

Conducting Empirical Legal Scholarship: The Advanced Course Day 1

Lee Epstein¹ Andrew D. Martin²

¹Northwestern University School of Law

²Washington University School of Law

February 4-6, 2011

Friday

- Introduction to Inference
- Linear Regression
- Regression Diagnostics

Saturday

- Effectively Communicating Research Results I
- Logit, Probit, and Maximum Likelihood
- Effectively Communicating Research Results II
- Cross-Sectional Models
- Matching Methods for Causal Inference



Sunday

- Finishing Up Matching Methods for Causal Inference
- Effectively Communicating Research Results III
- Questions and Wrap-Up

What is Inference?

The process of using facts we know (or have collected) to learn about facts we don't know.

Types of Inference in Empirical Research

Descriptive Inference

Learning about the world (a population) by studying a small piece of it (a sample).

Causal Inference

Learning about the effect of an event (the key causal variable) on an outcome (the dependent variable) by determining the causal effect—the goal of causal inference.

How Do We Estimate the Causal Effect?

What Would Researchers Do If They
Had NO Constraints?

How Do We Estimate the Causal Effect?

But We Are Constrained by the
Fundamental Problem of Causal
Inference. What Now?

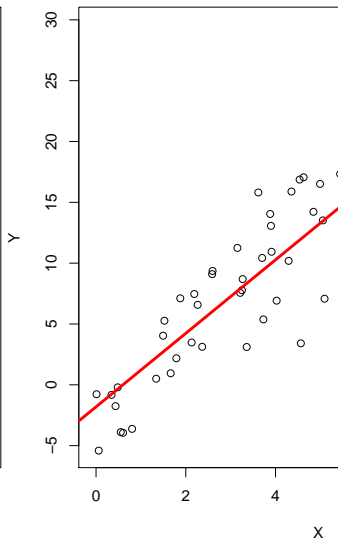
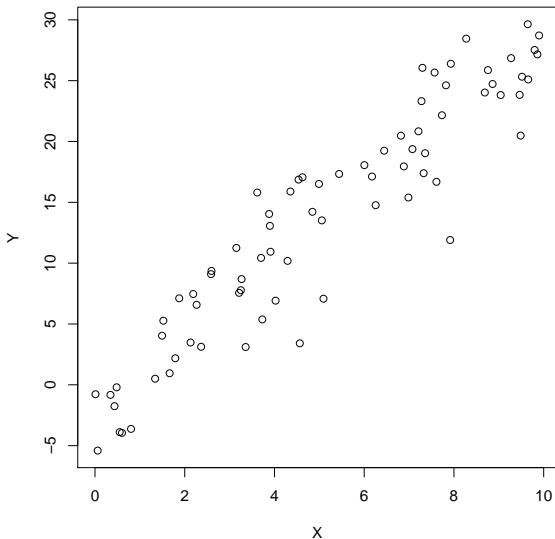
How Do We Estimate the Causal Effect?

- 1 Parametric Methods (e.g., regression analysis)
- 2 Nonparametric Methods (e.g., matching)

Variables

- The linear regression model can also be used to perform inference about the relationship between two variables.
- We are interested in learning about the **population** relationship by looking at a **sample** of data.
- Y is the dependent variable.
- X is the independent variable.
- Assume that both Y and X are continuous.

Visualization Using a Scatterplot



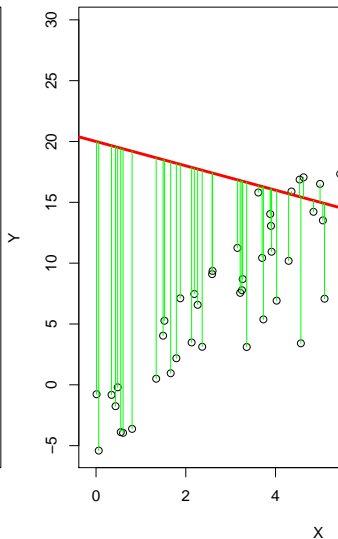
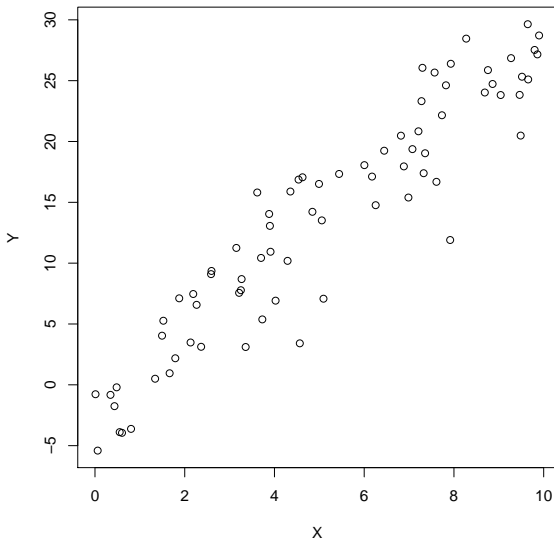
Estimating the Regression Line

- We are interested in the population relationship:

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

- The **parameter** α is the intercept. It is the expected value of Y when X equals zero.
- The **parameter** β is the slope. For a one unit increase in X , it is the expected increase in Y .
- Note that the slope captures the nature of the relationship. It could be positive, negative, or zero.

Finding the Ordinary Least Squares (OLS) Line



Estimation

- If the population regression were true, we would only observe fundamental variability.
- We will also encounter sampling variability which we will also have to take account of.
- We will assume that the distribution of the residuals is Normal (bell-shaped).
- This results in the following model:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

- The **parameter** σ^2 gauges the the variance of the points around the regression line.
- Any textbook provides the **statistics** that can be used to estimate these parameters: $a, b, \hat{\sigma}$.

Inference

- What would be the key hypothesis to test for the linear regression model?

$$H_0 : \beta = 0$$

- If we could reject the null, we could conclude that there is a statistically significant relationship between the dependent variable and the independent variable.
- Statistical software reports these p-values, which are computed from the Student-t distribution.

Assessing Model Fit

- We also might want to know, in general, how well a model fits.
- One quantity is the standard error of the estimate (SEE): $\hat{\sigma}$
- Another quantity is called the R^2 .
- R^2 is the percentage of explained variance.
- It also is the correlation coefficient to the second power.
- When a model is perfect, R^2 goes to one, and SEE goes to zero.

Linear Regression in Stata

- Nominations Data (`nom.dta`)
- Y = percent liberalism in civil liberties cases
- X = ideology as measured at time of confirmation
- Create a scatterplot
- Estimate a linear regression

Statistical Control

- We would like to expand the simple linear regression model to allow for statistical control.
- Now we have two (or more) independent variables: X_1 and X_2 .
- To find the effect of X_2 while controlling for X_1 , and vice versa, consider the following:

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i$$

Interpretation

- The intercept α coefficient is now the expected value of Y when both X_1 and X_2 are zero.
- There are now two partial slope coefficients— β_1 and β_2 —which capture the effect of a one unit change in X *while holding the other constant*.
- The measures of model fit are the same.
- Stata also provides us with an adjusted- R^2 and an F-test.
- The model can be extended to include as many explanatory variables as you want.
- This model is the workhorse of empirical social science (along with its cousins for other types of dependent variables).

Independent Variables

- The model can be easily adapted to deal with other types of independent variables.
- X can be a dichotomous (yes/no) variable.
- One can encode a nominal or ordinal variable using a series of dichotomous variables.
- One can deal with interactions by multiplying X_1 and X_2 and estimating a partial slope for the interaction (details tomorrow).
- The model can be generalized to deal with non-linearities.

Multiple Regression in Stata

- Nominations Data (`nom.dta`)
- $Y = \text{civlib}$
- $X_1 = \text{ideol}$
- $X_2 = \text{pparty}$
- Estimate a multiple regression

General Principles

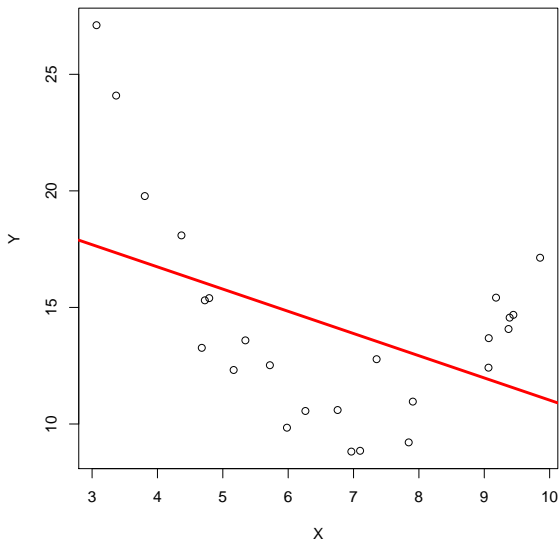
- “All models are wrong; some models are useful.” – George Box
- To some extent, every inference depends on the model structure.
- Some inferences are very model dependent; others quite robust.
- In the context of the linear regression model, there are a number of important assumptions.
- Some of these assumptions are testable; others are not.
- I’ll discuss each of the assumptions, talk about diagnostic tools, and possible alternatives if the assumptions are violated.

Assumption

- **Linearity [1].**
- The expected value of Y is a linear function of the independent variables.
- This figure shows the ordinary least squares fit when the assumption does not hold.
- The parameter estimates (and their standard errors) are not meaningful.

Linearity

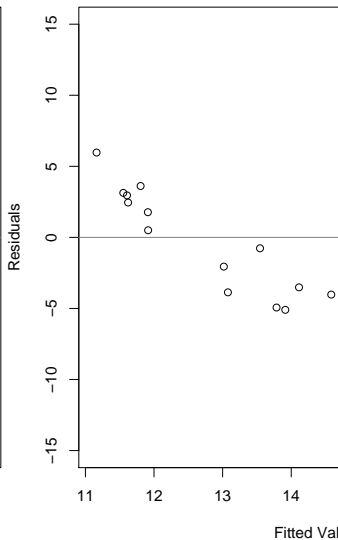
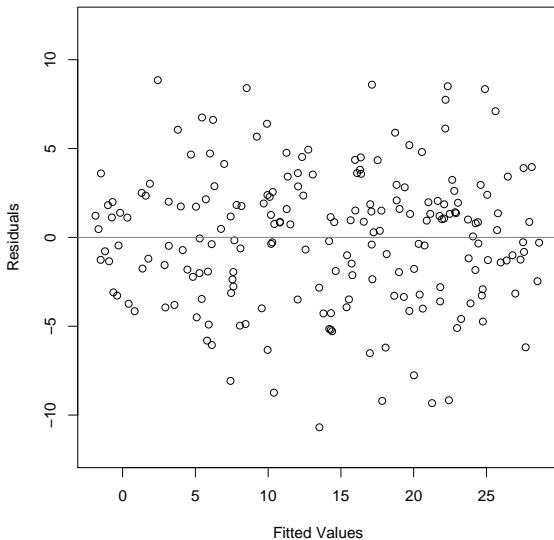
Linearity



Diagnostics

- For a simple linear regression, it is easy to eye-ball violations of non-linearity.
- What to do for a multiple regression model?
- Look at the residuals.
- One typically will plot the residuals against the fitted values (\hat{Y}_i) to detect patterns.
- Here are some examples of “well-behaved” and “ill-behaved” residuals.

Diagnostics



Solutions

- There is one solution...
- Respecify the model.
- One could transform X , Y , or both to yield a model that is linear in the parameters.
- A common example is to use a logarithmic transformation when dealing with a highly skewed variable; e.g., log-punitive damages.

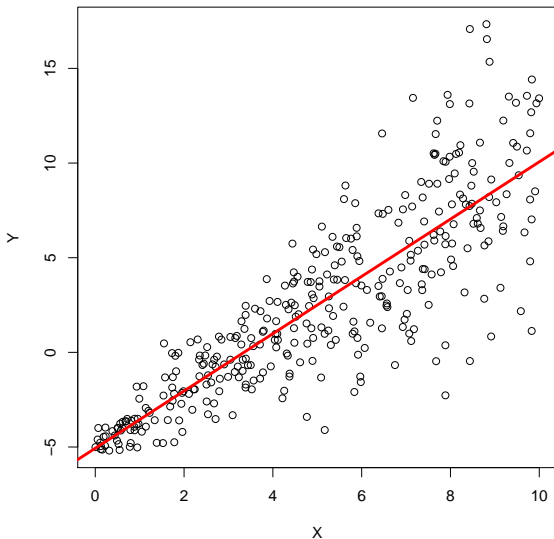
Examining Residuals in Stata

- Nominations Data (`nom.dta`)
- Y = percent liberalism in civil liberties cases
- X = ideology as measured at time of confirmation
- Estimate a regression model
- Examine the residuals

Constant Variance

- **Constant Variance [2].**; i.e., homoscedasticity.
- The variance of the errors is constant regardless of the value of the X s.
- This means it is reasonable to estimate a single parameter σ^2 rather than a separate σ_i^2 for each observation.
- A violation of this assumption is called **heteroscedasticity**.
- What happens if the assumption is violated?
- The estimates of α and β are correct, but the standard errors will be wrong.

Illustration of Heteroscedastic Errors



What Can Cause Heteroscedasticity?

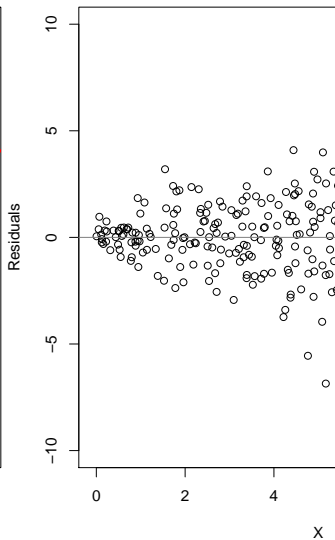
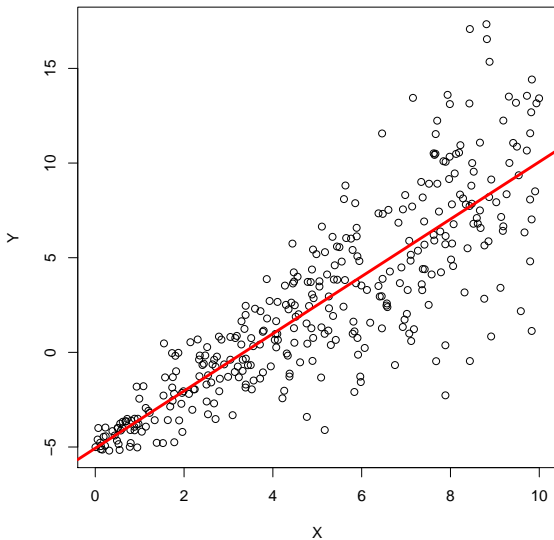
- Heterogeneous units of analysis.
- Data collected over time with learning.
- Improved data collection protocol.
- Improper model specification.

Diagnosing Heteroscedastic Errors

- One can just look at the scatterplot in a simple linear regression.
- For a multiple regression, we'll again look at the residuals.
- Here we will plot the residuals on each X_j and look for a pattern.
- There are also various hypothesis tests.
- Stata implements the Breusch-Pagan / Cook-Weisberg test (`hettest`). H_0 : constant error variance.
- With a small p-value, we'd reject the null hypothesis of homoscedasticity.

Homoscedastic Errors

Diagnosing Errors



Heteroscedasticity in Stata

- Nominations Data (`nom.dta`)
- Y = percent liberalism in civil liberties cases
- X = ideology as measured at time of confirmation
- Plot the residuals against X
- Perform a `hettest`

Solutions

- What should one do if we find heteroscedasticity?
- Old solutions: Weighted Least Squares (WLS) and Generalized Least Squares (GLS).
- The gold standard: Estimate a heteroscedastic regression (Stata's optional `regh` command).
- The next best solution: Huber-White ("robust") standard errors.
- These standard errors are shown to be correct even in the face of heteroscedasticity (with a sufficiently large sample).
- Easy to implement in Stata: `, robust`.
- One loses little by using these as a matter of course when fitting linear regression models.
- There are some issues when used for other models.

Normal Errors

- **Normal Errors [3].**
- The errors follow a Normal distribution.
- Alternatively, Y conditional on X follows a Normal distribution.
- It is conceivable that errors could follow any other distribution.
- The normality assumption is typically justified by appealing to the Central Limit Theorem.
- One could test this assumption by looking at a histogram of the errors, or performing an `sktest` for normality.
- Assessing normality in Stata.
- Solutions.

Independent Errors

- Independent Errors [4].
- For cross-sectional data, this will almost always be met if the data are a simple random sample from the population.
- For time-series data, this will likely not be met. The problem is called **autocorrelation** or **serial correlation**.
- When dealing with autocorrelated data, the α and β are correct, but the standard errors will be wrong.
- Common diagnostic tools: time series plot of residuals and Durbin-Watson test.
- When dealing with time series data, consult an expert.

Exogenous Regressors

- Exogenous X [5].
- In experimental studies, X is under the control of the experimenter, and is thus exogenous.
- Not so for observational studies.
- Example – explaining voteshare in congressional elections using campaign spending.
- Technically this assumption is that the residuals and X are uncorrelated.
- If the assumption does not hold, the estimates of α and β are inconsistent; i.e., wrong, even in large samples.

Thinking About Exogeneity

- There are some exogeneity tests, but for the most part they are not powerful.
- A better approach is to think about this as a research design problem.
- Always ask: Are my regressors exogenous?
- If not, then you need to turn to an alternative strategy.
- The most common strategy is called instrumental variables (two-stage least squares).

Instrumental Variables

- Instrumental variables works by finding exogenous variables Z that are correlated with X and uncorrelated with the errors.
- With an endogenous regressor $Cov(X_{1,i}, \varepsilon_i) \neq 0$
- Look for a “magic” instrument Z_i that can be used in place of $X_{1,i}$ that has these properties:
 - $Cov(Z_i, \varepsilon_i) = 0$
 - $Cov(Z_i, X_{1,i}) \neq 0$
 - Z_i only affects Y_i through $X_{1,i}$
- Two-stage least squares works by (essentially) “substituting” Z_i for $X_{1,i}$
- If the assumptions are correct, you can recover unbiased estimates of the slope and intercept parameters

The Final (and Toughest) Assumption

- **Model Correctly Specified [6].**
- This is an untestable assumption.
- What happens if we leave a theoretically important variable out of a model?
- The estimates of α and β are biased; this is called omitted variable bias.
- Finding other ills, such as heteroscedasticity, might suggest omitted variable bias.
- One implication – never “prune” models.
- What if I include irrelevant variables?
- Nothing except a loss of efficiency, as long as all variables are pre-treatment.

Multicollinearity

- Suppose that we include two variables in a multiple regression X_1 and X_2 that are perfectly correlated.
- For example, imagine ideology measured liberal/conservative and party measured Democrat/Republican.
- If they are perfectly correlated, then the regression model cannot be estimated.
- This is problem of **perfect multicollinearity**.
- Stata will jettison one of the offending variables.

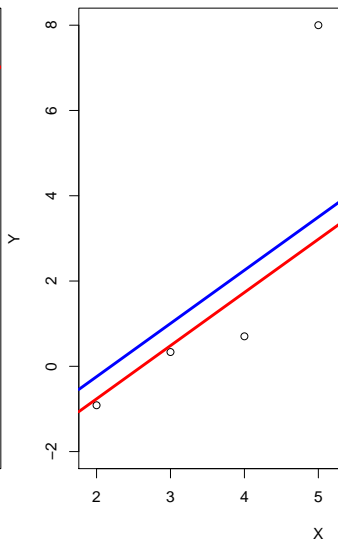
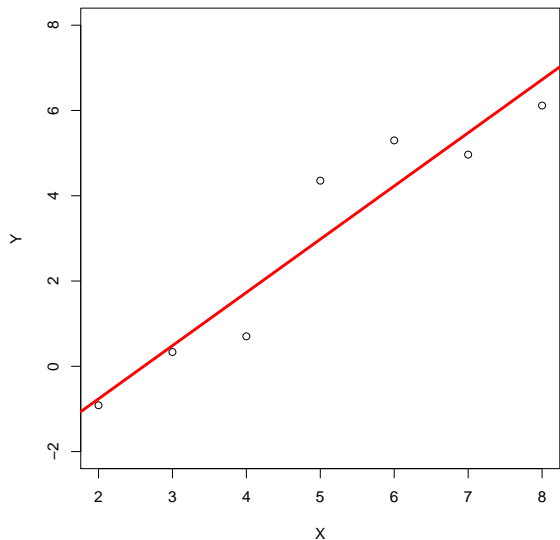
Multicollinearity

- Suppose that X_1 and X_2 are highly, but not perfectly, correlated.
- This is the so-called problem of **multicollinearity**.
- This is not, however, a statistical problem. All reported model estimates are correct!
- However, in this situation, we often encounter weird things, like high R^2 with all variables insignificant.
- There are a number of diagnostic tools (variance inflation factors, condition number of the design matrix, etc.).
- This should be thought about as a research design problem.

Robustness

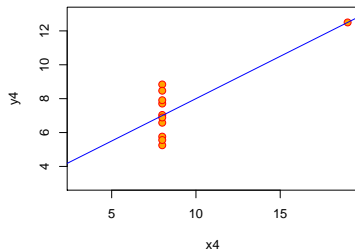
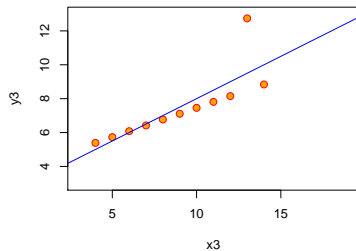
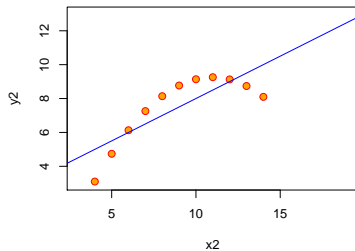
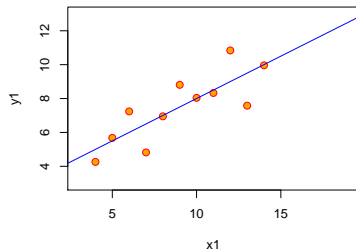
- Recall that the same mean \bar{Y} is not robust; changing one value can radically change our estimate of the population mean.
- Linear regression estimates are, in general, also not robust.
- When using a linear regression model, it is important to determine whether one's estimates are being driven by weird observations.
- An **outlier** is a data point that falls “far” from the regression line.
- A **leveraged outlier** is an outlier that causes the regression line to be quite different from what it otherwise would be.
- What causes these? Most common is coding mistakes, etc. But also could be just idiosyncratic observations.

Leverage Illustration



Outliers and Leverage

Anscombe's Quartet



Finding Leveraged Outliers

- There are a number of different statistics.
- The most commonly used is Cook's D statistics.
- These are computed by re-estimating the regression dropping each observation at a time.
- These are "large" if $D > 4/(N - K - 1)$.
- Computing Cook's D statistics in Stata.

Dealing with Leveraged Outliers

- Drop the troublesome observations?
- Search for data errors.
- Respecify the model with transformations; e.g., a logarithmic transformation of income.
- Robust regression (`rreg`). *Not* robust standard errors.
- There is a whole literature on robust statistics.

The Monte Carlo Principle

- Suppose we want to learn about some process; e.g., craps.
- We want to answer the question: What can I expect to win on a \$10 pass line bet? (The answer is \$9.72).
- We could approach this problem by doing some fancy probability calculations.
- We could also use a computer to simulate a large number of pass line bets.
- The long-run average of this process would be a simulation-based estimate of the true probability.

The Monte Carlo Principle

- The **Monte Carlo Principle** states that anything we want to learn about a random variable we can learn by sampling from its density.
- These methods were first introduced by Stan Ulam and Nick Metropolis.
- All you need is a fast computer and some algorithm to do the simulation.
- Why would we want to do this?
- For many problems, simulation is far easier than doing fancy math. We'll use simulation to help us with interpretation.

An Illustration of the Monte Carlo Principle

[Click Here for Mixture of Normals Quicktime Movie](#)

[IN CLASS EXERCISE]