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i'\,\beta)^2$$

$$\ln(L) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\,ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i'\,\beta)^2$$

- Note in particular that the β vector that maximizes the likelihood (or log likelihood – the optimizers must be identical because the log is a positive monotonic transformation) is the β vector that minimizes the sum of squared residuals.
- The log likelihood is also useful for hypothesis testing via likelihood-ratio tests.

Regression: A Wage Equation

- Under very general conditions we have asymptotically that:

$$-2\,[\ln(L_0) - \ln(L_1)] \sim \chi^2_d$$

- where ln L0 is the maximized log likelihood under the restrictions implied by the null hypothesis, ln L1 is the unrestricted log likelihood, and d is the number of restrictions imposed under the null hypothesis.
- t and F tests are likelihood ratio tests under a normality assumption.
- That's why they can be written in terms of minimized SSR's rather than maximized ln L's.

Regression: A Wage Equation

- **5. F statistic 199.626**
- We use the F statistic to test the hypothesis that the coefficients of all variables in the regression except the intercept are jointly zero.
- We test whether, taken jointly as a set, the variables included in the forecasting model have any explanatory value.
- This contrasts with the t statistics, which we use to examine the explanatory value of the variables one at a time.
- If no variable has explanatory value, the F statistic follows an F distribution with k − 1 and T − k degrees of freedom.

Regression: A Wage Equation

- The formula is

$$F = \frac{\left[\frac{SSR_{res} - SSR}{k-1}\right]}{\left[\frac{SSR}{N-k}\right]}$$

- where SSR_{res} is the sum of squared residuals from a restricted regression that contains only an intercept.
- Thus the test proceeds by examining how much the SSR increases when all the variables except the constant are dropped.
- If it increases by a great deal, there's evidence that at least one of the variables has explanatory content.

Regression: A Wage Equation

- **6. Prob(F statistic) 0.000000**
- The probability value for the F statistic gives the significance level at which we can just reject the hypothesis that the set of RHS variables has no predictive value. Here, the value is indistinguishable from zero, so we reject the hypothesis overwhelmingly.
- **7. S.E. of regression .492**
- If we knew the elements of β and predicted y_i using $x_i'\beta$, then our prediction errors would be the $\varepsilon'_i s$, with variance σ^2.
- An estimate of σ^2 tells us whether our prediction errors are likely to be large or small. The observed residuals, the $\varepsilon'_i s$, are effectively estimates of the unobserved population disturbances, the $\varepsilon'_i s$.

Regression: A Wage Equation

- Thus the sample variance of the e's, which we denote s^2 (read "s-squared"), is a natural estimator of σ^2:

$$s^2 = \frac{\sum_{i=1}^{N} e_i^2}{N - K}$$

- s^2 is an estimate of the dispersion of the regression disturbance and hence is used to assess goodness of fit of the model, as well as the magnitude of prediction errors that we're likely to make.

- The larger is s^2 , the worse the model's fit, and the larger the prediction errors we're likely to make. s^2 involves a degrees-of-freedom correction (division by N − K rather than by N −1, reflecting the fact that K regression coefficients have been estimated), which is an attempt to get a good estimate of the out-of-sample prediction error variance on the basis of the in-sample residuals.

Regression: A Wage Equation

- The standard error of the regression (SER) conveys the same information; it's an estimator of σ rather than σ^2, so we simply use s rather than s^2 .
- The formula is

$$SER = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N - K}}$$

- The standard error of the regression is easier to interpret than s^2, because its units are the same as those of the e's, whereas the units of s^2 are not.
- If the e's are in dollars, then the squared e's are in dollars squared, so s^2 is in dollars squared. By taking the square root at the end of it all, SER converts the units back to dollars.

Regression: A Wage Equation

- **8. R-squared .232**
- If an intercept is included in the regression, as is almost always the case, R-squared must be between zero and one. In that case, R-squared, usually written R^2, is the percent of the variance of y explained by the variables included in the regression. R^2 measures the in-sample success of the regression equation in predicting y; hence it is widely used as a quick check of goodness of fit, or predictability, of y based on the variables included in the regression.
- Here the R^2 is about 23% – well above zero but not great. The formula is

$$R^2 = 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

Regression: A Wage Equation

- We can write R^2 in a more roundabout way as

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N} e_i^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

- which makes clear that the numerator in the large fraction is very close to s^2, and the denominator is very close to the sample variance of y.

Regression: A Wage Equation

- **9. Adjusted R-squared .231**
- The interpretation is the same as that of R^2, but the formula is a bit different. Adjusted R^2 incorporates adjustments for degrees of freedom used in fitting the model, in an attempt to offset the inflated appearance of good fit if many RHS variables are tried and the "best model" selected.
- adjusted R^2 is a more trustworthy goodness-of-fit measure than R^2. As long as there is more than one RHS variable in the model fitted, adjusted R^2 is smaller than R^2; here, however, the two are extremely close (23.1% vs. 23.2%). Adjusted R^2 is often denoted $\bar{R}^2$; the formula is

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-K}\sum_{i=1}^{N} e_i^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where K is the number of RHS variables, including the constant term. Here the numerator in the large fraction is precisely s^2, and the denominator is precisely the sample variance of y.

Regression: A Wage Equation

- **10. Akaike info criterion 1.423**
- The Akaike information criterion, or AIC, is effectively an estimate of the out-of-sample forecast error variance, as is s^2, but it penalizes degrees of freedom more harshly.
- It is used to select among competing models. The formula is:

$$AIC = \left[e^{\left(\frac{2K}{N}\right)}\right] \left[\frac{\sum_{i=1}^{N} e_i^2}{N}\right]$$

- "smaller is better". That is, we select the model with smallest AIC.

Regression: A Wage Equation

- **11. Schwarz criterion 1.435**
- The Schwarz information criterion, or SIC, is an alternative to the AIC with the same interpretation, but a still harsher degrees-of-freedom penalty.
- Schwarz criterion is also known as the Bayesian Information Criteria, or BIC.
- The formula is:

$$AIC = \left[N^{\left(\frac{K}{N}\right)}\right] \left[\frac{\sum_{i=1}^{N} e_i^2}{N}\right]$$

- and “smaller is better”. That is, we select the model with smallest SIC.

Regression: A Wage Equation

A Bit More on AIC and SIC

- The AIC and SIC are tremendously important for guiding model selection in a ways that avoid data mining and in-sample overfitting.
- You will want to start using AIC and SIC immediately, so we provide a bit more information here. Model selection by maximizing R^2, or equivalently minimizing residual SSR, is ill-advised, because they don't penalize for degrees of freedom and therefore tend to prefer models that are "too big."
- Model selection by maximizing $\bar{R}^2$, or equivalently minimizing residual s^2, is still ill-advised, even though $\bar{R}^2$ and s^2 penalize somewhat for degrees of freedom, because they don't penalize harshly enough and therefore still tend to prefer models that are too big. In contrast, AIC and SIC get things just right. SIC has a wonderful asymptotic optimality property when the set of candidate models is viewed as fixed: Basically SIC "gets it right" asymptotically, selecting either the DGP (if the DGP is among the models considered) or the best predictive approximation to the DGP (if the DGP is not among the models considered).
- AIC has a different and also-wonderful asymptotic optimality property, known as "efficiency," when the set of candidate models is viewed as expanding as the sample size grows. In practice, the models selected by AIC and SIC rarely disagree.

Regression: A Wage Equation

12. Hannan-Quinn criter. 1.427

Hannan-Quinn is yet another information criterion for use in model selection.

13. Durbin-Watson stat. 1.926

The Durbin-Watsion (DW) statistic is used in time-series contexts, and we will study it later.

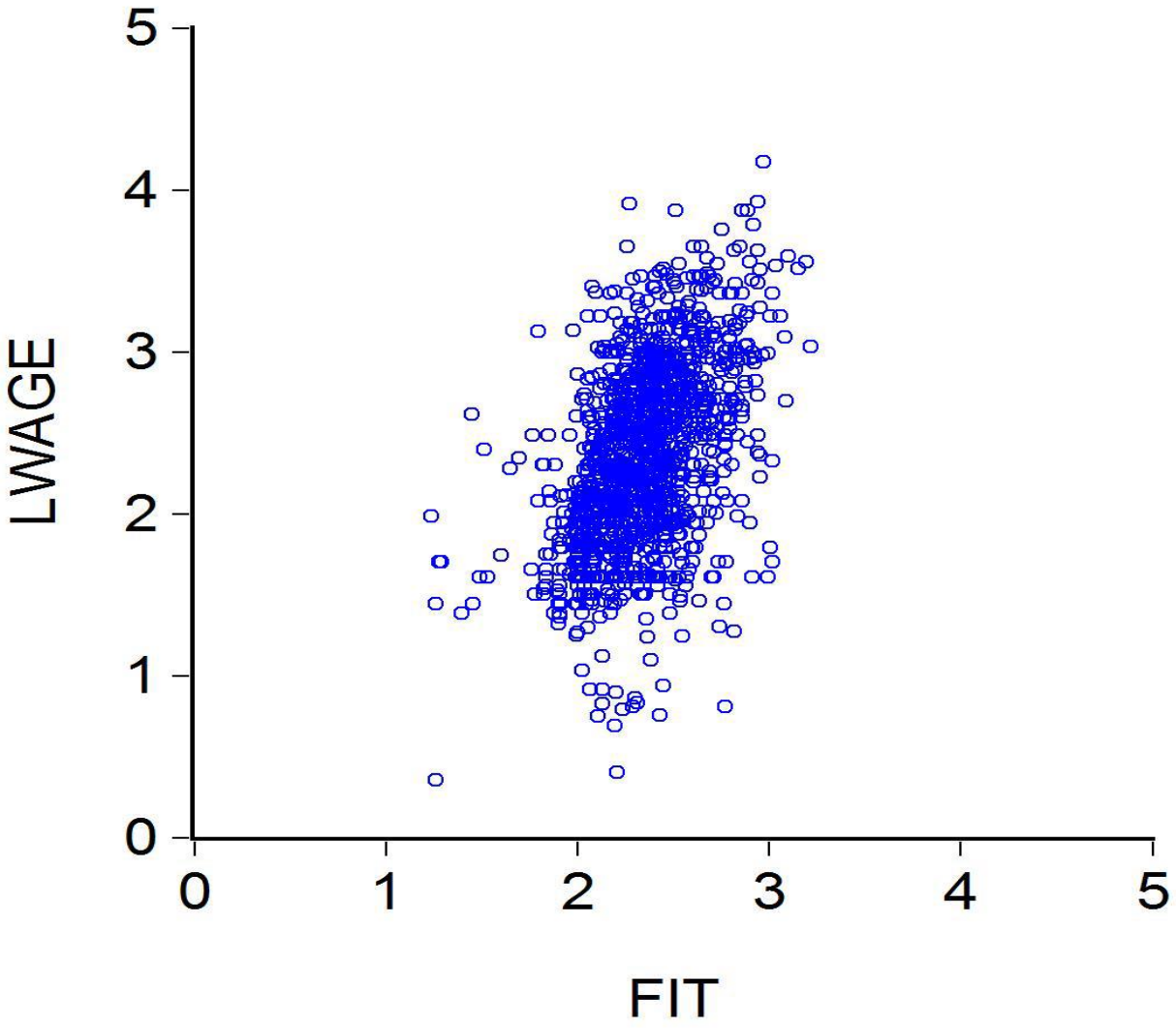
14. The Residual Scatter

The residual scatter is often useful in both cross-section and time-series situations. It is a plot of y vs $\hat{y}$.

A perfect fit ($R^2 = 1$) corresponds to all points on the 45 degree line, and no fit ($R^2 = 0$) corresponds to all points on a vertical line corresponding to $y = \bar{y}$.

In Figure below we show the residual scatter for the wage regression. It is not a vertical line, but certainly also not the 45 degree line, corresponding to the positive but relatively low R^2 of .23.

Regression: A Wage Equation

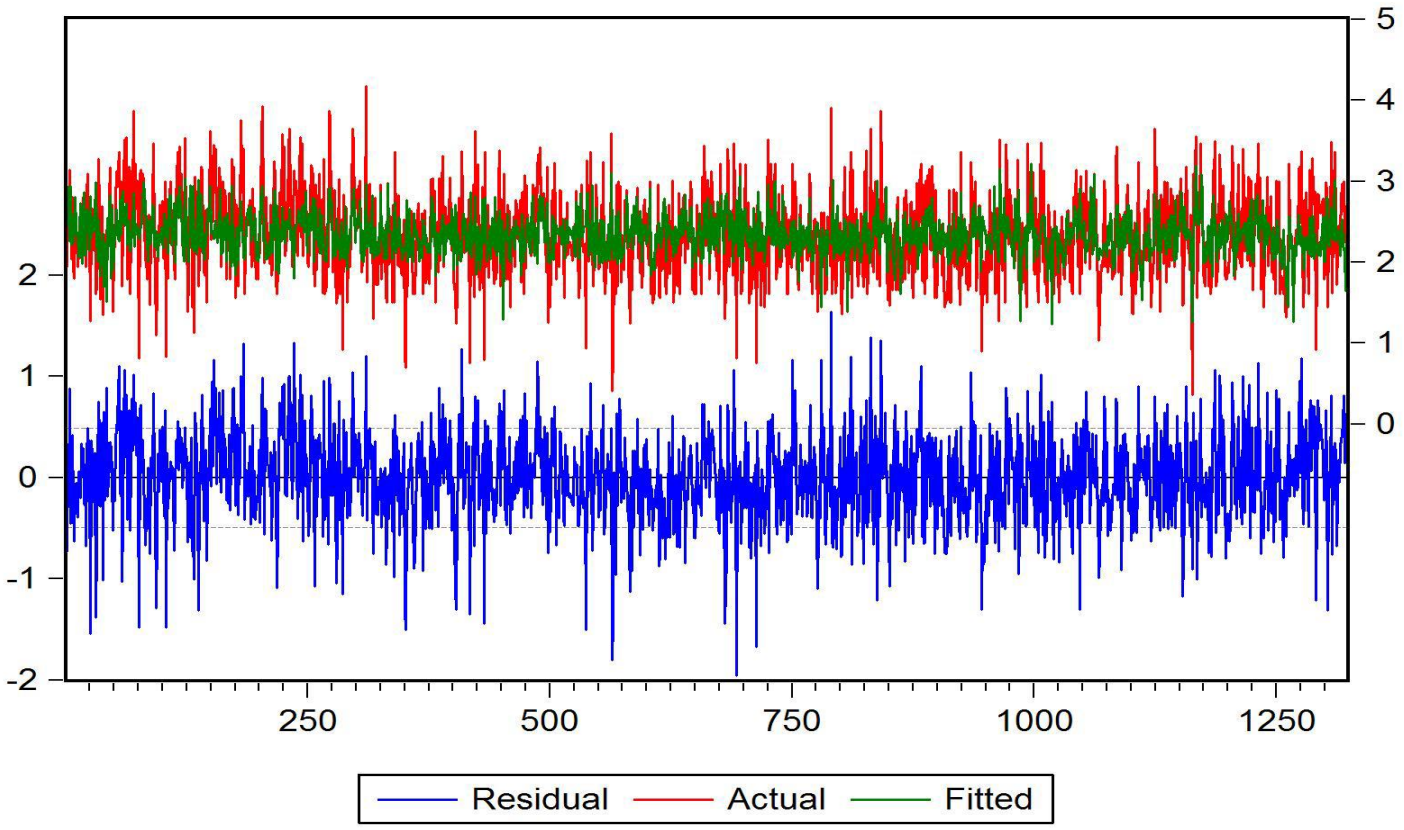


Regression: A Wage Equation

15. The Residual Plot

- In time-series settings, it's always a good idea to assess visually the adequacy of the model via time series plots of the actual data (y_i's), the fitted values ($\hat{y}_i$'s), and the residuals (e_i's). Often we'll refer to such plots, shown together in a single graph, as a residual plot.
- Note that even with many RHS variables in the regression model, both the actual and fitted values of y, and hence the residuals, are simple univariate series that can be plotted easily.
- The reason we examine the residual plot is that patterns would indicate violation of our iid assumption. In time series situations, we are particularly interested in inspecting the residual plot for evidence of serial correlation in the e_i's, which would indicate failure of the assumption of iid regression disturbances. More generally, residual plots can also help assess the overall performance of a model by flagging anomalous residuals, due for example to outliers, neglected variables, or structural breaks.
- Our wage regression is cross-sectional, so there is no natural ordering of the observations, and the residual plot is of limited value. But we can still use it, for example, to check for outliers.

Regression: A Wage Equation



In the figure, we show the residual plot for the regression of LWAGE on EDUC and EXPER. The actual and fitted values appear at the top of the graph; their scale is on the right. The fitted values track the actual values fairly well. The residuals appear at the bottom of the graph; their scale is on the left. It's important to note that the scales differ; the e_i's are in fact substantially smaller and less variable than either the y_i's or the $\hat{y}_i$'s. We draw the zero line through the residuals for visual comparison. No outliers are apparent.

Least Squares and Optimal Point Prediction

Least Squares and Optimal Point Prediction

- The linear regression DGP under the ideal conditions implies the conditional mean function,

$$E(y_i|x_{1i} = 1, x_{2i} = x_{2i}^*, \dots, x_{Ki} = x_{Ki}^*) = \beta_1 + \beta_2 x_{2i}^* + \cdots + \beta_K x_{Ki}^*$$

OR

$$E(y_i|x_i = x_i^*) = x_{Ki}^{*\prime}\beta$$

- AS a major goal in econometrics is predicting y. The question is "If a new person arrives with characteristics x*, what is my minimum mean squared error (MSE) prediction of her y?
- It turns out, very intuitively, that the answer is $E(y_i|x_i = x_i^*) = x_{Ki}^{*\prime}\beta$
- That is, "the conditional mean is the minimum MSE (point) predictor". (Indeed if it were anything else you'd surely be suspicious.)
- The non-operational version (i.e., pretending that we know β) is $E(y_i|x_i = x_i^*) = x_{Ki}^{*\prime}\beta$, and the operational version (using $\hat{\beta}_{LS}$) is $E(y_i|\widehat{x_i = x_i^*}) = x_{Ki}^{*\prime}\beta$

Least Squares and Optimal Point Prediction

- A very basic and powerful result.
- Notice that the β's in the conditional mean expression give the weights on the various x's for forming the optimal predictor. Hence under the IC, consistency of OLS ensures that asymptotically the operational point prediction (based on $\hat{\beta}_{LS}$) will use the right weights (based on β). That is, under the IC, LS is consistent for the right predictive weights. Now here's the really amazing thing (although it's obvious when you think about it):
- *Under great generality, in particular even if the IC fail, LS is still consistent for the right predictive weights, simply by virtue of the MSE-optimization problem that it solves directly.*
- The bottom line:
- *Forecasting is of central importance in economics, and LS regression delivers optimal forecasts under great generality.*

Least Squares and Optimal Point Prediction

- If LS provides optimal forecasts even without the IC, you might wonder why we introduced the IC. There are two key sets of reasons.
- **First,** even for standard forecasting situations of the form "If a new person arrives with characteristics x^*, what is my minimum-MSE prediction of her y," once we drop the IC, so that the fitted model does not necessarily match the true DGP, there is a crucial issue of what model to use. Many questions arise.
- Which x's should we include, and which should we exclude? Is a linear model really adequate, or should we incorporate some non-linearlity? And so on. For any given model, LS will deliver the optimal parameter configuration for forecasting, but again, a crucial issue is what features a "good" or "the best" model should incorporate.

Least Squares and Optimal Point Prediction

- **Second,,** what we've considered so far is called "non-causal" prediction. It exploits correlation between y and x to generate forecasts, but there is no presumption (or need) that x truly causes y in a deep scientific sense.

- (Remember, correlation does not imply causation!) But there is a causal form of prediction that differs from the one sketched thus far. In particular, thus far we've considered "If a new person arrives with characteristics $x*$, what is my minimum-MSE prediction of her y?", but we might alternatively be interested in predicting the effects of an active treatment, or intervention, or policy, along the lines of "If I randomly select someone and change her characteristics in some way, what is my minimum-MSE prediction of the corresponding change in her y?" It turns out that LS does not always perform well for such "causal prediction" questions. So when does LS perform well for causal prediction? Under the IC! Effectively LS solves both the non-causal and causal prediction problems under the IC (or, put differently, the two problems are identical under the IC), but when the IC fail LS continues to solve the non-causal prediction problem but fails for the causal prediction problem.

Least Squares and Optimal Point Prediction

- **Summarizing, here's what true:**
- 1. Non-causal prediction is important in economics
- 2. LS succeeds for non-causal prediction under great generality
- 3. Causal prediction is important in economics
- 4. LS fails for causal prediction unless the IC hold, so credible causal prediction is much harder.
- Given the combination of 1 and 2 above, it makes obvious sense to start with with non-causal prediction and treat it extensively, reserving 4 for separate treatment. That has been the successful strategy of econometrics for many decades, and it is very much at the center of modern "data science" and "machine learning".

Optimal Interval and Density Prediction

Optimal Interval and Density Prediction

- Prediction as introduced thus far is so-called point prediction (a single best – i.e., minimum MSE – guess).
- Forecasts stated as confidence intervals ("interval forecasts") are also of interest. The linear regression DGP under the IC implies the conditional variance function

$$var(y_i|x_i = x_i^*) = \sigma^2$$

- which we can use to form interval forecasts. The non-operational version is

$$y_i \in \left[x_{Ki}^{*\prime}\beta \pm 1.96\,\sigma\right];\ \ w.p.\,0.95$$

- and the operational version is

$$y_i \in \left[x_{Ki}^{*\prime}\beta \pm 1.96\,s\right];\ \ w.p.\,0.95$$

- Finally full density forecasts are of interest. The linear regression DGP under the IC implies the conditional density function

$$y_i|\ x_i = x_i^* \sim N(x_{Ki}^{*\prime}\beta, \sigma^2)$$

Optimal Interval and Density Prediction

- Hence a non-operational density forecast is

$$y_i | \, x_i = x_i^* \sim N(x_{Ki}^{*\prime}\beta, \sigma^2)$$

- and the operational version is

$$y_i | \, x_i = x_i^* \sim N(x_{Ki}^{*\prime}\beta, s^2)$$

- Notice that the interval and density forecasts rely for validity on more parts of the IC than do the point forecasts: Gaussian disturbances and constant disturbance variances – which makes clear in even more depth why violations of the IC are generally problematic even in non-causal forecasting situations.

Regression Output from a Predictive Perspective

Regression Output from a Predictive Perspective

- In light of our predictive emphasis, here we offer some predictive perspective on the regression statistics discussed earlier.
- The sample, or historical, mean of the dependent variable, $\bar{y}$, an estimate of the *unconditional* mean of y, is a benchmark forecast. It is obtained by regressing y on an intercept alone – no conditioning on other regressors.
- The sample standard deviation of y is a measure of the in-sample accuracy of the unconditional mean forecast $\bar{y}$.
- The OLS fitted values, $y_i = x'_i\hat{\beta}$ are effectively in-sample regression predictions.
- The OLS residuals, $e_i = y_i - \hat{y}_i$, are effectively in-sample prediction errors corresponding to use of those in-sample regression predictions.
- OLS coefficient signs and sizes relate to the weights put on the various x variables in forming the best in-sample prediction of y.

Regression Output from a Predictive Perspective

- OLS coefficient signs and sizes relate to the weights put on the various x variables in forming the best in-sample prediction of y.
- The standard errors, t statistics, and p-values let us do statistical inference as to which regressors are most relevant for predicting y.
- SSR measures "total" in-sample accuracy of the regression predictions. It is closely related to in-sample MSE:

$$MSE = \frac{1}{N} SSR = \frac{1}{N} \sum_{i=1}^{N} e_i^2$$

- ("average" in-sample accuracy of the regression predictions).
- The F statistic effectively compares the accuracy of the regression-based forecast to that of the unconditional-mean forecast. It helps us assess whether the x variables, taken as a set, have predictive value for y. That contrasts with the t statistics, which assess predictive value of the x variables one at a time.

Regression Output from a Predictive Perspective

- s^2 is just SSR scaled by $N - K$, so again, it's a measure of the in-sample accuracy of the regression-based forecast. It's like MSE, but corrected for degrees of freedom.
- R^2 and $\bar{R}^2$ effectively compare the in-sample accuracy of conditional-mean ($x'_i = \hat{\beta}$ and unconditional-mean $\bar{y}$ forecasts.
- R^2 is not corrected for d.f. and has MSE on top:

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - x_i'\,\hat{\beta}\right)^2}{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}$$

- In contrast, $\bar{R}^2$ is corrected for d.f. and has s^2 on top:

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-K}\sum_{i=1}^{N}\left(y_i - x_i'\,\hat{\beta}\right)^2}{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}$$

- Residual plots are useful for visually flagging neglected things that impact forecasting. Residual correlation (in time-series contexts) indicates that point forecasts could possibly be improved. Non-constant residual volatility indicates that interval and density forecasts could be possibly improved.

Multicollinearity

Multicollinearity

- Collinearity and multicollinearity don't really involve failure of the ideal conditions, but they nevertheless are sometimes issues and should be mentioned.
- Collinearity refers to two x variables that are highly correlated. But even if all pairwise correlations are small an x variable could nevertheless be highly correlated with a *linear combination* of other x variables.
- That raises the idea of multicollinearity, where an x variable is highly correlated with a *linear combination* of other x variables. Collinearity is of course a special case of multicollinearity, so henceforth we will simply speak of multicollinearity.

Multicollinearity

- **Perfect and Imperfect Multicollinearity**
- There are two types of multicollinearity, perfect and imperfect.
- Perfect multicollinearity refers to perfect correlation among some regressors, or linear combinations of regressors. Perfect multicollinearity is indeed a problem; the $X'X$ matrix is singular, so $(X'X)^{-1}$ does not exist, and the OLS estimator cannot even be computed!
- Perfect multicollinearity is disastrous, but it's unlikely to occur unless you do something really silly, like entering the same regressor twice. In any event the solution is trivial: simply drop one of the redundant variables.
- Imperfect multicollinearity, in contrast, occurs routinely but is not necessarily problematic, although in extreme cases it may require some attention.
- Imperfect collinearity/multicollinearity refers to (imperfect) correlation among some regressors, or linear combinations of regressors.

Multicollinearity

- Imperfect multicollinearity is not a "problem" in the sense that something was done incorrectly, and it is not a violation of the IC. Rather, it just reflects the nature of economic and financial data. But we still need to be aware of it and understand its effects.
- Telltale symptoms are large F and R2, yet small t's (large s.e.'s), and/or coefficients that are sensitive to small changes in sample period. That is, OLS has trouble parsing individual influences, yet it's clear that there is an overall relationship. OLS is in some sense just what the doctor ordered – orthogonal projection.
- It can be shown, and it is very intuitive, that

$$var(\hat{\beta}_k) = f\left(\underbrace{\sigma^2}_{+}, \underbrace{\sigma^2_{xk}}_{+}, \underbrace{R^2_k}_{+}\right)$$

- where R^2_k is the R^2 from a regression of x_k on all other regressors.
- In the limit, as $R^2_k \to 1$, $var(\hat{\beta}_k) \to \infty$, because x_k is then perfectly "explained" by the other variables and is therefore completely redundant. R^2_k is effectively a measure of the "strength" of the multicollinearity affecting β_k.

Multicollinearity

- We often measure the strength of multicollinearity by the "variance inflation factor",

$$VIF(\hat{\beta}_k) = \frac{1}{1 - R_k^2}$$

- which is just a transformation of R_k^2.

Beyond Fitting the Conditional Mean: Quantile regression

Beyond Fitting the Conditional Mean: Quantile regression

- Recall that the OLS estimator, $\hat{\beta}_{OLS}$, solves:

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_2 t \ - \cdots - \beta_K x_{kt})^2 = \min_{\beta} \sum_{i=1}^{N} (\varepsilon_i)^2$$

- The solution has a simple analytic closed-form expression, $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$, with wonderful properties under the IC (unbiased, consistent, Gaussian, MVUE). But other objectives are possible and sometimes useful.

- So-called quantile regression (QR) involves an objective function linear on each side of 0 but with (generally) unequal slopes. QR estimator $\hat{\beta}_{QR}$ minimizes "linlin loss," or "check function loss":

$$\min_{\beta} \sum_{i=1}^{N} linlin(\varepsilon_i)^2$$

- where:

$$linlin(e) = \begin{Bmatrix} a|e| & if\ e \leq 0 \\ b|e| & if\ e > 0 \end{Bmatrix} = a|e|I(e \leq 0) + b|e|I(e > 0)$$

Beyond Fitting the Conditional Mean: Quantile regression

- $I(.)$ stands for "indicator" variable where $I(x) = 1$ if x is true, and $I(x) = 0$ otherwise. "linlin" refers to linearity on each side of the origin.
- QR is not as simple as OLS, but it is still simple (solves a linear programming problem).
- A key issue is what, precisely, quantile regression fits. QR fits the d ·100% quantile:

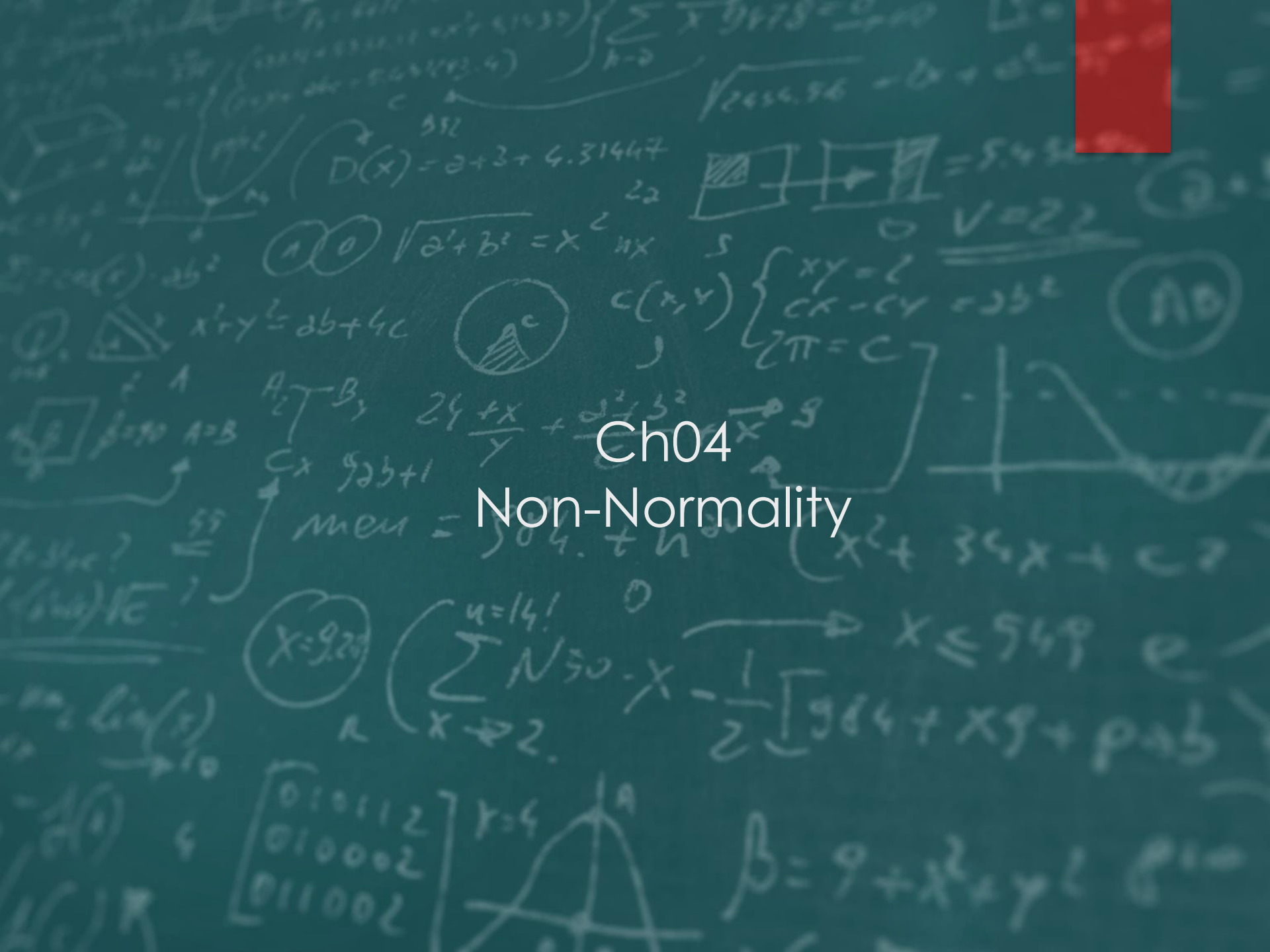
$$quantile_d(y|X) = x\beta$$

- where:

$$d = \frac{b}{a+b} = \frac{1}{1+\frac{a}{b}}$$

- This is an important generalization of regression (e.g., How do the wages of people in the far left tail of the wage distribution vary with education and experience, and how does that compare to those in the center of the wage distribution?)

Exercises



Ch04
Non-Normality

Non-Normality

Non-normality and outliers, which we introduce in this chapter, are closely related, because deviations from Gaussian behavior are often characterized by fatter tails than the Gaussian, which produce outliers. It is important to note that outliers are not necessarily "bad," or requiring "treatment."

Every data set must have some most extreme observation, by definition! Statistical estimation efficiency, moreover, increases with data variability.

The most extreme observations can be the most informative about the phenomena of interest. "Bad" outliers, in contrast, are those associated with things like data recording errors (e.g., you enter .753 when you mean to enter 75.3) or one-off events (e.g., a strike or natural disaster).

Results

- To understand the properties of OLS without normality, it is helpful first to consider the properties of the sample mean without normality.
- for a non-Gaussian simple random sample,

$$y_i \sim iid\ (\mu, \sigma^2), i = 1, \ldots, N$$

- we have that the sample mean is consistent, asymptotically normal, and asymptotically efficient, with

$$\bar{y} \underset{\sim}{a} N\left(\mu, \frac{\sigma^2}{N}\right)$$

- This result forms the basis for asymptotic inference. It is a Gaussian central limit theorem. We consistently estimate σ^2 using s^2.
- Now consider the linear regression under the IC except that we allow non- Gaussian disturbances. OLS remains consistent, asymptotically normal, and asymptotically efficient, with

$$\hat{\beta}_{ols} \underset{\sim}{a} N(\beta, V)$$

- We consistently estimate the covariance matrix V using $s^2(X'X)^{-1}$.

Assessing Normality

- There are many methods, ranging from graphics to formal tests.
- **1. QQ Plots**
- We introduced histograms earlier in Chapter 2 as a graphical device for learning about distributional shape. If, however, interest centers on the tails of distributions, QQ plots often provide sharper insight as to the agreement or divergence between the actual and reference distributions.
- The QQ plot is simply a plot of the quantiles of the standardized data against the quantiles of a standardized reference distribution (e.g., normal). If the distributions match, the QQ plot is the 45 degree line. To the extent that the QQ plot does not match the 45 degree line, the nature of the divergence can be very informative, as for example in indicating fat tails.

Assessing Normality

2. Residual Sample Skewness and Kurtosis

- Recall skewness and kurtosis, which we reproduce here for convenience:

$$S = \frac{E(y-\mu)^3}{\sigma^3}$$

$$K = \frac{E(y-\mu)^4}{\sigma^4}$$

- Obviously, each tells about a different aspect of non-normality. Kurtosis, in particular, tells about fatness of distributional tails relative to the normal.
- A simple strategy is to check various implications of residual normality, such as S = 0 and K = 3, via informal examination of $\hat{S}$ and $\hat{K}$.

Assessing Normality

3. The Jarque-Bera Test

The Jarque-Bera test (JB) effectively aggregates the information in the data about both skewness and kurtosis to produce an overall test of the joint hypothesis that $S = 0$ and $K = 3$, based upon $\hat{S}$ and $\hat{K}$.

The test statistic is

$$JB = \frac{N}{6}\left(\hat{S} + \frac{1}{4}\left(\hat{K} - 3\right)^2\right)$$

Under the null hypothesis of independent normally-distributed observations ($S = 0, K = 3$), JB is distributed in large samples as a χ^2 random variable with two degrees of freedom.

Outliers

- Outliers refer to big disturbances (in population) or residuals (in sample). Outliers may emerge for a variety of reasons, and they may require special attention because they can have substantial influence on the fitted regression line.
- On the one hand, OLS retains its magic in such outlier situations – it is Best Linear Unbiased Estimator (BLUE) regardless of the disturbance distribution. On the other hand, the fully-optimal (MVUE) estimator may be highly non-linear, so the fact that OLS remains BLUE is less than fully comforting. Indeed OLS parameter estimates are particularly susceptible to distortions from outliers, because the quadratic least-squares objective really hates big errors (due to the squaring) and so goes out of its way to tilt the fitted surface in a way that minimizes them.
- How to identify and treat outliers is a time-honored problem in data analysis, and there's no easy answer. If an outlier is simply a data-recording mistake, then it may well be best to discard it if you can't obtain the correct data. On the other hand, every dataset, even a perfectly "clean" dataset, has a "most extreme observation," but it doesn't follow that it should be discarded. Indeed the most extreme observations are often the most informative – precise estimation requires data variation.

Outliers

Outlier Detection

1. Graphics

One obvious way to identify outliers in bivariate regression situations is via graphics: one xy scatterplot can be worth a thousand words. In higher dimensions, the residual $\hat{y}y$ scatterplot remains invaluable, as does the residual plot of $y - \hat{y}$.

2. Leave-One-Out and Leverage

Another way to identify outliers is a "leave-one-out" coefficient plot, where we use the computer to sweep through the sample, leaving out successive observations, and examining differences in parameter estimates with observation various observations "in" vs. "out". That is, in an obvious notation, we examine and plot

$$\hat{\beta}_{ols}^{(-i)} - \hat{\beta}_{ols} = -\frac{1}{1-h_i}(X'X)^{-1}x_i'e_i$$

where h_i is the i-th diagonal element of the "*hat matrix*," $X(X'X)^{-1}X'$. Hence the estimated coefficient change $\hat{\beta}_{ols}^{(-i)} - \hat{\beta}_{ols}$ is driven by $\frac{1}{1-h_i}$. Also, h_i is called the observation-i **leverage**. h_i can be shown to be in [0, 1], so that the larger is h_i, the larger is $\hat{\beta}_{ols}^{(-i)} - \hat{\beta}_{ols}$.

Hence one really just needs to examine the leverage sequence, and scrutinize carefully observations with high leverage.

Robust Estimation

- Robust estimation provides a useful middle ground between completely discarding allegedly-outlying observations ("dummying them out") and doing nothing.
- Here we introduce outlier-robust approaches to regression. The first involves OLS regression, but on weighted data, an the second involves switching from OLS to a different estimator.
- **1. Robustness Iteration**
- Fit at robustness iteration 0:

$$\hat{y}^{(0)} = X\hat{\beta}^{(0)}$$

- Where

$$\hat{\beta}^{(0)} = arg\min_{\beta}\left[\sum_{i=1}^{N}(y_i - x_i'\beta)^2\right]$$

Robust Estimation

- Robustness weight at iteration 1:

$$\rho_i^{(1)} = S\left(\frac{e_i^{(0)}}{6\ med\ |e_i^{(0)}|}\right)$$

- Where $e_i^{(0)} = y_i - \hat{y}_i{}^{(0)}$
- and $S(z)$ is a function such that $S(z) = 1$ for $z \in [-1, 1]$ but down weights outside that interval.
- Fit at robustness iteration 1:

$$\hat{y}^{(1)} = X\hat{\beta}^{(1)}$$

- Where

$$\hat{\beta}^{(1)} = arg\ \min_{\beta}\left[\sum_{i=1}^{N} \rho_i^{(1)}(y_i - x_i'\beta)^2\right]$$

Continue as desired.

Robust Estimation

2. Least Absolute Deviations

- Recall that the OLS estimator solves
- $\min_{\beta} \sum_{i=1}^{N} (y_i - x_i'\beta)^2$
- Now we simply change the objective to

$$\min_{\beta} \sum_{i=1}^{N} (y_i - x_i'\beta) \quad \text{or} \qquad \min_{\beta} \sum_{i=1}^{N} (|\varepsilon_i|)$$

- That is, we change from squared-error loss to absolute-error loss.
- We call the new estimator "least absolute deviations" (LAD) and we write $\hat{\beta}_{LAD}$.
- By construction, $\hat{\beta}_{LAD}$ is not influenced by outliers as much as $\hat{\beta}_{OLS}$. Put differently, LAD is more robust to outliers than is OLS.

Robust Estimation

- Of course nothing is free, and the price of LAD is a bit of extra computational complexity relative to OLS. In particular, the LAD estimator does not have a tidy closed-form analytical expression like OLS, so we can't just plug into a simple formula to obtain it. Instead we need to use the computer to find the optimal β directly. If that sounds complicated, rest assured that
- it's largely trivial using modern numerical methods, as embedded in modern software.
- It is important to note that whereas OLS fits the conditional mean function

$$mean(y|X) = X\beta$$

- LAD fits the conditional median function (50% quantile): $median(y|X) = X\beta$
- The conditional mean and median are equal under symmetry and hence under normality, but not under asymmetry, in which case the median is a better measure of central tendency. Hence LAD delivers two kinds of robustness to non-normality: it is robust to outliers and robust to asymmetry.

Wage Determination

Wage Determination

- Here we show some empirical results that make use of the ideas sketched above. There are many tables and figures appearing at the end of the chapter.
- We do not refer to them explicitly, but all will be clear upon examination.
- We run WAGE → c, EDUC, EXPER. We show the regression results, the residual plot, the residual histogram and statistics, the residual Gaussian QQ plot, the leave-one-out plot, and the results of LAD estimation. The residual plot shows lots of positive outliers, and the residual histogram and Gaussian QQ plot indicate right-skewed residuals.
- **1. LWAGE**
- Now we run LWAGE → c, EDUC, EXPER. Again we show the regression results, the residual plot, the residual histogram and statistics, the residual Gaussian QQ plot, the leave-one-out plot, and the results of LAD estimation. Among other things, and in sharp contrast to the results for WAGE and opposed to LWAGE, the residual histogram and Gaussian QQ plot indicate approximate residual normality.

Wage Determination

Equation: UNTITLED Workfile: DATAWAGES::Datawa...

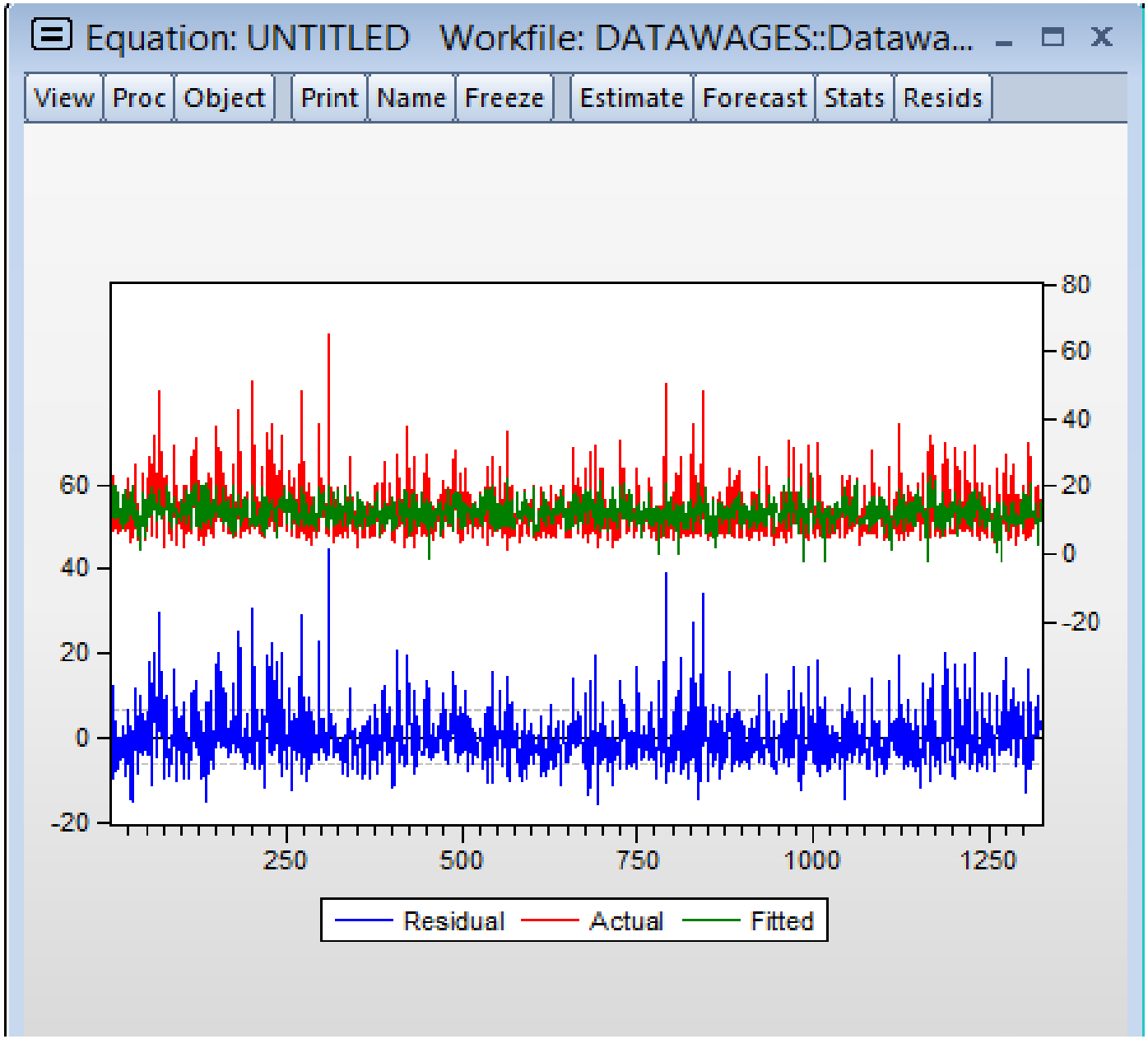
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: WAGE
Method: Least Squares
Date:
Sample: 1 1323
Included observations: 1323

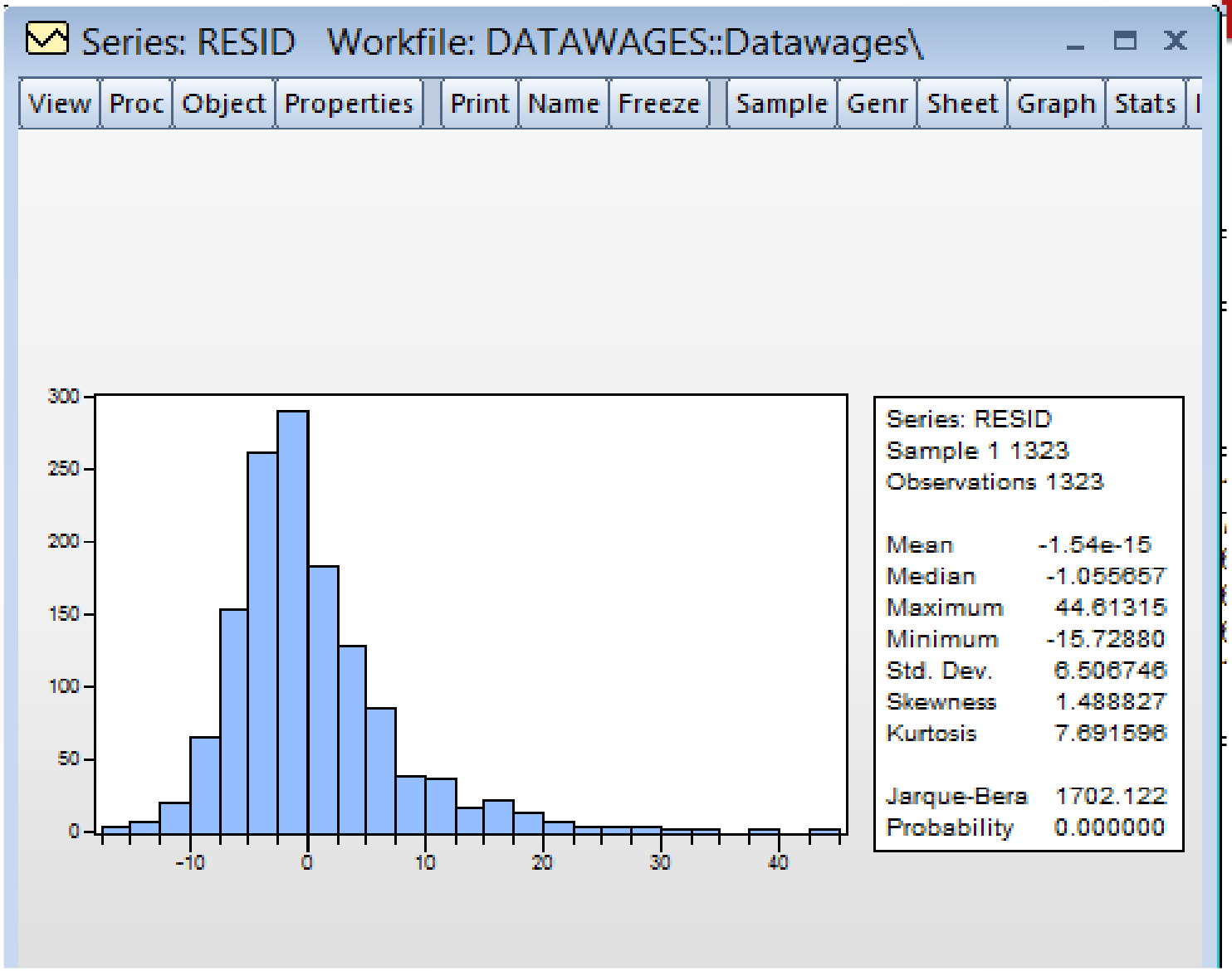
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-6.891639	0.996367	-6.916771	0.0000
EDUC	1.219441	0.066726	18.27527	0.0000
EXPER	0.160537	0.015390	10.43148	0.0000

R-squared	0.223589	Mean dependent var	12.18780
Adjusted R-squared	0.222413	S.D. dependent var	7.384450
S.E. of regression	6.511674	Akaike info criterion	6.587335
Sum squared resid	55970.50	Schwarz criterion	6.599099
Log likelihood	-4354.522	Hannan-Quinn criter.	6.591745
F-statistic	190.0658	Durbin-Watson stat	1.933125
Prob(F-statistic)	0.000000		

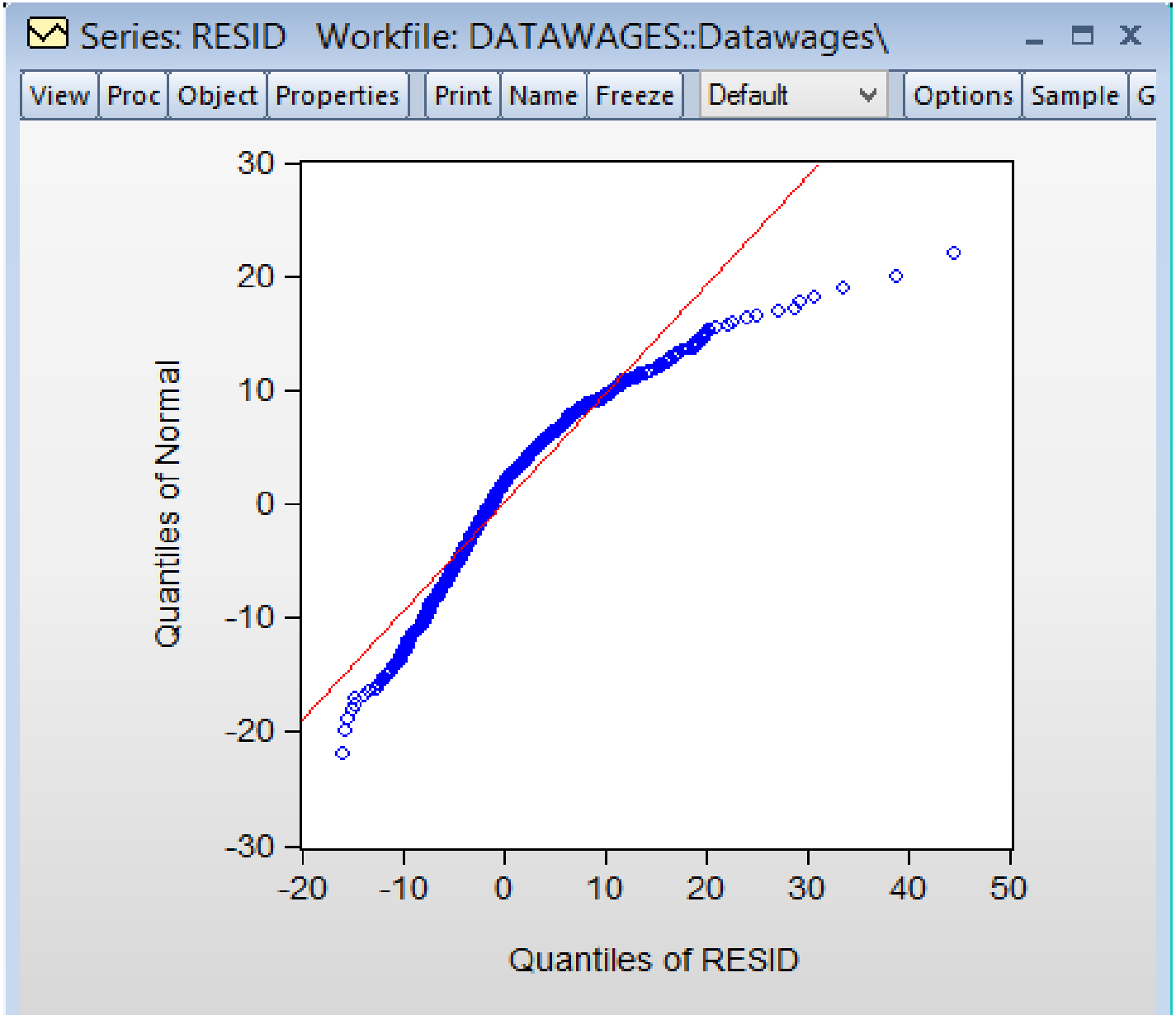
Wage Determination



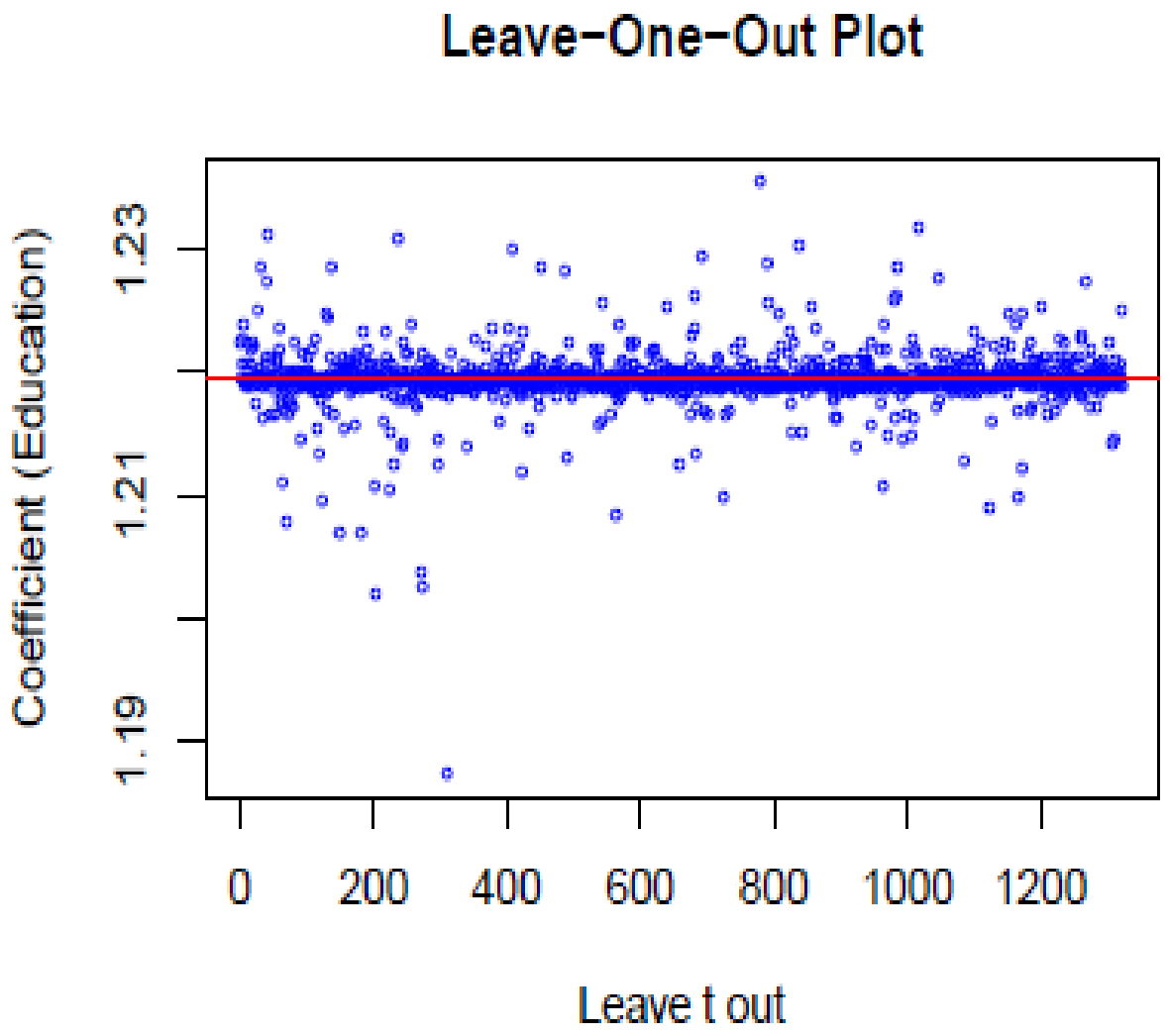
Wage Determination



Wage Determination



Wage Determination



Wage Determination

Equation: UNTITLED Workfile: DATAWAGES::Datawa...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: WAGE
Method: Quantile Regression (Median)
Date:
Sample: 1 1323
Included observations: 1323
Huber Sandwich Standard Errors & Covariance
Sparsity method: Kernel (Epanechnikov) using residuals
Bandwidth method: Hall-Sheather, bw=0.088501
Estimation successfully identifies unique optimal solution

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-6.030000	1.107366	-5.445356	0.0000
EDUC	1.050000	0.087201	12.04110	0.0000
EXPER	0.170000	0.013905	12.22564	0.0000

Pseudo R-squared	0.138727	Mean dependent var	12.18780
Adjusted R-squared	0.137422	S.D. dependent var	7.384450
S.E. of regression	6.636805	Objective	3026.949
Quantile dependent var	10.00000	Restr. objective	3514.505
Sparsity	11.65584	Quasi-LR statistic	334.6342
Prob(Quasi-LR stat)	0.000000		

Wage Determination

Equation: UNTITLED Workfile: GRAPHS::Untitled\

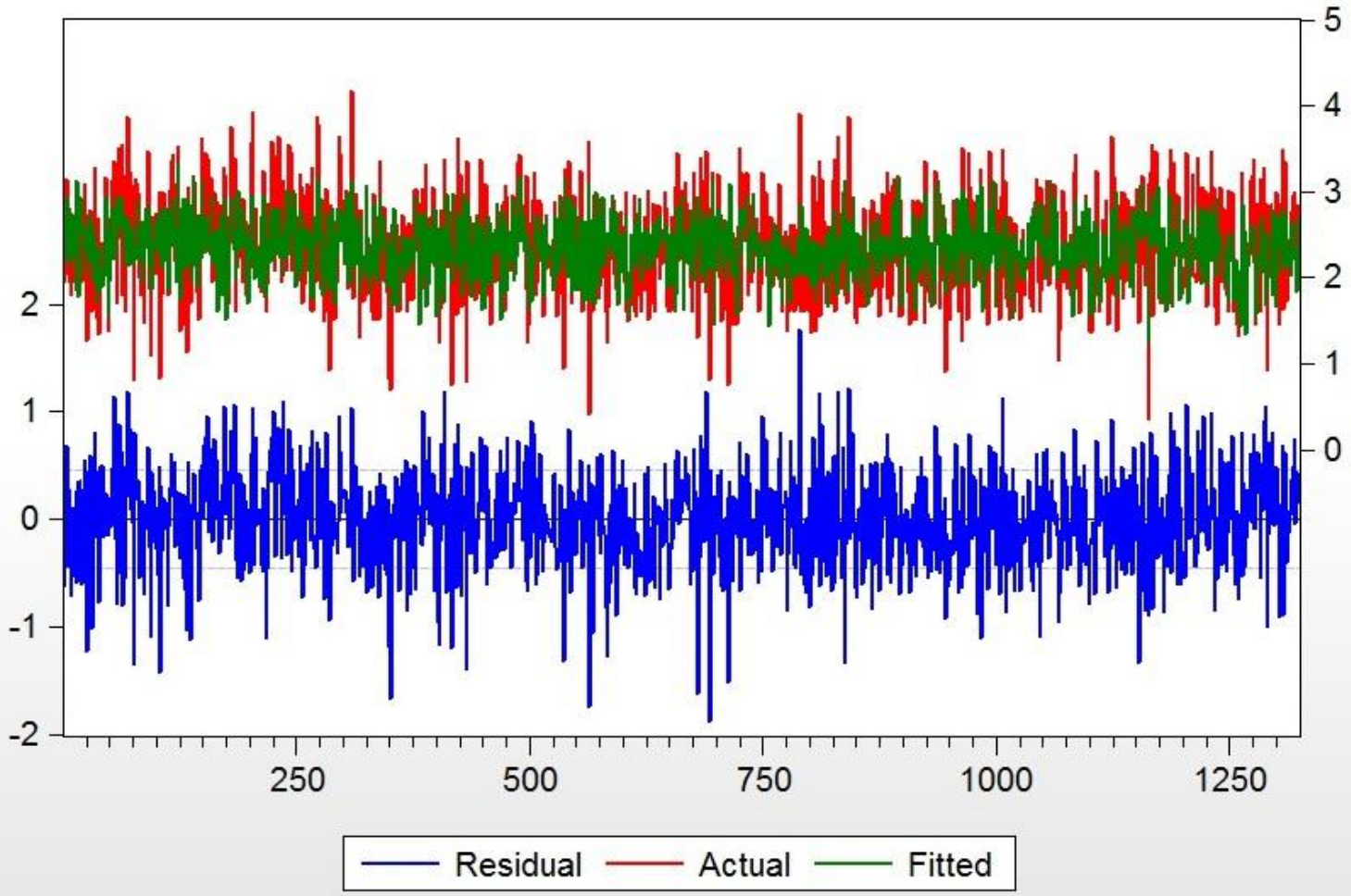
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LWAGE
Method: Least Squares
Date:
Sample (adjusted): 1 1323
Included observations: 1323 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.867382	0.075331	11.51431	0.0000
EDUC	0.093229	0.005045	18.48002	0.0000
EXPER	0.013104	0.001164	11.26208	0.0000

R-squared	0.232224	Mean dependent var	2.341995
Adjusted R-squared	0.231061	S.D. dependent var	0.561435
S.E. of regression	0.492318	Akaike info criterion	1.422881
Sum squared resid	319.9376	Schwarz criterion	1.434644
Log likelihood	-938.2358	Hannan-Quinn criter.	1.427291
F-statistic	199.6260	Durbin-Watson stat	1.926045
Prob(F-statistic)	0.000000		

Wage Determination



Exercises

Ch05
Group Heterogeneity and Indicator Variables

Group Heterogeneity and Indicator Variables

From one perspective we continue working under the IC.

From another we now begin relaxing the IC, effectively by recognizing RHS variables that were omitted from, but should not have been omitted from, our original wage regression.

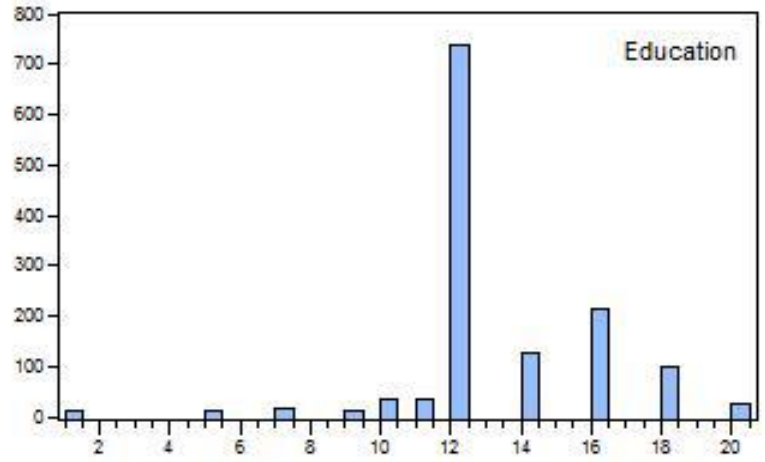
0-1 Dummy Variables

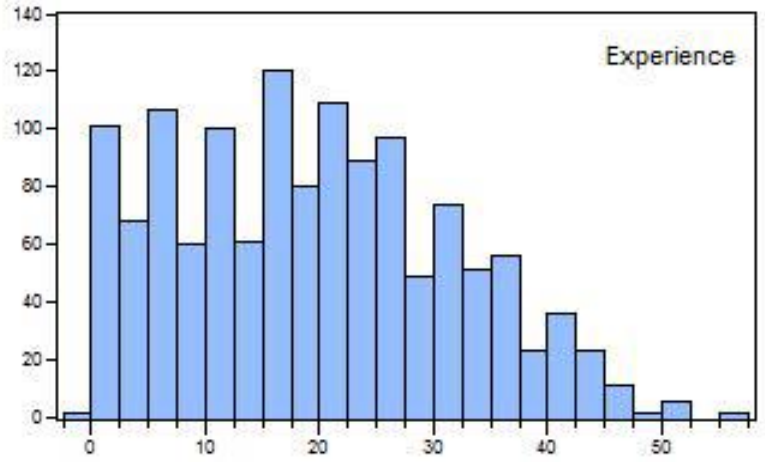
- A dummy variable, or indicator variable, is just a 0-1 variable that indicates something, such as whether a person is female, non-white, or a union member. We use dummy variables to account for such "group effects," if any. We might define the dummy UNION, for example, to be 1 if a person is a union member, and 0 otherwise. That is,

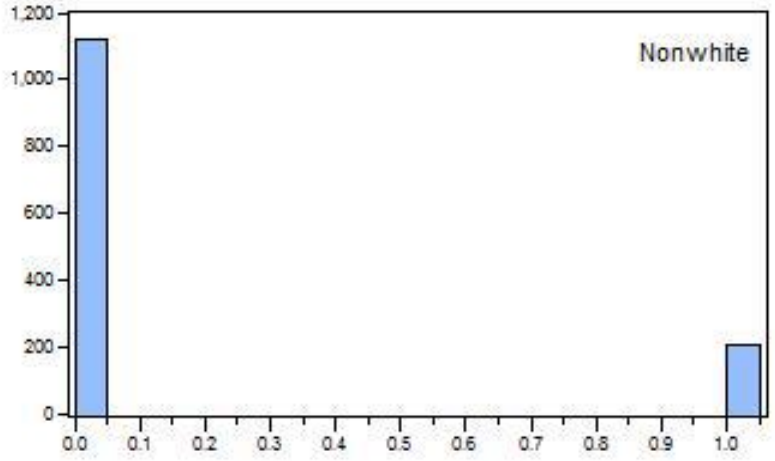
- $$UNION_i = \begin{cases} 1 & \text{if observation } i \text{ corresponds to a union member} \\ 0 & otherwise. \end{cases}$$

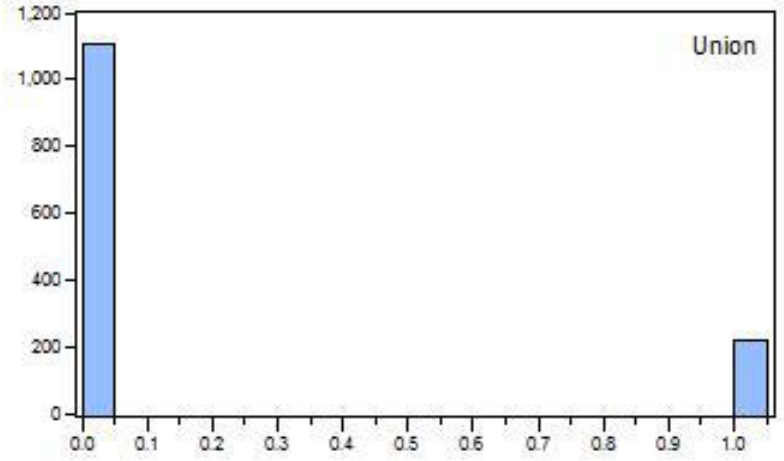
- In Figure below we show histograms and statistics for all potential determinants of wages. Education (EDUC) and experience (EXPER) are standard continuous variables, although we measure them only discretely (in years);

0-1 Dummy Variables









0-1 Dummy Variables

- we have examined them before and there is nothing new to say. The new variables are 0-1 dummies, UNION (already defined) and NONWHITE, where

- $\text{NONWHITE}_i = \begin{cases} 1 & \text{if observation } i \text{ corresponds to a non} - \text{white person} \\ 0 & otherwise. \end{cases}$

- Note that the sample mean of a dummy variable is the fraction of the sample with the indicated attribute. The histograms indicate that roughly one-fifth of people in our sample are union members, and roughly one-fifth are non-white.

- We also have a third dummy, FEMALE, where

- $\text{FEMALE}_i = \begin{cases} 1 & \text{if observation } i \text{ corresponds to aa female} \\ 0 & otherwise. \end{cases}$

- We don't show its histogram because it's obvious that FEMALE should be approximately 0 w.p. $1/2$ and 1 $w.p.\ 1/2$, which it is.

0-1 Dummy Variables

- Sometimes dummies like UNION, NONWHITE and FEMALE are called intercept dummies, because they effectively allow for a different intercept for each group (union vs. non-union, non-white vs. white, female vs. male).
- The regression intercept corresponds to the "base case" (zero values for all dummies) and the dummy coefficients give the extra effects when the respective dummies equal one. For example, in a wage regression with an intercept and a single dummy (UNION, say), the intercept corresponds to non-union members, and the estimated coefficient on UNION is the extra effect (up or down) on LWAGE accruing to union members.
- Alternatively, we could define and use a full set of dummies for each category (e.g., include both a union dummy and a non-union dummy) and drop the intercept, reading off the union and non-union effects directly.
- In any event, never include a full set of dummies and an intercept. Doing so would be redundant because the sum of a full set of dummies is just a unit vector, but that's what the intercept is. If an intercept is included, one of the dummy categories must be dropped.

Group Dummies in the Wage Regression

Equation: UNTITLED Workfile: GRAPHS::Untitled\

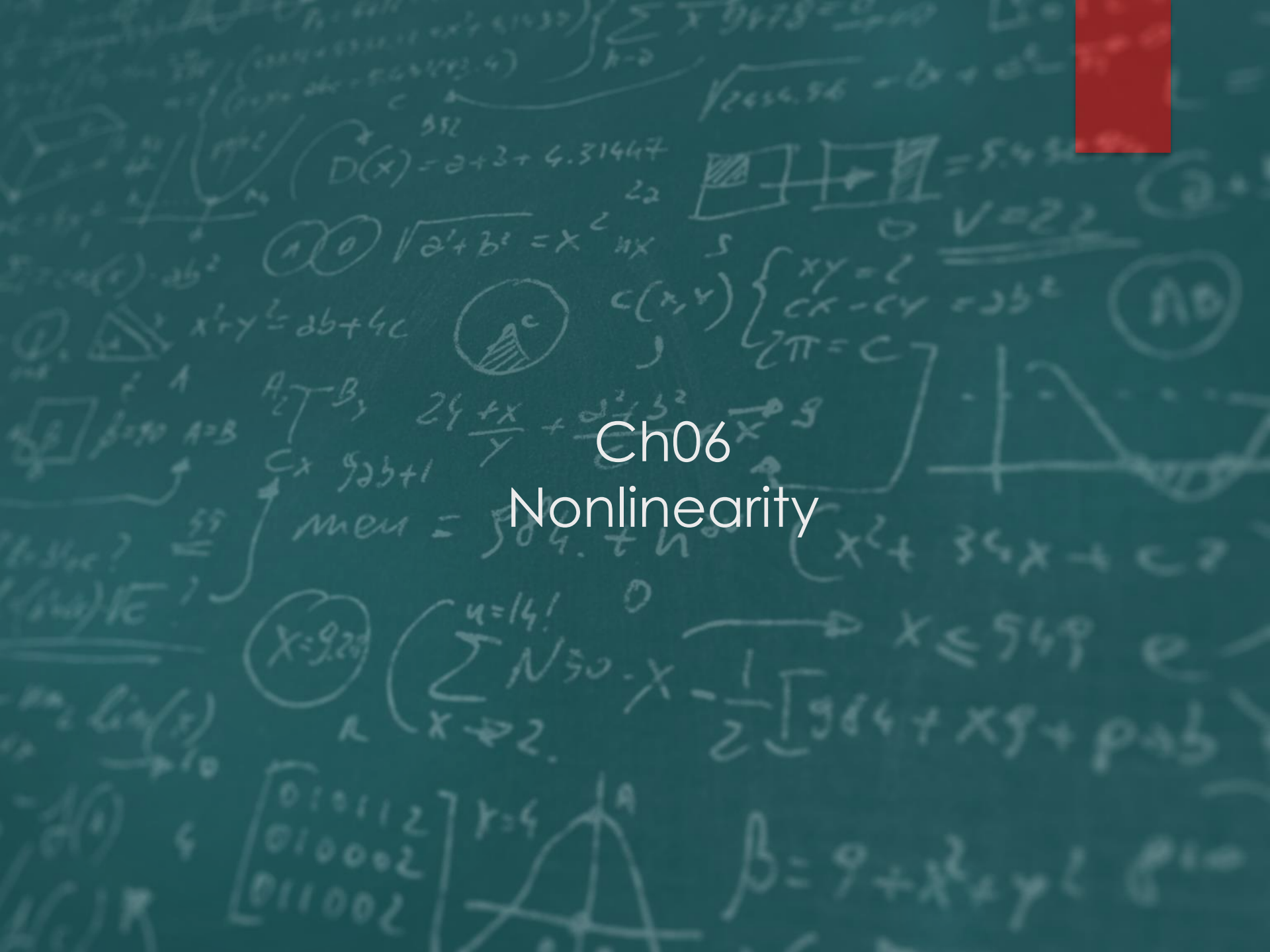
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LWAGE
Method: Least Squares
Date:
Sample (adjusted): 1 1323
Included observations: 1323 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.000385	0.073180	13.67013	0.0000
EDUC	0.090809	0.004814	18.86314	0.0000
EXPER	0.012707	0.001119	11.35624	0.0000
FEMALE	-0.237535	0.025965	-9.148397	0.0000
NONWHITE	-0.085286	0.035786	-2.383199	0.0173
UNION	0.223392	0.035307	6.327126	0.0000

R-squared	0.307856	Mean dependent var	2.341995
Adjusted R-squared	0.305229	S.D. dependent var	0.561435
S.E. of regression	0.467973	Akaike info criterion	1.323712
Sum squared resid	288.4212	Schwarz criterion	1.347239
Log likelihood	-869.6356	Hannan-Quinn criter.	1.332532
F-statistic	117.1568	Durbin-Watson stat	1.910120
Prob(F-statistic)	0.000000		

Exercises



Ch06
Nonlinearity

Nonlinearity

In general there is no reason why the conditional mean function should be linear. That is, the appropriate functional form may not be linear. Whether linearity provides an adequate approximation is an empirical matter. Non-linearity is related to non-normality, which we studied in chapter 4. In particular, in the multivariate normal case, the conditional mean function is linear in the conditioning variables. But once we leave the terra firma of multivariate normality, anything goes. The conditional mean function and disturbances may be linear and Gaussian, non-linear and Gaussian, linear and non-Gaussian, or non-linear and non-Gaussian. In the Gaussian case, because the conditional mean is a linear function of the conditioning variable(s), it coincides with the linear projection. In non-Gaussian cases, however, linear projections are best viewed as approximations to generally non-linear conditional mean functions. That is, we can view the linear regression model as a linear approximation to a generally nonlinear conditional mean function. Sometimes the linear approximation may be adequate, and sometimes not.

1. Logarithms

- Models can be non-linear but nevertheless linear in non-linearly-transformed variables. A leading example involves logarithms, to which we now turn. This can be very convenient. Moreover, coefficient interpretations are special, and similarly convenient.
- Logs turn multiplicative models additive, and they neutralize exponentials. Logarithmic models, although non-linear, are nevertheless "linear in logs."
- In addition to turning certain non-linear models linear, they can be used to enforce non-negativity of a left-hand-side variable and to stabilize a disturbance variance. (More on that later.)

1. Log-Log Regression

- First, consider log-log regression. We write it out for the simple regression case, but of course we could have more than one regressor. We have

$$y_i = \beta_1 + \beta_2 \ln(x_i) + \varepsilon_i$$

- y_i is a non-linear function of the xi , but the function is linear in logarithms, so that ordinary least squares may be applied.

1. Logarithms

- To take a simple example, consider a Cobb-Douglas production function with output a function of labor and capital,

$$y_i = AL_i^{\alpha} K_i^{\beta} \exp(\varepsilon_i)$$

- Direct estimation of the parameters A, α, β would require special techniques. Taking logs, however, yields

$$\ln(y_i) = \ln(A) + \alpha \ln(L_i) + \beta \ln(K_i) + \varepsilon_i$$

- This transformed model can be immediately estimated by ordinary least squares. We simply regress $\ln(y_i)$ on an intercept, $\ln(L_i)$ and $\ln(K_i)$. Such log-log regressions often capture relevant non-linearities, while nevertheless maintaining the convenience of ordinary least squares.

- Note that the estimated intercept is an estimate of $\ln(A)$ (not A, so if you want an estimate of A you must exponentiate the estimated intercept), and the other estimated parameters are estimates of α and β, as desired.

1. Logarithms

- Recall that for close y_i and x_i, $(\ln y_i - \ln x_i)$ is approximately the percent difference between y_i and x_i. Hence the coefficients in log-log regressions give the expected percent change in $E(y_i|x_i)$ for a one-percent change in xi, the so-called elasticity of y_i with respect to x_i.

2. Log-Lin Regression

- Second, consider log-lin regression, in which $\ln y_i = \beta x_i + \varepsilon_i$. We have a log on the left but not on the right. The classic example involves the workhorse model of exponential growth:

$$Y_t = Ae^{rt}\varepsilon_i$$

- It's non-linear due to the exponential, but taking logs yields

$$\ln(y_t) = \ln(A) + rt + \varepsilon_t$$

- which is linear. The growth rate r gives the approximate percent change in $E(y_t|t)$ for a one-unit change in time (because logs appear only on the left).

1. Logarithms

3. Lin-Log Regression

- Finally, consider lin-log Regression:

$$y_i = \beta \ln(x_i) + \varepsilon_t$$

- It's a bit exotic but it sometimes arises. β gives the effect on $E(y_i|x_i)$ of a one-percent change in x_i, because logs appear only on the right.

Box-Cox and GLM

1. Box-Cox

The Box-Cox transformation generalizes log-lin regression. We have

$$\mathrm{B}(y_i) = \beta_1 + \beta_2 x_i + \varepsilon_t$$

Where $\mathrm{B}(y_i) = \frac{y_t^{\lambda} - 1}{\lambda}$

Hence $E(y_i|x_i) = \mathrm{B}^{-1}(\beta_1 + \beta_2 x_i)$

Because $\lim_{\lambda \to 0} \left(\frac{y_t^{\lambda} - 1}{\lambda} \right) = \ln(y_i)$

the Box-Cox model corresponds to the log-lin model in the special case of $\lambda = 0$.

Box-Cox and GLM

2. generalized linear model (GLM)

The so-called "generalized linear model" (GLM) provides an even more flexible framework. Almost all models with left-hand-side variable transformations are special cases of those allowed in the generalized linear model (GLM). In the GLM, we have

$$G(y_i) = \beta_1 + \beta_2 x_i + \varepsilon_t$$

So that $E(y_i|x_i) = G^{-1}(\beta_1 + \beta_2 x_i)$

Wide classes of "link functions" G can be entertained. Log-lin regression,for example, emerges when G(yi) = ln(yi), and Box-Cox regression emerges when $G(y_i) = \frac{y_t^{\lambda}-1}{\lambda}$

Intrinsically Non-Linear Models

- Sometimes we encounter intrinsically non-linear models. That is, there is no way to transform them to linearity, so that they can then be estimated simply by least squares, as we have always done so far.
- As an example, consider the logistic model,

$$y_i = \frac{1}{a+br^{x_i}} + \varepsilon_t$$

- with $0 < r < 1$. The precise shape of the logistic curve of course depends on the precise values of a, b and r, but its "S-shape" is often useful.
- The key point for our present purposes is that there is no simple transformation of y that produces a model linear in the transformed variables.

Intrinsically Non-Linear Models

1. Nonlinear Least Squares

- The least squares estimator is often called “ordinary” least squares, or OLS.
- As we saw earlier, the OLS estimator has a simple closed-form analytic expression, which makes it trivial to implement on modern computers. Its computation is fast and reliable.
- The adjective “ordinary” distinguishes ordinary least squares from more laborious strategies for finding the parameter configuration that minimizes the sum of squared residuals, such as the non-linear least squares (NLS) estimator. When we estimate by non-linear least squares, we use a computer to find the minimum of the sum of squared residual function directly, using numerical methods, by literally trying many (perhaps hundreds or even thousands) of different β values until we find those that appear to minimize the sum of squared residuals. This is not only more laborious (and hence slow), but also less reliable, as, for example, one may arrive at a minimum that is local but not global.

Intrinsically Non-Linear Models

- Why then would anyone ever use non-linear least squares as opposed to OLS? Indeed, when OLS is feasible, we generally do prefer it. For example, in all regression models discussed thus far OLS is applicable, so we prefer it. Intrinsically non-linear models can't be estimated using OLS, but they can be estimated using non-linear least squares. We resort to non-linear least squares in such cases.
- Intrinsically non-linear models obviously violate the linearity assumption of the IC. But the violation is not a big deal. Under the remaining IC (thatis, dropping only linearity), $\hat{}$ βNLS has a sampling distribution similar to that under the IC.

Series Expansions

- There is really no such thing as an intrinsically non-linear model. In the bivariate case we can think of the relationship as

$$y_i = g(x_i, \varepsilon_i)$$

or slightly less generally as $y_i = f(x_i) + \varepsilon_i$.

- First consider Taylor series expansions of $f(x_i)$. The linear (first-order) approximation1 is

$$f(x_i) \approx \beta_1 + \beta_2 x_i + \beta_3 x_i^2$$

- In the multiple regression case, Taylor approximations also involves interaction terms. Consider, for example, a function of two regressors, $f(x_i, z_i)$.

- The second-order Taylor approximation is:

$$f(x_i, z_i) \approx \beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 x_i^2 + \beta_5 z_i^2 + \beta_6 x_i z_i$$

- The final term picks up interaction effects. Interaction effects are also relevant in situations involving dummy variables. There we capture interactions by including products of dummies

Series Expansions

- The ultimate point is that even so-called "intrinsically non-linear" models are themselves linear when viewed from the series-expansion perspective. In principle, of course, an infinite number of series terms are required, but in practice nonlinearity is often quite gentle (e.g., quadratic) so that only a few series terms are required. From this viewpoint non-linearity is in some sense really an omitted-variables problem.

- One can also use Fourier series approximations:

$$f(x_i) \approx \beta_1 + \beta_2 \sin(x_i) + \beta_3 \cos(x_i) + \beta_4 \sin(2x_i) + \beta_5 \cos(2x_i) + \cdots$$

- and one can also mix Taylor and Fourier approximations by regressing not only on powers and cross products ("Taylor terms"), but also on various sines and cosines ("Fourier terms"). Mixing may facilitate parsimony.

A Final Word on Nonlinearity and the IC

- It is of interest to step back and ask what parts of the IC are violated in our various non-linear models.
- Models linear in transformed variables (e.g., log-log regression) actually don't violate the IC, after transformation. Neither do series expansion models, if the adopted expansion order is deemed correct, because they too are linear in transformed variables.
- The series approach to handling non-linearity is actually very general and handles intrinsically non-linear models as well, and low-ordered expansions are often adequate in practice, even if an infinite expansion is required in theory. If series terms are needed, a purely linear model would suffer from misspecification of the X matrix (a violation of the IC) due to the omitted higher-order expansion terms. Hence the failure of the IC discussed in this chapter can be viewed either as:
- 1. The linearity assumption ($E(y|X) = X'\beta$) is incorrect, or
- 2. The linearity assumption ($E(y|X) = X'\beta$) is correct, but the assumption that X is correctly specified (i.e., no omitted variables) is incorrect, due to the omitted higher-order expansion terms.

Selecting a Non-Linear Model

1. t and F Tests, and Information Criteria

- One can use the usual t and F tests for testing linear models against nonlinear alternatives in nested cases, and information criteria (AIC and SIC) for testing against non-linear alternatives in non-nested cases.
- To test linearity against a quadratic alternative in a simple regression case, for example, we can simply run $y \rightarrow c, x, x^2$ and perform a t-test for the relevance of x^2.
- And of course, use AIC and SIC as always.

Selecting a Non-Linear Model

2. The RESET Test

- Direct inclusion of powers and cross products of the various x variables in the regression can be wasteful of degrees of freedom, however, particularly if there are more than just one or two right-hand-side variables in the regression and/or if the non-linearity is severe, so that fairly high powers and interactions would be necessary to capture it.
- In light of this, a useful strategy is first to fit a linear regression $y_i \rightarrow c, x_i$ and obtain the fitted values $\hat{y}_i$. Then, to test for non-linearity, we run the regression again with various powers of $\hat{y}_i$ included,

$$y_i \rightarrow c, x_i, \hat{y}_i{}^2, \dots, \hat{y}_i{}^m$$

- Note that the powers of $\hat{y}_i$ are linear combinations of powers and cross products of the x variables – just what the doctor ordered. There is no need to include the first power of $\hat{y}_i$, because that would be redundant with the included x variables. Instead we include powers $\hat{y}_i{}^2, \hat{y}_i{}^3$ ... Typically a small
- m is adequate. Significance of the included set of powers of $\hat{y}_i$ can be checked using an F test. This procedure is called RESET (Regression Specification Error Test).

Non-Linearity in Wage Determination

- For convenience we reproduce in Figure 6.1 the results of our current linear wage regression,

$$LWAGE \rightarrow c, EDUC, EXPER, FEMALE, UNION, NONWHITE$$

- The RESET test from that regression suggests neglected non-linearity; the p-value is .03 when using $\hat{y}_i^{\,2}$ *and* $\hat{y}_i^{\,3}$ in the RESET test regression.

Equation: UNTITLED Workfile: GRAPHS::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LWAGE
Method: Least Squares
Date:
Sample (adjusted): 1 1323
Included observations: 1323 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.000385	0.073180	13.67013	0.0000
EDUC	0.090809	0.004814	18.86314	0.0000
EXPER	0.012707	0.001119	11.35624	0.0000
FEMALE	-0.237535	0.025965	-9.148397	0.0000
NONWHITE	-0.085286	0.035786	-2.383199	0.0173
UNION	0.223392	0.035307	6.327126	0.0000

R-squared	0.307856	Mean dependent var	2.341995
Adjusted R-squared	0.305229	S.D. dependent var	0.561435
S.E. of regression	0.467973	Akaike info criterion	1.323712
Sum squared resid	288.4212	Schwarz criterion	1.347239
Log likelihood	-869.6356	Hannan-Quinn criter.	1.332532
F-statistic	117.1568	Durbin-Watson stat	1.910120
Prob(F-statistic)	0.000000		

Non-Linearity in Wage Determination

- Non-Linearity in *EDUC* and *EXPER*: *Powers and Interactions*
- Given the results of the RESET test, we proceed to allow for non-linearity.
- $LWAGE \rightarrow c, EDUC, EXPER, FEMALE, UNION, NONWHITE, EDUC^2, EXPER^2, EDUC,* EXPER$

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LWAGE
Method: Least Squares
Date:
Sample: 1 1323
Included observations: 1323

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.473236	0.240586	1.967017	0.0494
EDUC	0.109673	0.028918	3.792608	0.0002
EXPER	0.064422	0.007652	8.419060	0.0000
EDUC2	0.000501	0.000895	0.559994	0.5756
EXPER2	-0.000705	8.86E-05	-7.962263	0.0000
EDU_EXP	-0.001789	0.000429	-4.173423	0.0000
FEMALE	-0.237696	0.025506	-9.319335	0.0000
UNION	0.202955	0.034569	5.870998	0.0000
NONWHITE	-0.095028	0.034931	-2.720476	0.0066

R-squared	0.343072	Mean dependent var	2.341995
Adjusted R-squared	0.339073	S.D. dependent var	0.561435
S.E. of regression	0.456433	Akaike info criterion	1.276028
Sum squared resid	273.7465	Schwarz criterion	1.311318
Log likelihood	-835.0925	Hannan-Quinn criter.	1.289257
F-statistic	85.77745	Durbin-Watson stat	1.894409

Non-Linearity in Wage Determination

- Two of the non-linear effects are significant. The impact of experience is decreasing, and experience seems to trade off with education, insofar as the interaction is negative.
- Non-Linearity in FEMALE, UNION and NONWHITE: Interactions Just as continuous variables like EDUC and EXPER may interact (and we found that they do), so too may discrete dummy variables.
- For example, the wage effect of being female and non-white might not simply be the sum
- of the individual effects.
- We would estimate it as the sum of coefficients on the individual dummies FEMALE and NONWHITE plus the coefficient on the interaction dummy FEMALE*NONWHITE.
- In Figure below we show results for
- $LWAGE \rightarrow EDUC, EXPER, UNION, FEMALE, NONWHITE, FEMALE * UNION, FEMALE * NONWHITE, UNION * NONWHITE.$
- The dummy interactions are insignificant.

Non-Linearity in Wage Determination

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Sample: 1 1323
Included observations: 1323

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.482967	0.240926	2.004623	0.0452
EDUC	0.109522	0.029003	3.776211	0.0002
EXPER	0.064269	0.007654	8.396570	0.0000
EDUC2	0.000517	0.000900	0.573929	0.5661
EXPER2	-0.000701	8.87E-05	-7.904460	0.0000
EDU_EXP	-0.001796	0.000429	-4.185878	0.0000
FEMALE	-0.252921	0.029659	-8.527539	0.0000
UNION	0.200937	0.046575	4.314297	0.0000
NONWHITE	-0.161501	0.055077	-2.932246	0.0034
FEM_UNI	-0.012956	0.070740	-0.183153	0.8547
FEM_NON	0.110319	0.070093	1.573909	0.1157
UNI_NON	0.033202	0.089258	0.371975	0.7100

R-squared	0.344357	Mean dependent var	2.341995
Adjusted R-squared	0.338856	S.D. dependent var	0.561435
S.E. of regression	0.456507	Akaike info criterion	1.278605
Sum squared resid	273.2109	Schwarz criterion	1.325658
Log likelihood	-833.7970	Hannan-Quinn criter.	1.296244
F-statistic	62.59682	Durbin-Watson stat	1.891544

Non-Linearity in Continuous and Discrete Variables Simultaneously

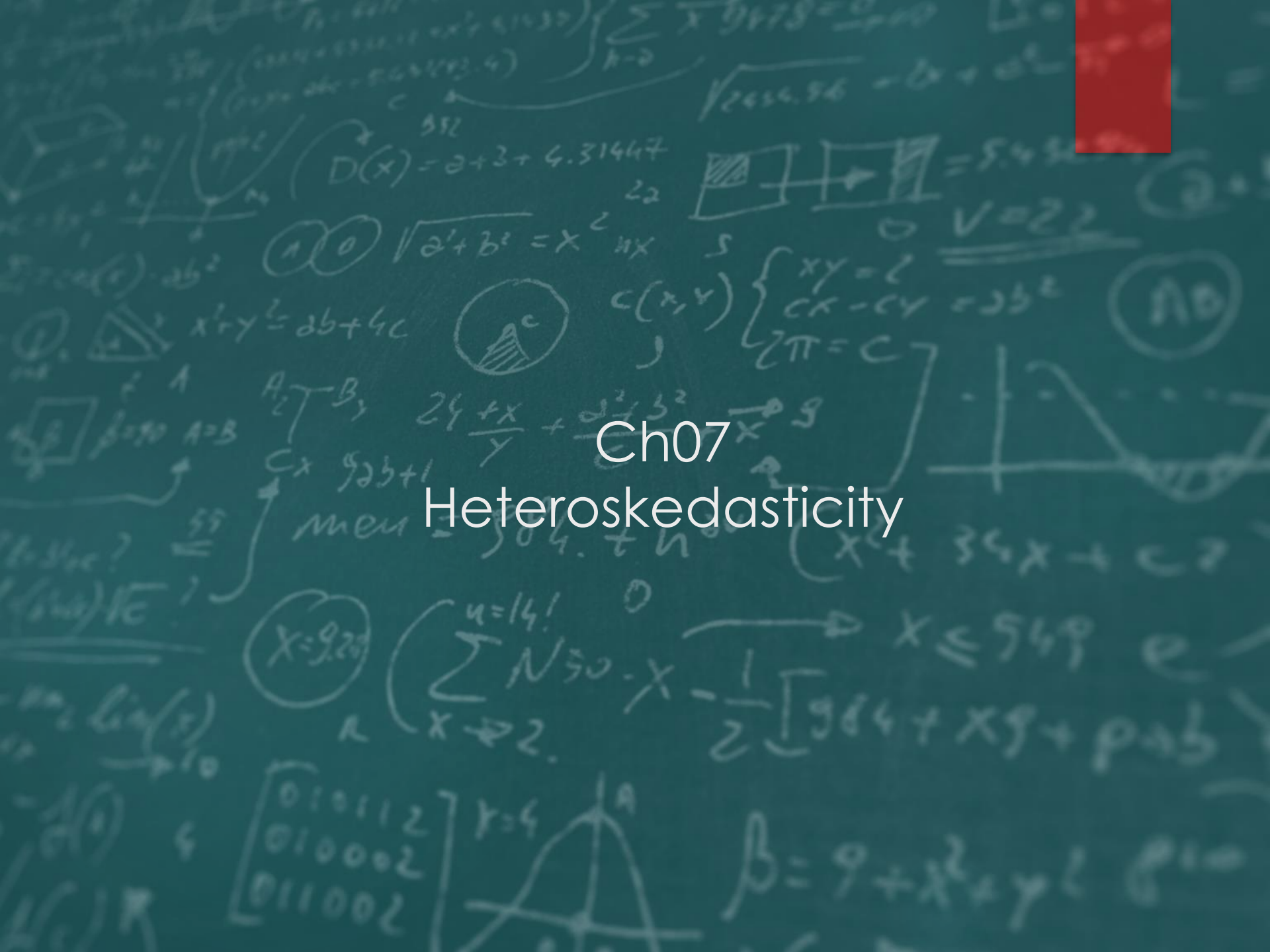
- Let's incorporate powers and interactions in *EDUC* and *EXPER*, and *interactions* in *FEMALE*, *UNION* and *NONWHITE*.
- The results for

$LWAGE$

$\rightarrow EDUC, EXPER, EDUC^2, EXPER^2, EDUC * EXPER, FEMALE, UNION, NONWHITE, FEMALE * UNION, FEMALE * NONWHITE, UNION * NONWHITE.$

- The dummy interactions remain insignificant.
- Note that we could explore additional interactions among *EDUC*, *EXPER* and the various dummies.

Exercises



Ch07
Heteroskedasticity

Heteroskedasticity

- We continue exploring issues associated with possible failure of the ideal conditions.
- This chapter's issue is "Do we really believe that disturbance variances are constant?"
- As always, consider: $\varepsilon \sim N(0, \Omega)$.
- Heteroskedasticity corresponds to Ω diagonal but $\Omega \neq \sigma^2 I$

$$\Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_N^2 \end{pmatrix}$$

- Heteroskedasticity can arise for many reasons. A leading cause is that σ_i^2 may depend on one or more of the x_i's. A classic example is an "Engel curve", a regression relating food expenditure to income. Wealthy people have much more discretion in deciding how much of their income to spend on food, so their disturbances should be more variable, as routinely found.

Consequences of Heteroskedasticity for Estimation, Inference, and Prediction

- As regards point estimation, OLS remains largely OK, insofar as parameter estimates remain consistent and asymptotically normal. They are, however, rendered inefficient. But consistency is key. Inefficiency is typically inconsequential in large samples, as long as we have consistency.
- As regards inference, however, heteroskedasticity wreaks significant havoc. Standard errors are biased and inconsistent. Hence t statistics do not have the t distribution in finite samples and do not even have the N(0, 1) distribution asymptotically.
- Finally, as regards prediction, results vary depending on whether we're talking about point or density prediction. Our earlier feasible point forecasts constructed under homoskedasticity remain useful under heteroskedasticity.
- Because parameter estimates remain consistent, we still have

$$E\left(y_i|\widehat{x_i = x_i^*}\right) \rightarrow E(y_i|x_i = x_i^*)$$

- In contrast, our earlier feasible density forecasts do not remain useful, because under heteroskedasticity it is no longer appropriate to base them on "identical σ's for different people". Now we need to base them on "different σ's for different people".

Detecting Heteroskedasticity

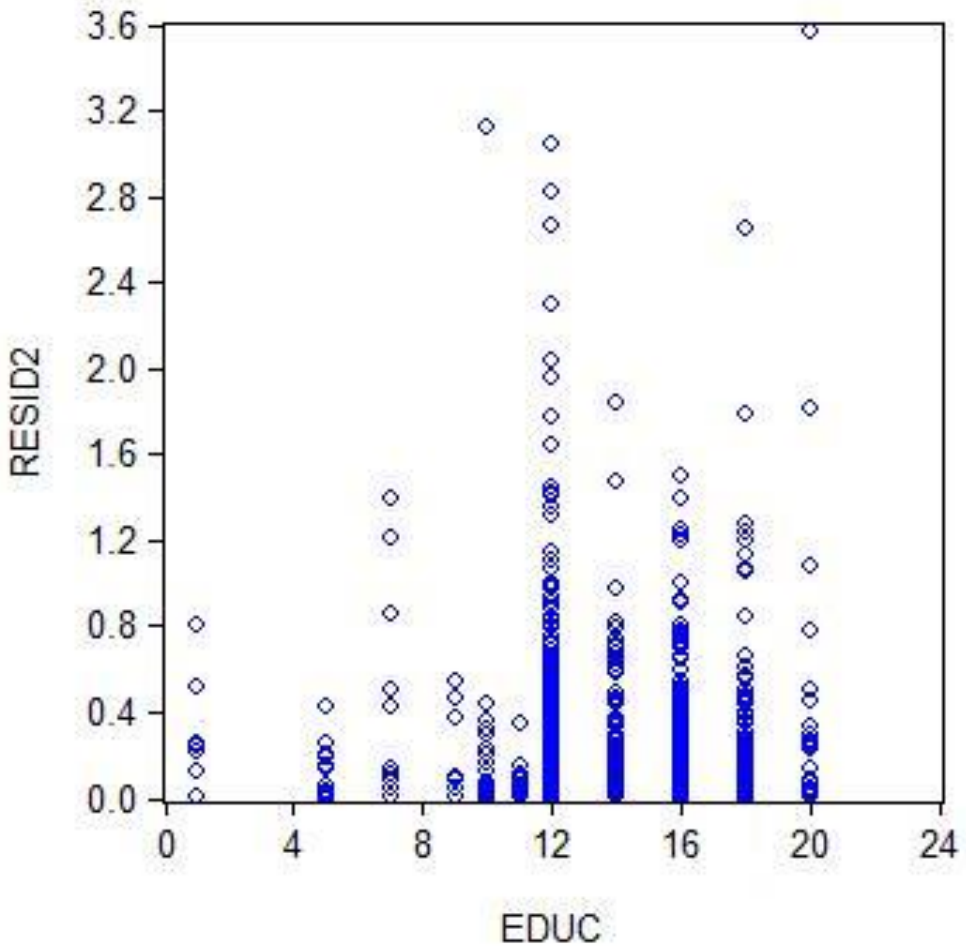
- We will consider both graphical heteroskedasticity diagnostics and formal heteroskedasticity tests. The two approaches are complements, not substitutes.

1. Graphical Diagnostics

- The first thing we can do is graph e_i^2 against x_i, for various regressors, looking for relationships. This makes sense because e_i^2 is effectively a proxy for σ_i^2.

- In Figure below we graph the squared residuals against EDUC. There is apparently a positive relationship, although it is noisy. This makes sense, because very low education almost always leads to very low wage, whereas high education can produce a larger variety of wages (e.g., both neurosurgeons and college professors are highly educated, but neurosurgeons typically earn much more).

Detecting Heteroskedasticity



Detecting Heteroskedasticity

1. Formal Tests

An important limitation of the graphical method for heteroscedasticity detection is that it is purely pairwise (we can only examine one x at a time), whereas the disturbance variance might actually depend on more than one x.

Formal tests let us blend the information from multiple x's, and they also let us assess statistical significance.

The Breusch-Pagan-Godfrey Test (BPG)

The BPG test proceeds as follows:

1. Estimate the OLS regression, and obtain the squared residuals,

2. Regress the squared residuals on all regressors

3. To test the null hypothesis of no relationship, examine $(N \cdot R^2)$ from this regression. It can be shown that in large samples $(N \cdot R^2) \sim \chi^2_{K-1}$ under the null of homoscedasticity, where K is the number of regressors in the test regression.

We show the BPG test results in Figure below.

Detecting Heteroskedasticity

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	5.414870	Prob. F(7,1315)	0.0000
Obs*R-squared	37.06628	Prob. Chi-Square(7)	0.0000
Scaled explained SS	49.66045	Prob. Chi-Square(7)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date:
Sample: 1 1323
Included observations: 1323

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.170309	0.097349	-1.749473	0.0804
EDUC	0.024074	0.006787	3.547204	0.0004
EXPER	0.011701	0.005183	2.257616	0.0241
EXPER2	-5.53E-05	6.52E-05	-0.849150	0.3960
EDU_EXP	-0.000478	0.000277	-1.725513	0.0847
FEMALE	-0.009757	0.018708	-0.521530	0.6021
UNION	-0.079648	0.025523	-3.120623	0.0018
NONWHITE	0.000486	0.025794	0.018829	0.9850

Detecting Heteroskedasticity

White's Test

White's test is a simple extension of BPG, replacing the linear BPG test regression with a more flexible (quadratic) regression:

1. Estimate the OLS regression, and obtain the squared residuals

2. Regress the squared residuals on all regressors, squared regressors, and pairwise regressors cross products

3. To test the null hypothesis of no relationship, examine $(N \cdot R^2)$ from this regression. It can be shown that in large samples $(N \cdot R^2) \sim \chi^2_{K-1}$ under the null of homoscedasticity, where K is the number of regressors in the test regression.

We show the BPG test results in Figure below.

Equation: UNTITLED Workfile: WAGESWFTEMP::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Heteroskedasticity Test: White

F-statistic	2.431488	Prob. F(29,1293)	0.0000
Obs*R-squared	68.41804	Prob. Chi-Square(29)	0.0000
Scaled explained SS	91.66473	Prob. Chi-Square(29)	0.0000

Dealing with Heteroskedasticity

- We will consider both adjusting standard errors and adjusting density forecasts
- **1. Adjusting Standard Errors**
- Using advanced methods, one can obtain consistent standard errors, even when heteroscedasticity is present. Mechanically, it's just a simple regression option.
- Even if you're only interested in prediction, you still might want to use robust standard errors, in order to do credible inference regarding the contributions of the various x variables to the point prediction.
- In Figure below we show the final wage regression with robust standard errors.
- Although the exact values of the standard errors change, it happens in this case that significance of all coefficients is preserved.

Dealing with Heteroskedasticity

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LWAGE
Method: Least Squares
Date:
Sample: 1 1323
Included observations: 1323
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.360535	0.130391	2.765026	0.0058
EDUC	0.125028	0.009890	12.64118	0.0000
EXPER	0.066130	0.006967	9.491284	0.0000
EXPER2	-0.000710	8.86E-05	-8.004870	0.0000
EDU_EXP	-0.001905	0.000412	-4.623006	0.0000
FEMALE	-0.239352	0.025499	-9.386559	0.0000
UNION	0.202574	0.031386	6.454196	0.0000
NONWHITE	-0.094903	0.034074	-2.785164	0.0054

R-squared	0.342915	Mean dependent var	2.341995
Adjusted R-squared	0.339418	S.D. dependent var	0.561435
S.E. of regression	0.456313	Akaike info criterion	1.274755
Sum squared resid	273.8119	Schwarz criterion	1.306124
Log likelihood	-835.2503	Hannan-Quinn criter.	1.286514
F-statistic	98.03775	Durbin-Watson stat	1.894273

Dealing with Heteroskedasticity

2. Adjusting Density Forecasts

Recall operational density forecast under the ideal conditions (which include, among other things, Gaussian homoscedastic disturbances):

$$y_i|x_i = x^* \sim N(x^{*\prime}\hat{\beta}_{LS}, S^2)$$

Now, under heteroscedasticity (but maintaining normality), we have the natural extension,

$$y_i|x_i = x^* \sim N(x^{*\prime}\hat{\beta}_{LS}, \hat{\sigma}_*^{\,2})$$

where $\hat{\sigma}_*^{\,2}$ is the fitted value from the BPG or White test regression evaluated at x^*.

Exercises