Fundamentals of Econometrics

MEPS - Preparatory and Orientation Week WS 2012/2013

Kristin Bernhardt

8. - 12. October 2012

Overview

1. Introduction to Statistics and Econometrics
2. Review of Statistical Theory
3. Linear Regression Model

Course Outline

Econometric Literature

- Stock, Watson 'Introduction to Econometrics', 3rd edition (2011)
- Wooldridge, 'Introductory Econmetrics - A modern Approach', 4th edition (2008)
- Greene, 'Econometric Analysis', 7th edition (2010)

Introduction to statistics and econometrics

Economics (theory) suggests important relationships, often with policy implications, but virtually never suggests quantitative magnitudes of causal effects.

- What is the *quantitative* effect of increasing interest rates on consumption?
- How does another year of education change earnings?
- What is the price elasticity of cigarettes?
- etc.

Data types

Cross sectional data:

- We observe different objects (e.g. individuals, companies, countries) at one point in time.

Time series data:

- We observe one object at different points in time.

Panel data:

- We observe different objects at different points in time.

The probability framework for statistical inference

Population

- The group or collection of all possible entities of interest (school districts).
- We will think of populations as infinitely large (inf is an approximation to 'very big').

Random variable Y

- Numerical values of an random outcome (district average test score, district STR).

Probability distribution of Y

- The probabilities of different values of Y that occur in the population, for ex. $\Pr[Y=650]$, when Y is discrete.
- or: The probabilities of sets of these values, for ex. $\Pr[640 \leq Y \leq 660]$, when Y is continuous.

The probability framework for statistical inference

Mean

- Expected value (expectation) of Y.
- $E(Y) = \mu_y$
- Long-run average value of Y over repeated realizations of Y.

Variance

- Averaged squared deviation of the random variable from the expected value.
- $E[(Y - E(Y))^2] = \sigma_Y^2$
- Measure of the squared spread of the distribution.

Standard deviation

- Square root of variance.
- $\sqrt{\sigma_Y^2} = \sigma_Y$
- Same unit as Y.

The probability framework for statistical inference

Skewness

- Measure of asymmetry of a distribution.
- Skewness $= 0$; distribution is symmetric.
- Skewness $> (<) \, 0$; distribution has long right (left) tail.

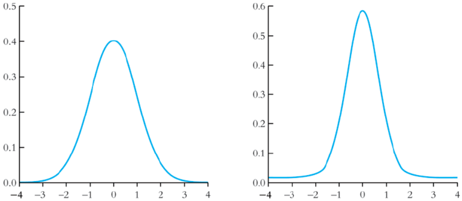
Kurtosis

- Measure of mass in tails.
- Measure of probability of large values.
- Kurtosis $= 3$; normal distribution.
- Kurtosis > 3; heavy tails ('leptokurtotic').

Skewness Kurtosis

$$\frac{E[(y-\mu_y)^3]}{\sigma_y^3} \quad \frac{E[(y-\mu_y)^4]}{\sigma_y^4}$$

$\mu_Y = 0,$
$\sigma_Y = 1$



(a) Skewness = 0, kurtosis = 3 (b) Skewness = 0, kurtosis = 20

The probability framework for statistical inference

2 random variables: joint distributions and covariance

- Random variables X and Z have *joint distribution*, Pr(X=x, Z=z).
- The *covariance* between X and Z is

$$cov_{XZ} = \sigma_{XZ} = E[(X - \mu_X)(Z - \mu_Z)]$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) \quad (1)$$

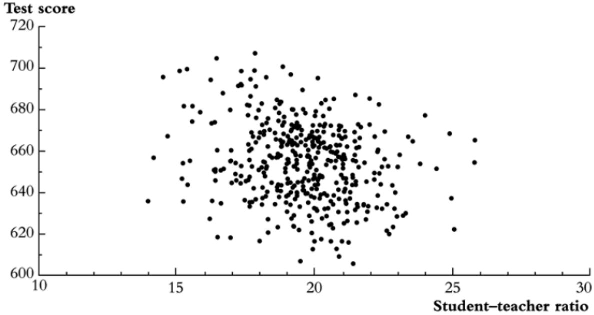
  The covariance is a measure of the linear association between X and Z; its units are the units of X times the units of Z.

- $cov(X, Z) > 0$ means a positive relation between X and Z.
- If X and Z are independently distributed, then $cov(X, Z) = 0$.
- The covariance of a random variable v with itself is its variance σ_X^2.

The covariance between Test Score and STR is negative, so the *correlation* is....

FIGURE 4.2 Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student–teacher ratio and test scores: The sample correlation is −0.23.



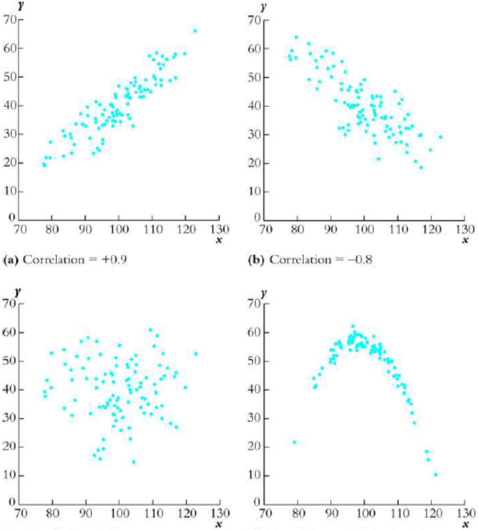
The probability framework for statistical inference

The correlation coefficient is defined in terms of the covariance

$$corr(X, Z) = \frac{cov(X,Z)}{\sqrt{var(X)var(Z)}} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z}$$

- $-1 \leq corr(X, Z) \leq 1$
- $corr(X, Z) = 1$ means perfect positive linear association
- $corr(X, Z) = -1$ means perfect negative linear association
- $corr(X, Z) = 0$ means no linear association

The correlation coefficient measures only linear association.



(a) Correlation = +0.9

(b) Correlation = −0.8

(c) Correlation = 0.0

(d) Correlation = 0.0 (quadratic)

Conditional distributions

- The distributions of Y, given value(s) of some other random variable.
- Example: the distribution of test scores, given that $STR < 20$

Conditional expectations and conditional moments

- Conditional mean = mean of conditional distribution

  $=E(Y|X = x)$ **(important concept and notation)**
- Conditional variance = variance of conditional distribution.
- Example: $E(Testscores|STR < 20)$= the mean of test scores among districts with small class sizes.

Estimation

Law of Large Numbers

- For large samples, the sample mean is with large probability close to μ_y.
  - Large sample, i.e. $n \leftarrow \infty$
  - With large probability, i.e. $p \leftarrow 1$ for $n \leftarrow \infty$
  - Close to μ_Y, i.e. for large enough n, the deviation of sample mean from μ_Y is small.

Estimation

Central Limit Theorem

If $(Y_V, ..., Y_n)$ are i.i.d. and $0 < \sigma_Y^2 < \infty$, then when n is large the distribution of $\bar{Y}$ is well approximated by a normal distribution.

- $\bar{Y}$ is approximately distributed $N(\mu_Y, \frac{\sigma_Y^2}{n})$, 'normal distribution with mean μ_Y and variance $\frac{\sigma_Y^2}{n}$.
- $\frac{\sqrt{n}(\bar{Y}-\mu_Y)}{\sigma_Y}$ is Y approximately distributed $N(0,1)$, (standard normal).
- That is 'standardized' $\bar{Y} = \frac{\bar{Y}-E(\bar{Y})}{\sqrt{var(\bar{Y})}} = \frac{\bar{Y}-\mu_Y}{\frac{\sigma_Y}{\sqrt{n}}}$ is approximately distributed as $N(0,1)$.

Basic Idea

We want to find the regression line that fits our scatter plot best.

Basic Idea

- The slope of the regression line is the expected effect on Y of a unit change in X.
- In our example, class size and test score.
- With the regression model, we determine
  - whether there is a (statistically significant) relation between X and Y,
  - how strong a relation might be,
  - whether there is a causal effect between X and Y?

Basic Idea

Estimation:

- How should we draw a line through the data to estimate the slope?
    - Ordinary Least Squares (OLS).

Hypothesis testing:

- How to test if the slope is zero, i.e. there is no effect of X on Y?

Confidence intervals:

- How to construct a confidence interval for the slope?

Basic Idea

- The regression line: $Testscore = b_0 + b_1 STR$
  - b_1 = slope of the regression line
  - $= \frac{\partial Testscore}{\partial STR}$
  - = change in test score for a unit change in STR.
- We would like to know the value of b_1.
- Since we do not know b_1, we will estimate it, using data.

The simple linear regression model

$$Y_i = b_0 + b_1 X_i + u_i \text{ with } i = 1, \ldots, n$$

- We have n observations, $(X_i, Y_i), i = 1, \ldots, n$.
- X is the *independent variable* or *regressor*, also *explanatory variable*.
- Y is the *dependent variable* or *regressand*, also *explained variable*.
- $b_0 =$ *intercept*
- $b_1 =$ *slope*
- $u_i =$ the regression error

The regression error consists of omitted factors. In general, these omitted factors are other factors that influence Y, other than the variable X. The regression error also includes error in the measurement of Y.

The Ordinary Least Squares Estimation

How can we estimate b_0 and b_1 from the data?

- We will focus on the least squares ('ordinary least squares' or 'OLS') estimator of the unknown parameters β_0 and β_1.
- The OLS estimator solves,

$$min_{b_0,b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2 \tag{2}$$

- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction ('predicted value') based on the estimated line.
- The result are the OLS estimators of b_0 and b_1.

The Ordinary Least Squares Estimation

The OLS Estimator, Predicted Values and Residuals

The OLS estimators fo the slope β_1 and the intercept β_0 are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2} \tag{3}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{4}$$

The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, i = 1, ..., n \tag{5}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, i = 1, ..., n. \tag{6}$$

The OLS Estimator, Predicted Values and Residuals

- The estimated intercept $(\hat{\beta}_0)$, slope $(\hat{\beta}_1)$, and residual $(\hat{u}_i)$ are computed from a sample of n observations of X_i and Y_i, $i = 1, ..., n$.
- These are estimates of the unknown true population intercept (β_0), slope (β_1), and the error term (u_i).

Measures of Fit

Two regression statistics provide complementary measures of how well the regression line 'fits' or explains the data:

1. **Regression** R^2
   measures the fraction of the variance of Y that is explained by X; it is unitless and ranges between zero (no fit) and one (perfect fit).

2. **Standard error of the regression (SER)**
   measures the magnitude of a typical regression residual in the units of Y.

Measures of Fit

$$Y_i = \hat{Y}_i + \hat{u}_i$$

Variance decomposition:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{u_i^2} \quad (7)$$

$$\text{TSS} = \text{ESS} + \text{SSR}$$

- TSS= total sum of squares
- ESS = explained sum of squares
- SSR= sum of square residuals

(Here we use $\bar{\hat{Y}} = \bar{Y}$ and $\bar{\hat{u}} = 0$).

Regression R^2

Regression R^2 = fraction of variation of Y is explained by X.

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{SSR}{TSS} \tag{8}$$

- $0 \leq R^2 \leq 1$
- $R^2 = 1$ means ESS=TSS and SSR=0, i.e. all data points are on the regression line.
- $R^2 = 0$ measn ESS=0 ($\beta_1 = 0$), i.e. no variation is explained.
- For regression with a single X, R^2 = the square of the correlation coefficient between X and Y.

$$R^2 = [corr(X, Y)]^2$$

Standard Error of Regression SER

The SER measures the spread of the distribution of u.

$$SER = \sqrt{s_{\hat{u}}^2} = \sqrt{\frac{1}{n-2}\sum_{i=2}^{n}(\hat{u}_i - \bar{\hat{u}})^2} = \sqrt{\frac{1}{n-2}\sum_{i=2}^{n}(\hat{u_i^2}} \qquad (9)$$

- The SER is (almost) the sample standard deviation of the OLS residuals.
- The SER has the units of u, which are the units of Y.
- It measures the average 'size' of the OLS residual (the average 'mistake' made by the OLS regression line).
- Why $n-2$? Degrees of freedom correction by number of estimated estimators. (In large samples it is irrelevant, whether division by n, $n-1$ or $n-2$).