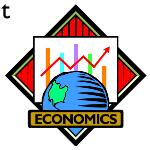
EconS 450

Forecasting – part 3

Forecasting with Regression

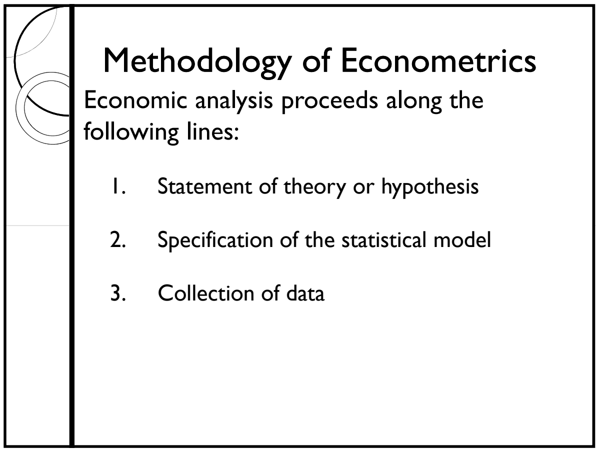
Using regression to study economic relationships is called *econometrics*

econo	=	of or pertaining to the economy
metrics	=	measurement



Econometrics

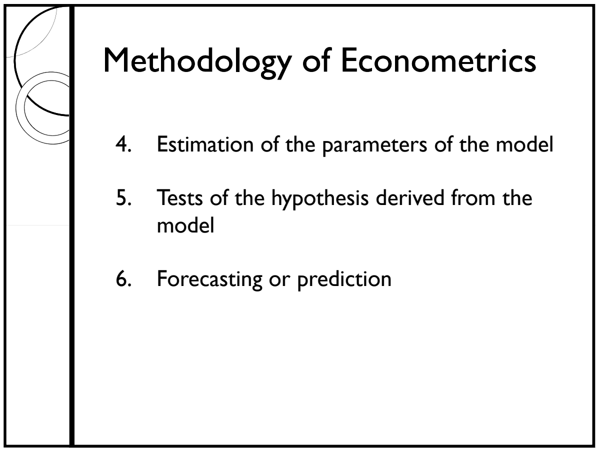
What we attempt to do by calculating regression parameters is find a model that best mimics the true underlying data generating process.



Methodology of Econometrics

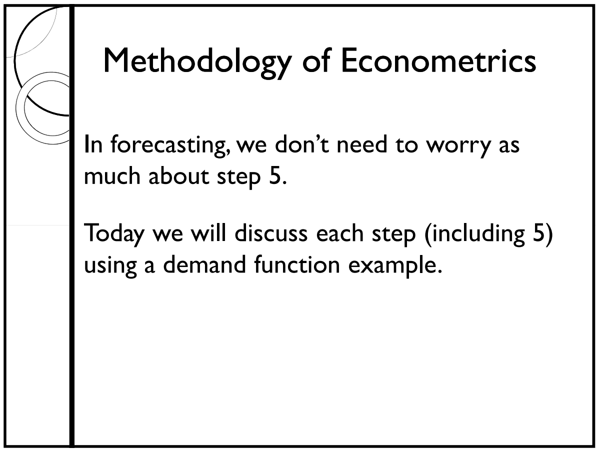
Economic analysis proceeds along the following lines:

1. Statement of theory or hypothesis
2. Specification of the statistical model
3. Collection of data



Methodology of Econometrics

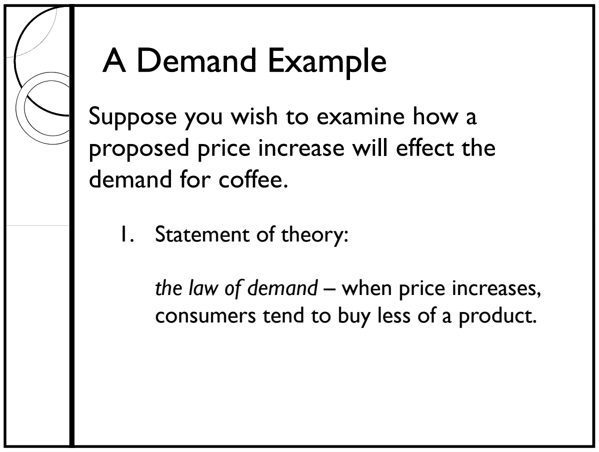
4. Estimation of the parameters of the model
5. Tests of the hypothesis derived from the model
6. Forecasting or prediction



Methodology of Econometrics

In forecasting, we don't need to worry as much about step 5.

Today we will discuss each step (including 5) using a demand function example.

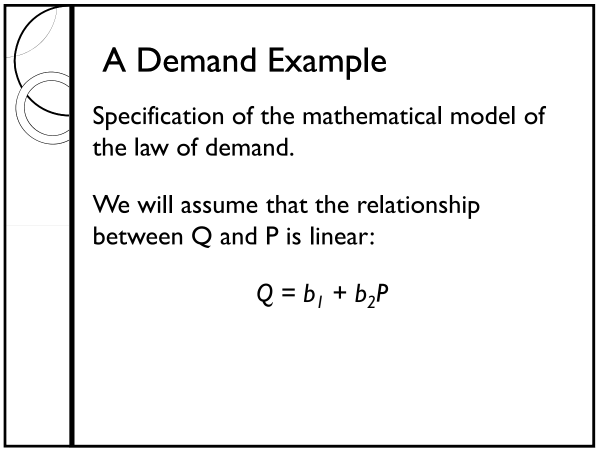


A Demand Example

Suppose you wish to examine how a proposed price increase will effect the demand for coffee.

1. Statement of theory:

   the law of demand – when price increases, consumers tend to buy less of a product.

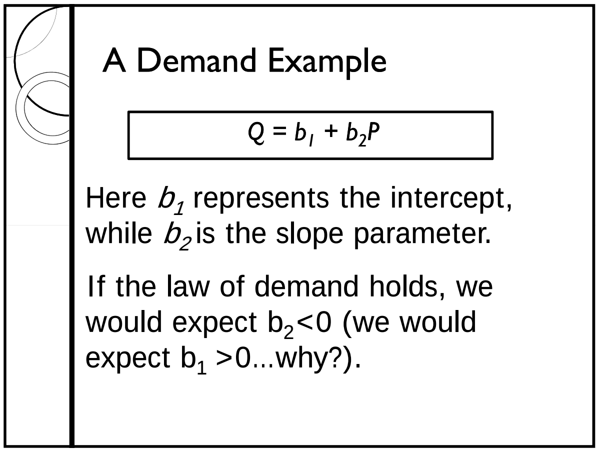


A Demand Example

Specification of the mathematical model of the law of demand.

We will assume that the relationship between Q and P is linear:

$$Q = b_1 + b_2P$$

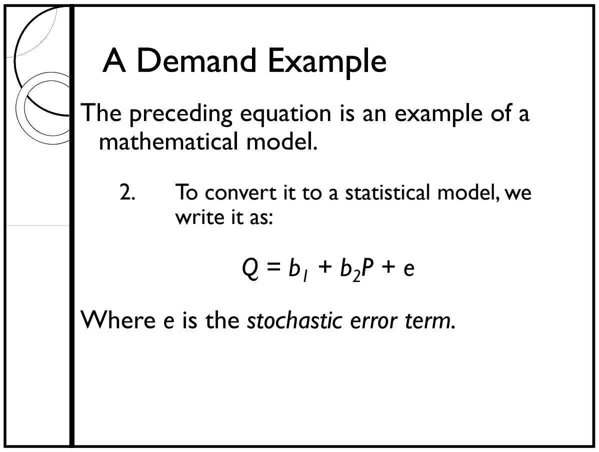


A Demand Example

$$\boxed{Q = b_1 + b_2P}$$

Here b_1 represents the intercept, while b_2 is the slope parameter.

If the law of demand holds, we would expect $b_2<0$ (we would expect $b_1 >0$…why?).



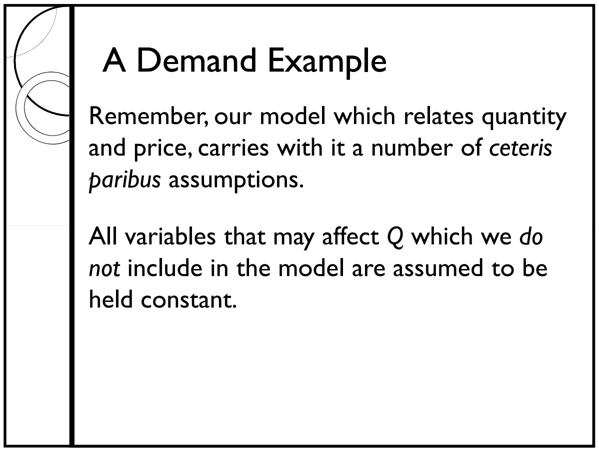
A Demand Example

The preceding equation is an example of a mathematical model.

2. To convert it to a statistical model, we write it as:

$$Q = b_1 + b_2P + e$$

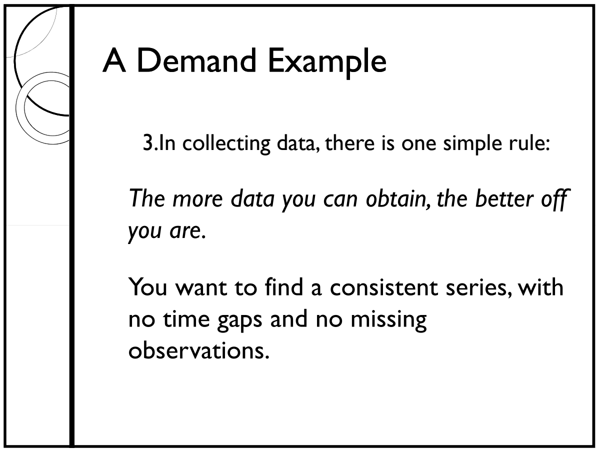
Where e is the *stochastic error term.*



A Demand Example

Remember, our model which relates quantity and price, carries with it a number of *ceteris paribus* assumptions.

All variables that may affect *Q* which we *do not* include in the model are assumed to be held constant.

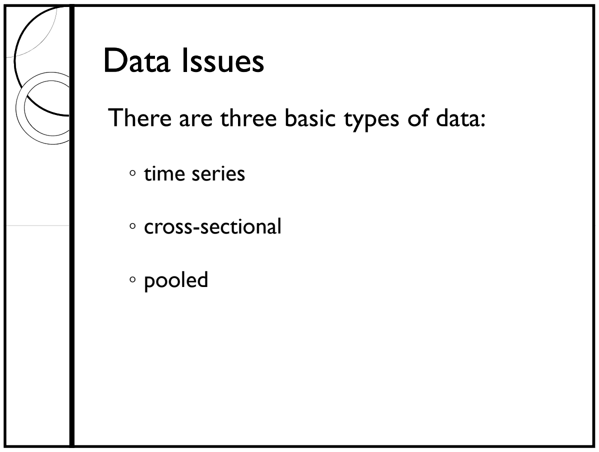


A Demand Example

3.In collecting data, there is one simple rule:

The more data you can obtain, the better off you are.

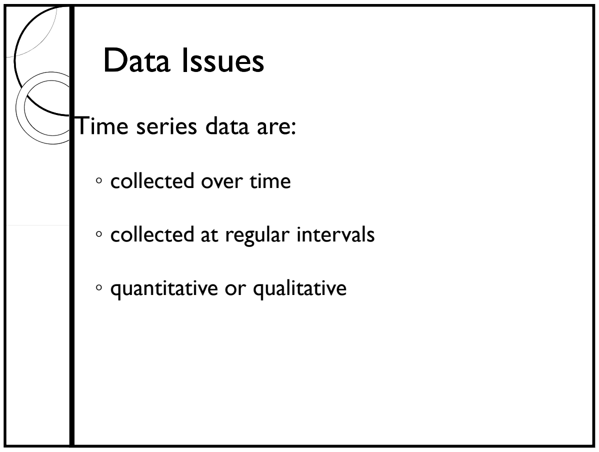
You want to find a consistent series, with no time gaps and no missing observations.



Data Issues

There are three basic types of data:

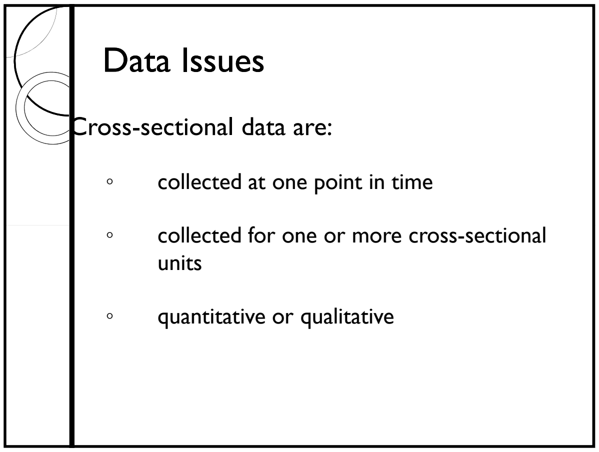
- time series
- cross-sectional
- pooled



Data Issues

Time series data are:

- collected over time
- collected at regular intervals
- quantitative or qualitative



Data Issues

Cross-sectional data are:

- collected at one point in time
- collected for one or more cross-sectional units
- quantitative or qualitative

Data Issues

Pooled data are:

- time series collected for two or more cross sectional units
- quantitative or qualitative

Data Issues

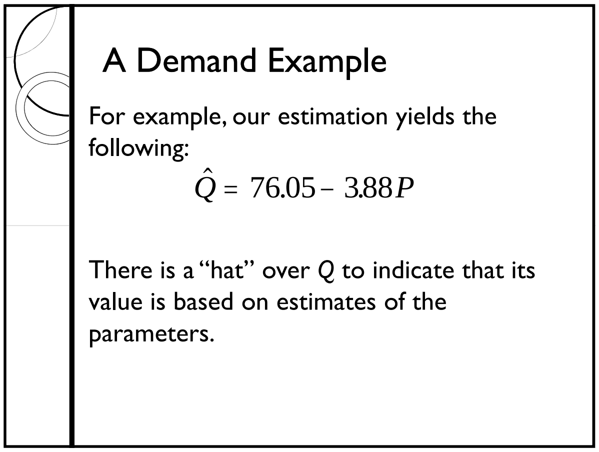
When forecasting, we are mostly concerned with *time series data.*

When estimating supply or demand, you may wish to use either time series or pooled data.

A Demand Example

4.Once we have the model and data, we can obtain *parameter estimates* as discussed in the last lecture.

The estimates are values which *minimize the sum of squared errors* (where the error refers to the value of the error term for each period).

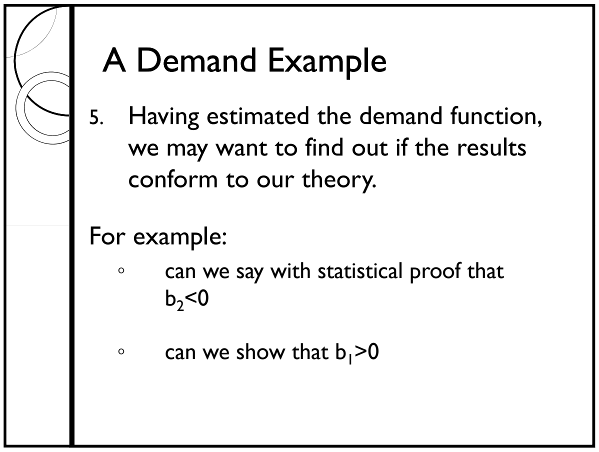


A Demand Example

For example, our estimation yields the following:

$$\hat{Q} = 76.05 - 3.88P$$

There is a "hat" over *Q* to indicate that its value is based on estimates of the parameters.

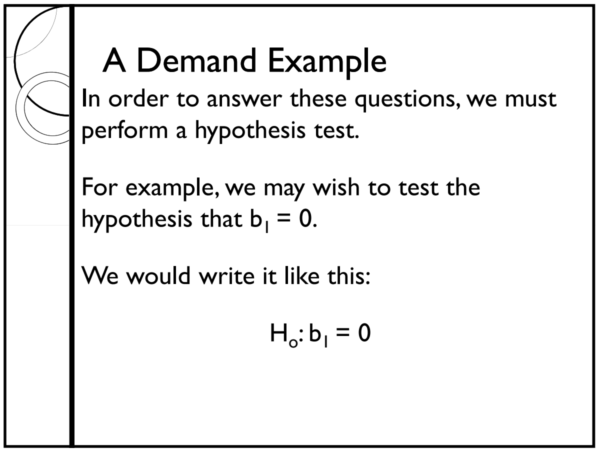


A Demand Example

5. Having estimated the demand function, we may want to find out if the results conform to our theory.

For example:

- can we say with statistical proof that $b_2<0$
- can we show that $b_1>0$



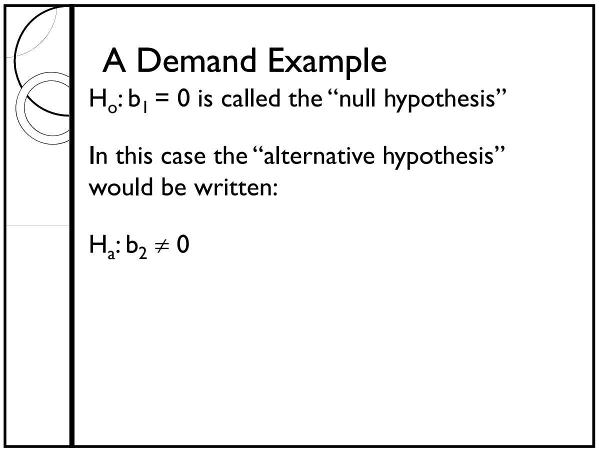
A Demand Example

In order to answer these questions, we must perform a hypothesis test.

For example, we may wish to test the hypothesis that $b_1 = 0$.

We would write it like this:

$$H_o: b_1 = 0$$

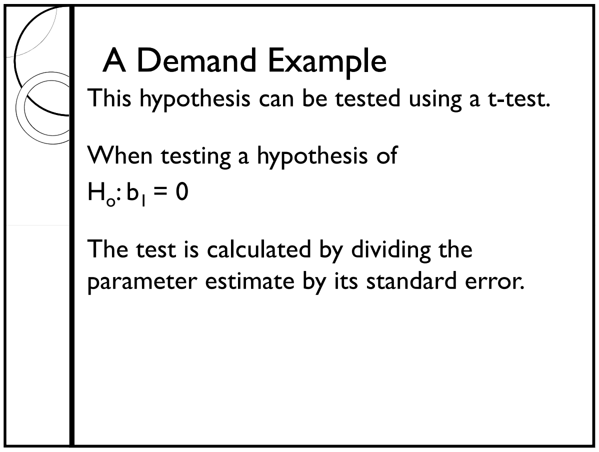


A Demand Example

H_o: $b_1 = 0$ is called the "null hypothesis"

In this case the "alternative hypothesis" would be written:

H_a: $b_2 \neq 0$

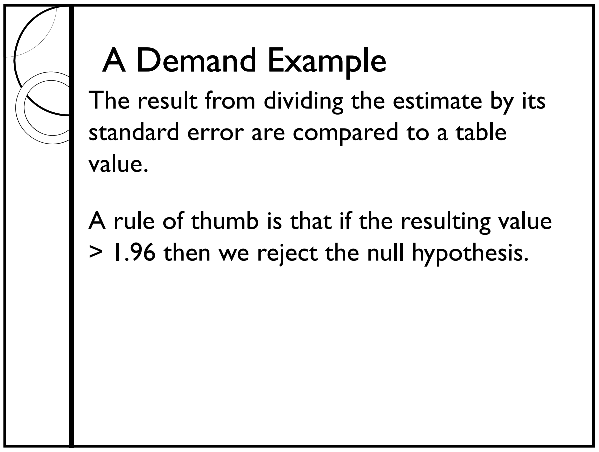


A Demand Example

This hypothesis can be tested using a t-test.

When testing a hypothesis of
H_o: $b_1 = 0$

The test is calculated by dividing the parameter estimate by its standard error.



A Demand Example

The result from dividing the estimate by its standard error are compared to a table value.

A rule of thumb is that if the resulting value > 1.96 then we reject the null hypothesis.

A Demand Example

6. Forecasting

Suppose that given our estimated demand function, we want to know what quantity would be demanded if price = $4.50

A Demand Example

A forecast could be obtained from:

$\hat{Q}$ = 76.05 – 3.88 (4.50) = 58.59

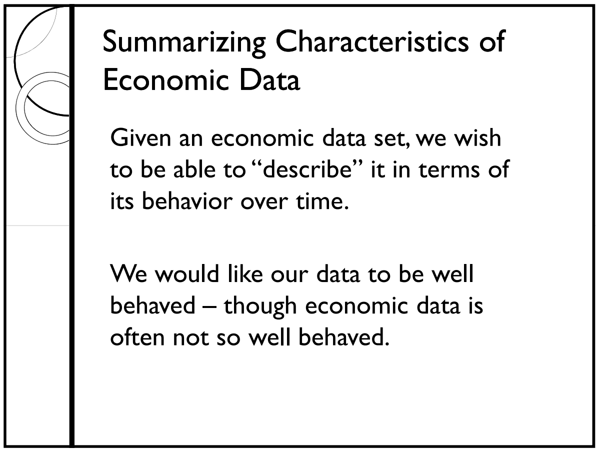
When forecasting time series, you will want to re-estimate the parameters with each subsequent realization and forecast only one-step-ahead.

Basic Regression

The most basic regression approach is known as Ordinary Least Squares (OLS).

Under certain assumptions the OLS estimator has some very attractive statistical properties.

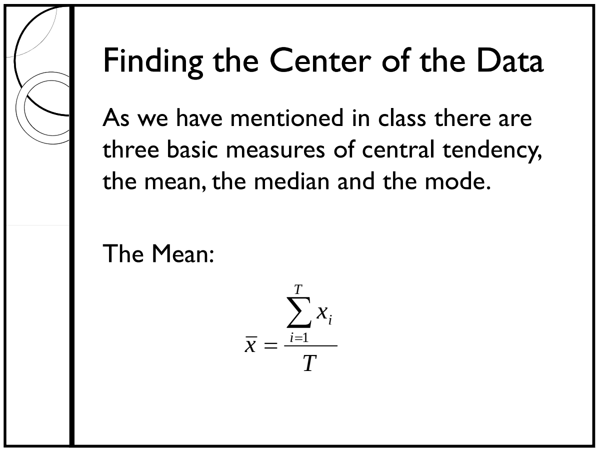
The "Least Squares" portion of the name indicates that the OLS estimator is one that minimizes squared errors.



Summarizing Characteristics of Economic Data

Given an economic data set, we wish to be able to "describe" it in terms of its behavior over time.

We would like our data to be well behaved – though economic data is often not so well behaved.

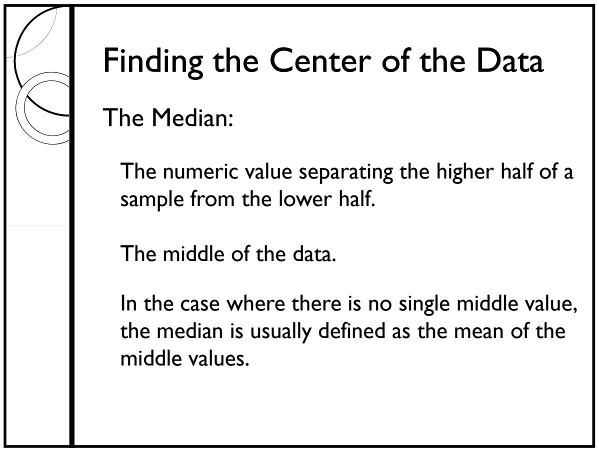


Finding the Center of the Data

As we have mentioned in class there are three basic measures of central tendency, the mean, the median and the mode.

The Mean:

$$\bar{x} = \frac{\sum_{i=1}^{T} x_i}{T}$$



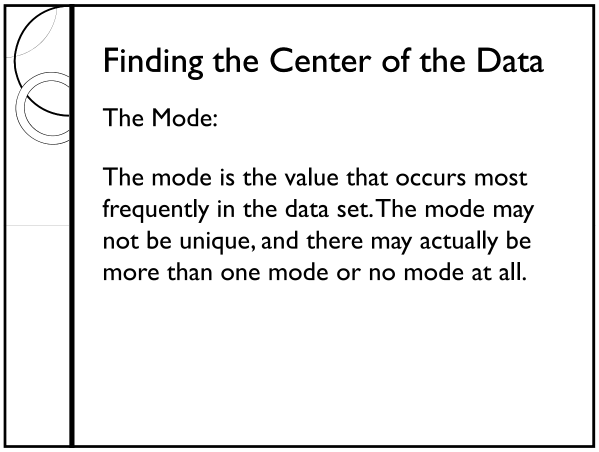
Finding the Center of the Data

The Median:

The numeric value separating the higher half of a sample from the lower half.

The middle of the data.

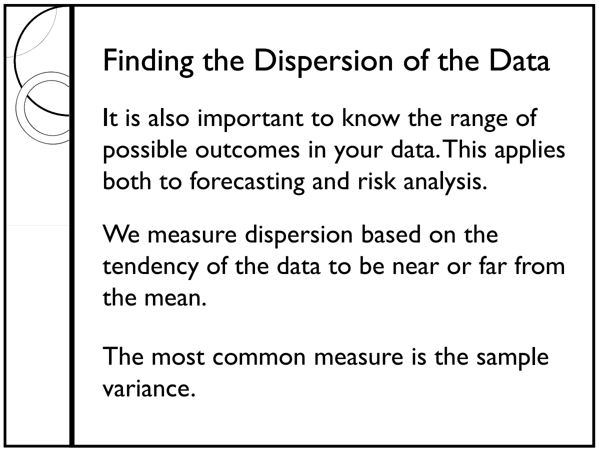
In the case where there is no single middle value, the median is usually defined as the mean of the middle values.



Finding the Center of the Data

The Mode:

The mode is the value that occurs most frequently in the data set. The mode may not be unique, and there may actually be more than one mode or no mode at all.

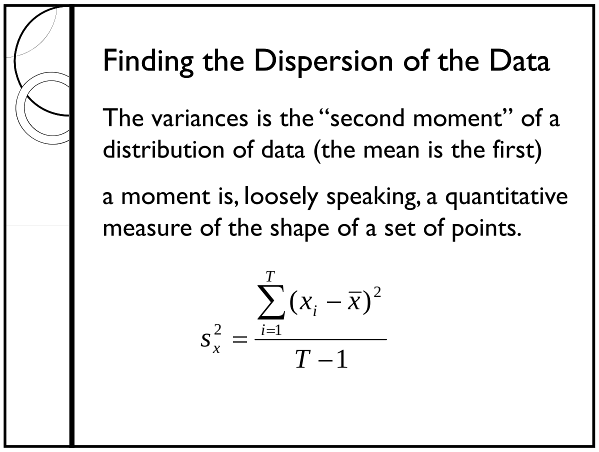


Finding the Dispersion of the Data

It is also important to know the range of possible outcomes in your data. This applies both to forecasting and risk analysis.

We measure dispersion based on the tendency of the data to be near or far from the mean.

The most common measure is the sample variance.



Finding the Dispersion of the Data

The variances is the "second moment" of a distribution of data (the mean is the first)

a moment is, loosely speaking, a quantitative measure of the shape of a set of points.

$$s_x^2 = \frac{\sum_{i=1}^{T}(x_i - \overline{x})^2}{T-1}$$

Finding the Dispersion of the Data

Two things should be obvious from the formula for the variance:

- The larger the variance, the greater the dispersion
- The variance is dependent on the scale of the mean (in other words you can't directly compare variances for two series of data with widely different mean values).

A Basic Regression Model

Now suppose that the data are available in time series and that

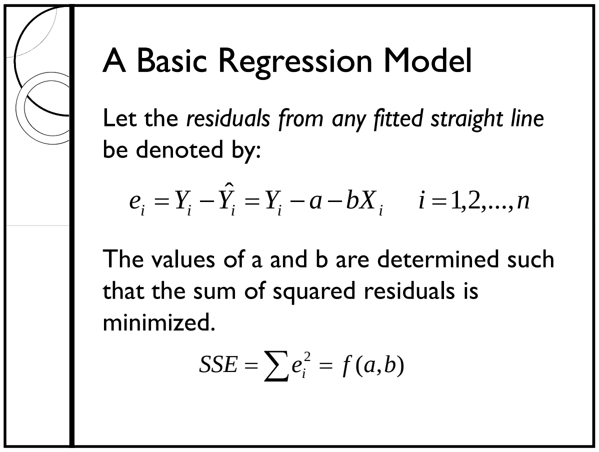
X_t = aggregate real disposable personal income in year t

Y_t = aggregate real vacation expenditure in year t.

A Basic Regression Model

Whether the data are time series or cross sectional, the simplest version of the two-variable model is:

$$Y_i = \alpha + \beta X_i + e_i$$



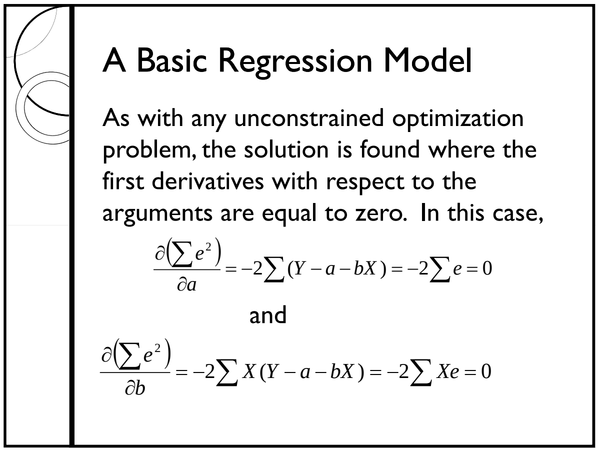
A Basic Regression Model

Let the *residuals from any fitted straight line* be denoted by:

$$e_i = Y_i - \hat{Y}_i = Y_i - a - bX_i \qquad i = 1,2,\ldots,n$$

The values of a and b are determined such that the sum of squared residuals is minimized.

$$SSE = \sum e_i^2 = f(a,b)$$



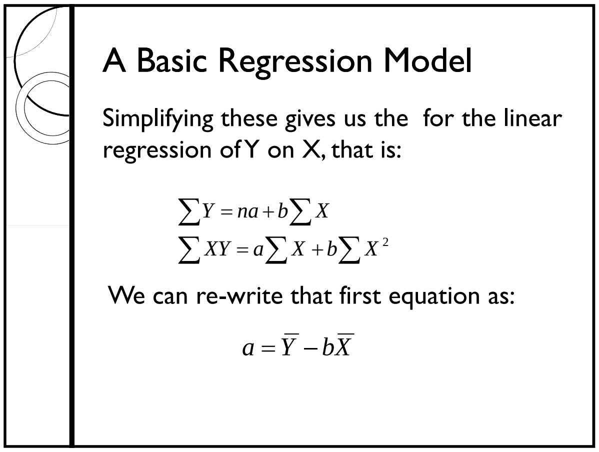
A Basic Regression Model

As with any unconstrained optimization problem, the solution is found where the first derivatives with respect to the arguments are equal to zero. In this case,

$$\frac{\partial\left(\sum e^2\right)}{\partial a} = -2\sum (Y - a - bX) = -2\sum e = 0$$

and

$$\frac{\partial\left(\sum e^2\right)}{\partial b} = -2\sum X(Y - a - bX) = -2\sum Xe = 0$$



A Basic Regression Model

Simplifying these gives us the for the linear regression of Y on X, that is:

$$\sum Y = na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

We can re-write that first equation as:

$$a = \overline{Y} - b\overline{X}$$

A Basic Regression Model

Substituting for a in the second normal equation gives:

$$b = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Thus, the least-squares slope can be estimated from the equation, and the result substituted in to calculate the intercept.

A Basic Regression Model

The least-squares line has three important properties. It minimizes sum of squared residuals (errors), it passes through the mean point $(\overline{X}, \overline{Y})$ and the least squares residuals have zero correlation in the sample with the values of X

A Basic Regression Model

Fortunately, we don’t have to memorize all of these formulas, or try to translate them to a spreadsheet.

Excel has a built in Regression program that can be found (if it is installed) under the “Data” tab by selecting “Data Analysis”

How “Good” is Your Model?

One measure (albeit an imperfect one) of the “goodness” of a model is the *coefficient of determination* or R^2.

R^2 indicates the proportion of the total variation in the dependent variable explained by the independent variables.

How “Good” is Your Model?

R^2 can be likened to an *in-sample* measure of forecasting accuracy.

- How well $\hat{y}_i$ predicts y_i
- Can only be measured in-sample

How “Good” is Your Model?

Perhaps a better measure of the goodness of your model is how well it forecasts out-of-sample.

You can think of MSE as a kind of out-of-sample R^2.