A Review of Estimation Models

Su, Chapter 6

Chapter Goals

1. Provide an Overview of Linear Regressions and OLS
2. Familiarize Students With Regression Techniques
3. Link Economic Theory, Economic Data and Regressions
4. Demonstrate How To Use Excel To Run Regressions
5. Explain How To Interpret Regression Results

Definitions

- *Population*: The complete set of pertinent data that contains certain characteristics of interest
- *Sample*: A subset of data drawn from a population - ideally contains the same characteristics of interest as the population
- *Parameter*: Any measurable characteristic of a population
- *Statistic or Estimate*: Any value computed entirely from the sample and used to estimate a population parameter
- *Estimator*: The method used to obtain an *estimate*
- *Regression Model*: A mathematical representation of the relationship between two or more variables
- *Simple Regression Analysis*: A linear or nonlinear regression model that has only two unknown parameters $Y = B_0 + B_1 X \ \ (6.1a)$
- *Multiple Regression Analysis*: A linear or nonlinear regression model that has more than two unknown parameters

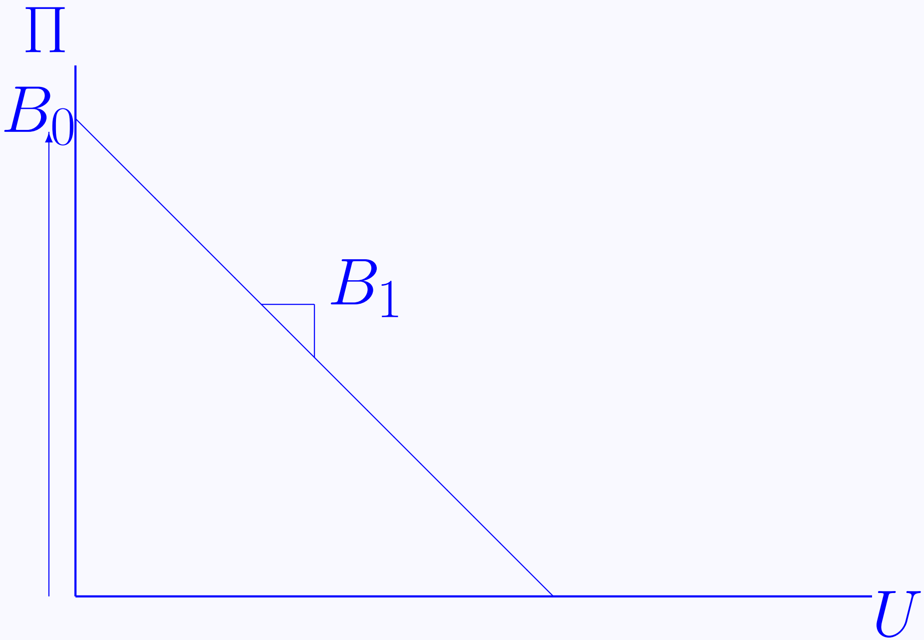
A Population Regression Model

- Suppose we are interested in studying the relationship between that inflation rate and the unemployment rate in an economy
- The Phillips Curve is an economic model that describes this relationship.
- According to this theory, the Phillips Curve relationship can be described by

$$\Pi = B_0 + B_1 U$$

  where Π is the inflation rate and U is the unemployment rate
- The relationship between Π and U is linear so B_0 is the intercept parameter and B_1 the slope parameter

Slope and Intercept Parameters



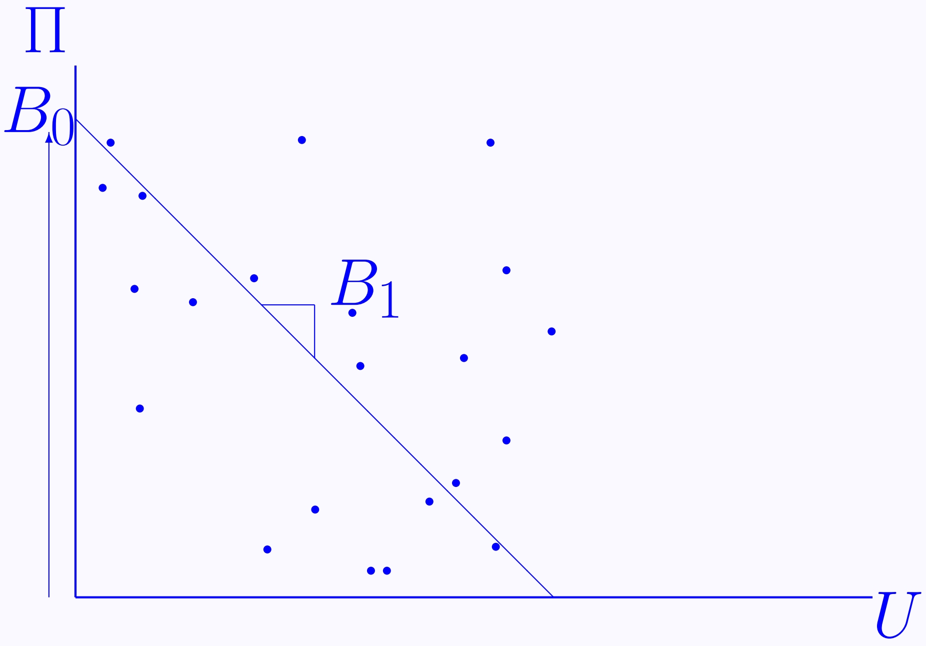
Economic Interpretation of Regression Line

- This straight line is called the Phillips Curve Relationship in Economics
- Given an Unemployment Rate (U_0) in an economy, can expect the Inflation Rate to be P_0, or

$$P_0 = E(P|U_0) = B_0 + B_1 U_0$$

- Points are observations - all inflation/unemployment combinations observed in a period of time
- The Population Regression Line is a hypothetical straight line passing through the distribution, cutting it in half
- The Population is all observations - the population is not observed

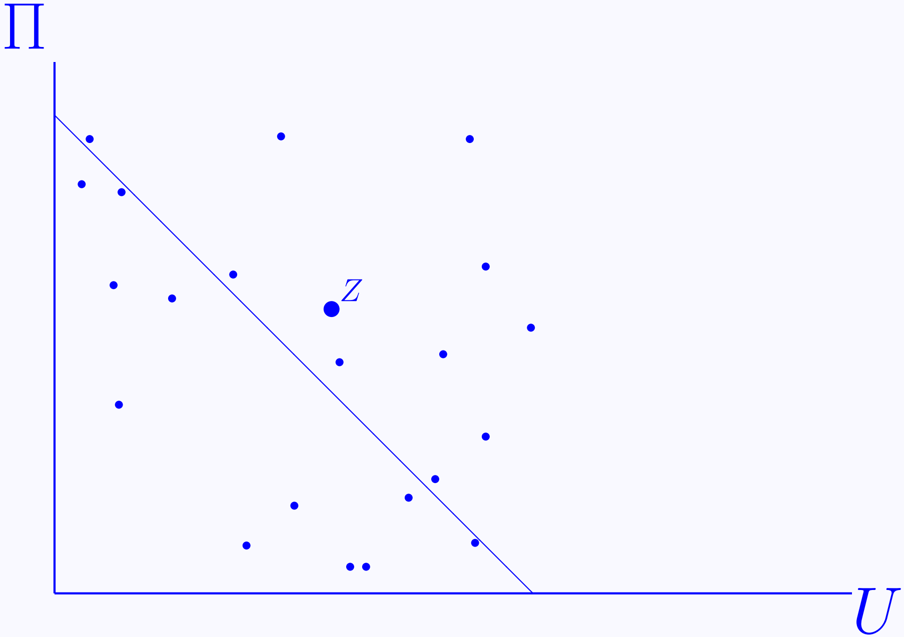
Regression Line and Observations



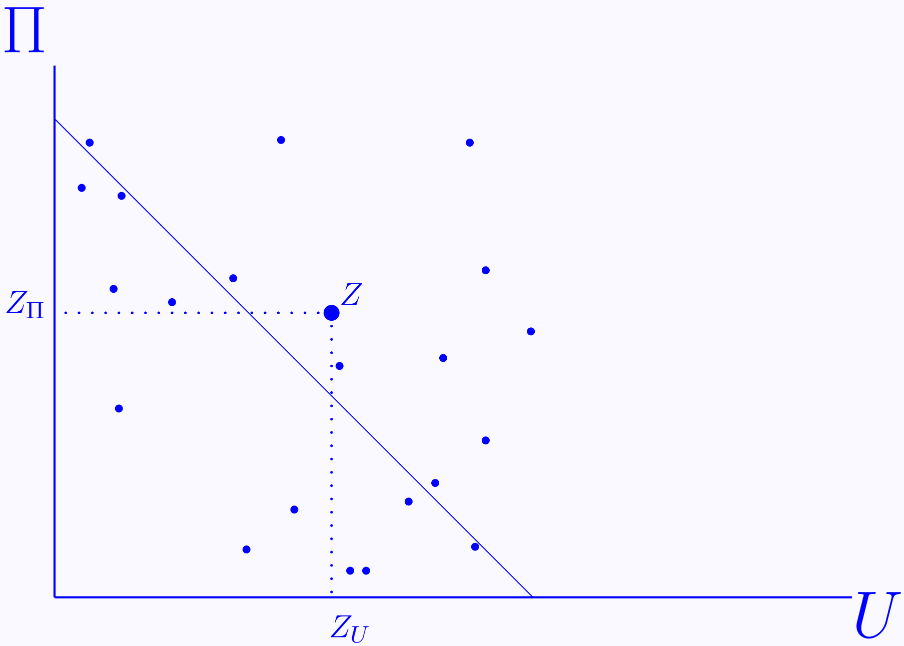
Observations and Deviations

- Observations are defined by their corresponding values of Π and U
- For example the observation Z would correspond to (Π_Z, U_Z)
- Observations deviate from the Population Regression Line both vertically and horizontally
- Define the vertical distance between any observation and the Population Regression Line as u_i
- Note that for each observation a "deviation" (u_i) can easily be calculated
- For observation Z this deviation is u_z

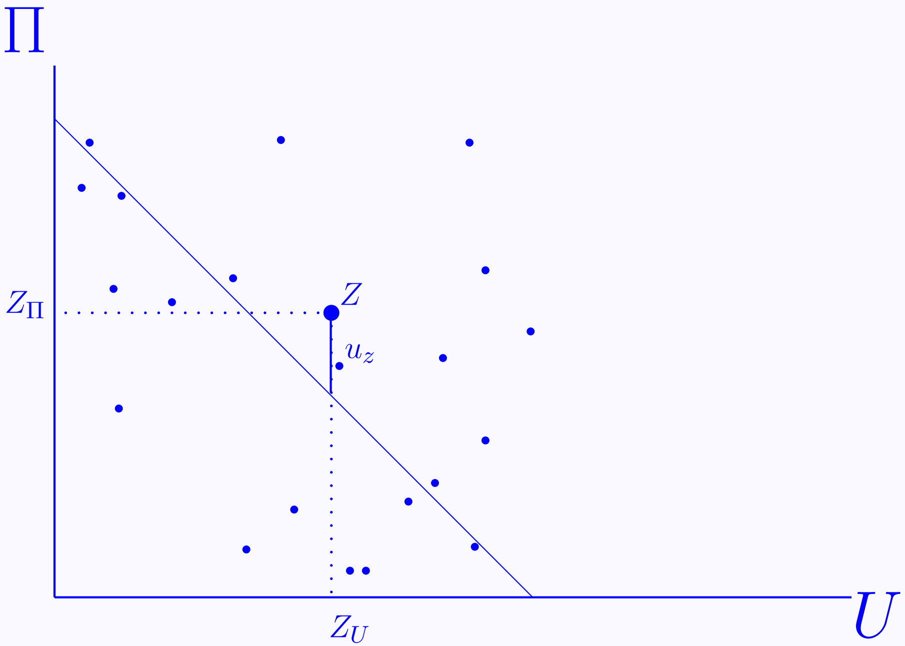
Regression Line and Deviations



Regression Line and Deviations



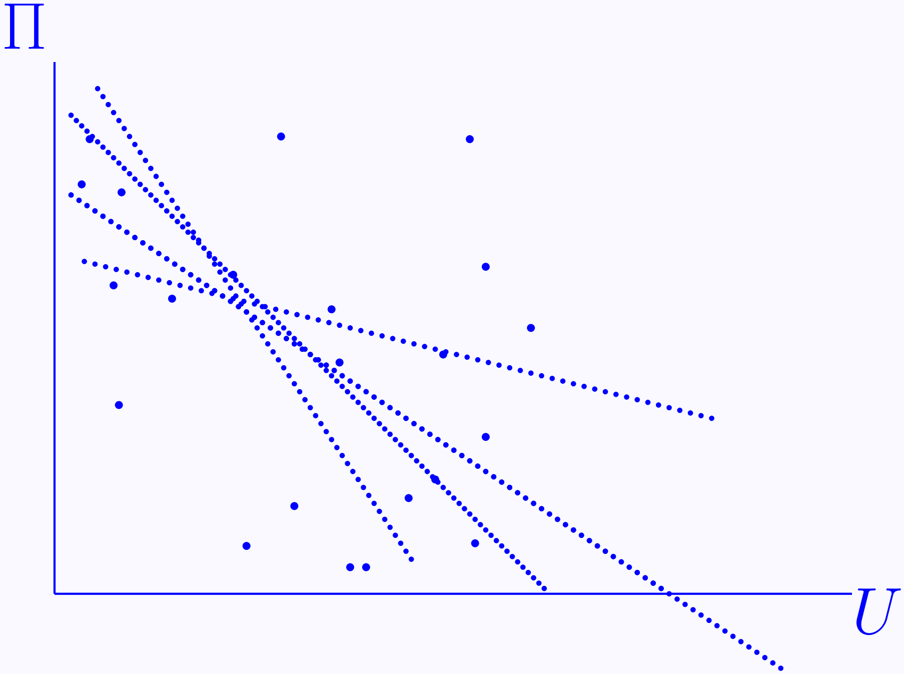
Regression Line and Deviations



The *Sample* Regression Model

- The parameters of the Population Regression Model are unknown
- They must be estimated by applying statistical methods to a **sample** of observations from the **population**
- If the sample is large enough, its distribution will be similar to the distribution of the population $Y_i = \beta_0 + \beta_1 X_i + e_i$
- For the Phillips Curve example $\Pi_i = \beta_0 + \beta_1 U_i + e_i$
- The is represent observations in the sample
- β_0 and β_1 are estimates of the population parameters B_0 and B_1
- The e_is are the sample disturbance term or sample errors - the distance between each observation and the sample regression line
- These are observed, not the u_is from the population

Regression Lines and Parameters



How Can The “Right” Sample Regression Line Be Located?

- By putting it “as close as possible” to all the sample points.
- What does “as close as possible” mean?

How Can The "Right" Sample Regression Line Be Located?

- By putting it "as close as possible" to all the sample points.
- What does "as close as possible" mean?
- Ordinary Least Squares (OLS):Technique to find the minimum sum of squared deviations, i.e. the regression line

OLS “Solutions” to the Minimization Problem for Regression Coefficients

- Note that

$$\Pi_i = \beta_0 + \beta_1 U_i + e_i \longrightarrow e_i = \Pi_i - \beta_0 - \beta_1 U_i$$

- Minimization Problem

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (\Pi - \beta_0 - \beta_1 U)^2$$

- Solution (“Normal Equations”):

$$\beta_0 = \frac{\sum U \sum U\Pi - \sum \Pi \sum U^2}{(\sum U)^2 - n \sum U^2}$$

$$\beta_1 = \frac{\sum U \sum \Pi - n \sum U\Pi}{(\sum U)^2 - n \sum U^2}$$

“Normal Equations”

$$\beta_0 = \frac{\sum U \sum U\Pi - \sum \Pi \sum U^2}{(\sum U)^2 - n \sum U^2}$$

$$\beta_1 = \frac{\sum U \sum \Pi - n \sum U\Pi}{(\sum U)^2 - n \sum U^2}$$

Three Types of Regression Parameters

- **Regression Coefficients**:

  β_0 and β_1 the slope and intercept - Location

- **Disturbance Variance**:

  σ_u^2, or σ_e^2 measures goodness of fit

- **Coefficients of Correlation and Determination**:

  r and R^2 measures association

Disturbance Variances

- Refers to the variance of the error term in the regression model - the e_is.
- **Variance or Standard Deviation**: Measures the spread or dispersion of a sample or population
- Always measured from the central location/mean/expected value
- **Population Variance**: For any Population Z, the Variance of the Population is the sum of squares of Z divided by population size

$$var(Z) = \sigma_Z^2 = E[Z_i - E(Z_i)]^2 = \frac{\sum(Z_i - \mu_Z)^2}{N}$$

- **Sample Variance**: For any Sample X, the Variance of the Sample is the sum of squares of X divided by sample size

$$var(X) = s_X^2 = E[X_i - E(X_i)]^2 = \frac{\sum(X_i - \mu_Z)^2}{N-1}$$

Standard Error of Estimate

- Want to measure the spread of the actual values around the regression line
- "Goodness of Fit" measurement
- Must use sample disturbances or residuals

$$\sigma_e^2 = \frac{\sum (e - E[e])^2}{n-2} = \frac{\sum e^2}{n-2}$$

Standard Errors of Estimated Slope and Intercept

- β_0 and β_1 are random variables, need to know about their distribution in order to understand how accurate they are
- Means are population parameters (B_0 and B_1)
- Variances:

$$VAR(\beta_0) = \frac{\sigma_e^2}{\sum x^2}$$

$$VAR(\beta_1) = \sigma_e^2 \frac{\sum X^2}{n \sum x^2}$$

$$x = X - E[X]$$

Measures of Association

- **Covariance** (between two random variables X and Y):Measures association but not causality

$$Cov(X,Y) = E[X - E(X)][Y - E(Y)] = \frac{\sum(X_i - \mu_X)(Y_i - \mu_Y)}{N-1}$$

- **Coefficient of Correlation** (between two random variables X and Y): Measures association, addresses problem that covariance is affected by units of measurement

$$corr(X,Y) = r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Properties of the Regression Parameters: Disturbance Variance

- The disturbance term (u_i in the population regression model, e_i in the sample regression model) represents unexplained variation in Y
- Where does this unexplained variation come from?

1. Unpredictable randomness
2. Factors that affect Y which were left out of regression
3. Measurement error in data

- Assumptions about the behavior of the error term
  - Mean zero $E(u_i) = 0$
  - Constant variance $E(u_i u_{i+j}) = \sigma_u^2 \; if \; j = 0$
  - No serial correlation $E(u_i u_{i+j}) = 0 \; if \; j > 0$
  - Independent of X $E(u_i x_i) = 0$
  - Normally distributed $u_i \; N(0, \sigma_u^2)$

Properties of the Estimated Slope and Intercept Parameters

- The population regression parameters do not vary, but the sample regression parameters do
- Think about taking many small samples from a population (say randomly drawn samples of 100 from a population with 1,000,000 observations)
- Would get many OLS estimates. The OLS estimates of slope and intercept term are random variables
- Usually have only one observation
- Need to know about the distribution in order to make inferences about these estimated parameters
- How do we describe distributions? Mean and variance

Recapitulation

- Regression line: $Y_i = \beta_0 + \beta_1 X_i$
- An observation, Point Z, deviates vertically from the regression line
- This vertical distance is e_Z
- Can calculate a e_i for each observation, called “deviations” or “regression errors”
- The regression error is a random variable, and we assume

1. $E(e) = 0$
2. $var(e) = \sigma^2$
3. $cov(e_i, e_j) = 0$
4. $e \sim N(0, \sigma^2)$

Estimating the variance σ^2

- Disturbances are $e_i = y_i - \beta_0 - \beta_1 x_i$
- Use the e_i's to estimate variance

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

Distribution of Regression Parameters

- β_0 and β_1 are observations of random variables
- Need to describe their distribution - mean and variance
- Means are the population parameters (B_0 and B_1) - *unbiasedness*
- **Variance of** β_0: Given that $\beta_0 = \overline{y} - \beta_1\overline{x}$

$$var(\beta_0) = \hat{\sigma}^2 \frac{\sum x^2}{T\sum(x_i - \overline{x})^2}$$

- **Variance of** β_1:

$$var(\beta_1) = E[\beta_1 - E[\beta_1]]^2 = \frac{\hat{\sigma}^2}{\sum(x_1 - \overline{x})^2}$$

note that although β_1 depends on y, it's variance does not

- **Covariance**:

$$cov(\beta_0, \beta_1) = \hat{\sigma}^2 \frac{\overline{x}}{\sum(x_i - \overline{x})^2}$$

Probability Distribution of Least Squares Estimators

$$\beta_0 \sim N\left[B_0, \hat{\sigma}^2 \frac{\sum x^2}{T\sum(x_i - \overline{x})^2}\right]$$

$$\beta_1 \sim N\left[B_1, \frac{\hat{\sigma}^2}{\sum(x_1 - \overline{x})^2}\right]$$

Properties of Regression Parameters

- **Linear**: The OLS model is clearly a linear model
- **Unbiased**: The expected value of the regression parameters is are population parameters
- **Best**: In the sense that among all linear unbiased estimators, the variance of the OLS estimator is smallest
- **Gauss-Markov Theorem**: Proves that OLS is the “best” linear unbiased estimator

Multiple Regression

- More than one variable can affect another
- Keynesian Investment Function:

$$I = \gamma_0 + \gamma_1 Y - \gamma_2 R$$

- Need to be able to sort out different effects of R and Y on I
- Use Multiple regression - more than one explanatory variable

$$y_t = B_0 + B_1 x_{t1} + B_2 x_{t2} + e_t$$

- The explanatory variables affect y separately: $\frac{\partial y_t}{\partial x_{t2}} = B_2$ and $\frac{\partial y_t}{\partial x_{t1}} = B_1$
- Estimates of B_1 and B_2 depend on both x_{t1} and x_{t2}

A General Statistical Model

- $y_t = B_1 + B_2 x_{2t} + B_3 x_{3t} + \ldots + B_k x_{kt} + e_t$

1. $E(e_t) = 0$
2. $var(e_t) = \sigma^2$
3. $cov(e_t, e_s) = 0 \; for \; t \neq s$
4. $e_t \sim N(0, \sigma^2)$

- Error variance estimation ("standard error of the estimate"): $\hat{\sigma}^2 = \frac{\sum \hat{e}_t^2}{T-K}$
- Variances of estimates

$$var(b_2) = \frac{\hat{\sigma}^2}{(1-r_{23}^2)\sum(x_{t2}-\overline{x_2})^2}$$
$$var(b_3) = \frac{\hat{\sigma}^2}{(1-r_{23}^2)\sum(x_{t3}-\overline{x_3})^2}$$
$$r_{23} = \frac{\sum(x_{t2}-\overline{x}_2)(x_{t3}-\overline{x}_3)}{\sqrt{\sum(x_{t2}-\overline{x}_2)}\sqrt{\sum(x_{t3}-\overline{x}_3)}}$$

Variance-Covariance Matrix

- For the statistical model

$$y_t = B_0 + B_1 x_{t1} + B_2 x_{t2} + e_t$$

- The OLS estimators b_1, b_2 and b_3 have a variance-covariance matrix

$$\begin{bmatrix} var(b_1) & cov(b_1, b_2) & cov(b_1, b_3) \\ cov(b_2, b_1) & var(b_2) & cov(b_2, b_3) \\ cov(b_3, b_1) & cov(b_3, b_2) & var(b_3) \end{bmatrix}$$

Analysis of Variance tables

- Generated by regression packages, show the variation in the dependent variable and other important measures of variation in regression models
- Have a common format

Source of Variation	DF	Sum of Squares	Mean Square
Explained	K	ESS	$\frac{ESS}{K}$
Unexplained	$T-K$	RSS	$\frac{RSS}{T-K}$
Total	T	TSS	$\frac{TSS}{T}$

Goodness-of-fit

- For multiple regression models, measure of goodness of fit is R^2, the *coefficient of determination*

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_t - \overline{y}_t)^2}{\sum(y_t - \overline{y}_t)^2}$$

- Note that $0 \leq R^2 \leq 1$
- **SSR**: "Regression Sum of Squares" - the amount of variation in y explained by the entire regression model
- **SST**: "Total Sum of Squares" - the total variation in y
- Also can use *adjusted* R^2 to adjust for multiple regressors

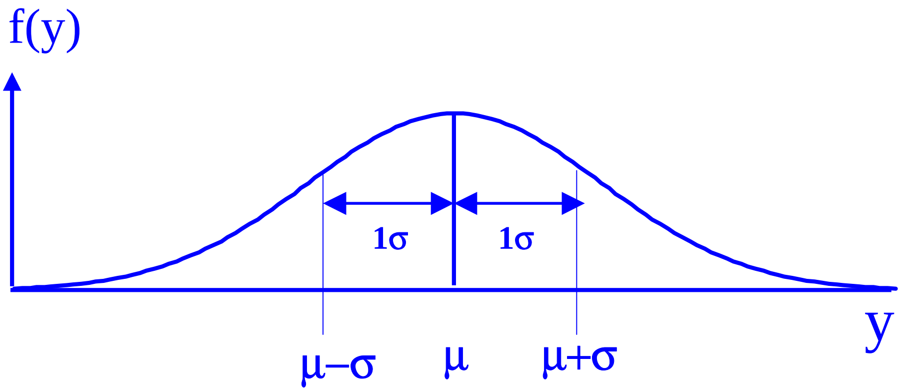
$$\overline{R^2} = 1 - \frac{SSE/(T-K)}{SST/(T-1)}$$

Statistical Inference

- Recall that the OLS estimators (b_1, b_2, etc.) are *random variables*
- This implies that we can make statistical inferences about the values of these parameters
- In particular, we can learn how close or far these values might be from some important values, like zero, or some value given to us by economic theory
- One distribution used for these tests is the *normal distribution*
- This distribution has two parameters: μ mean and σ^2 variance
- Graphically it is a “bell curve”
- The probability distribution function (pdf) is

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}$$

The Normal Distribution



The Standardized Normal Distribution

- "Standardizes" a normally distributed random variable by subtracting off the mean and dividing by the standard deviation of the original random variable
- Usually denoted by the variable Z
- Expression is

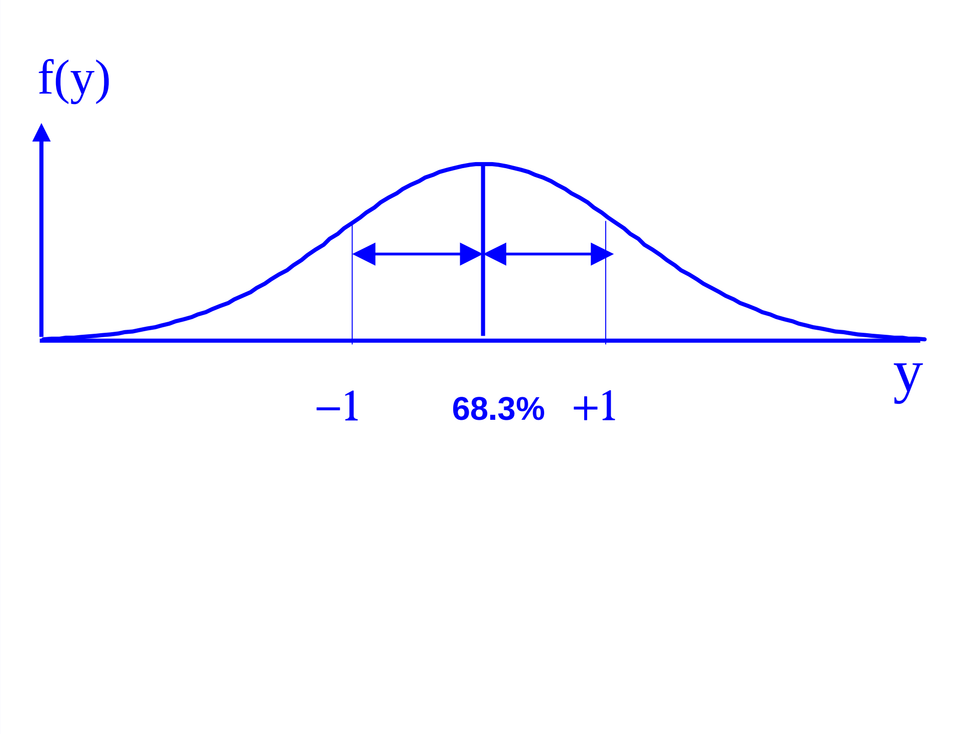
$$Z = \frac{y - \mu}{\sigma}$$

  where y is a normally distributed random variable, μ is the mean of y and σ is the standard deviation of y

- $Z \sim N(0, 1)$
- Probability distribution function is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

The Standard Normal Distribution



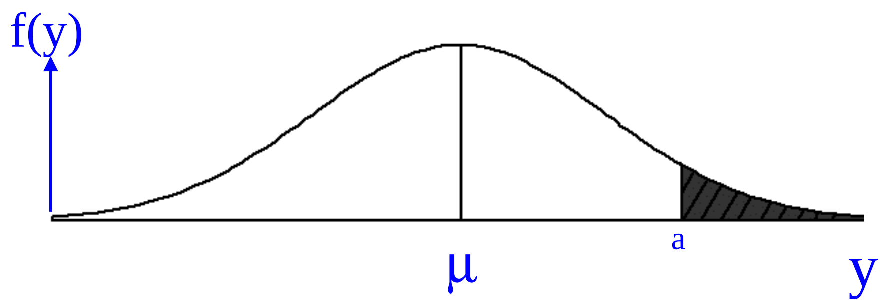
Probabilities in Continuous Distributions

- Recall that probabilities are expressed as areas under the curve in these continuous distributions
- Statistical inference of the type "Given the mean and standard deviation of a normally distributed random variable, what is the probability of observing a value larger than a" is typically done
- Formally, for a normally distributed random variable Y with mean μ and standard deviation σ

$$P[Y \geq a] = P[\left[\frac{Y-\mu}{\sigma} \geq \frac{a-\mu}{\sigma}\right] = P[\left[Z \geq \frac{a-\mu}{\sigma}\right]$$

- Use standard normal statistical tables or spreadsheets to compute this

Probabilities in Continuous Distributions



Student's t Distribution

- Since generally the population variance of an OLS estimator b_k , $var(b_k)$, is unknown, we estimate it with $v\hat{a}r(b_k)$ which uses $\hat{\sigma}^2$ instead of σ^2
- This is the basis of most hypothesis testing in regression models
- The test statistic

$$t = \frac{b_k - \beta_k}{\sqrt{v\hat{a}r(b_k)}} = \frac{b_k - \beta_k}{s.e.(b_k)}$$

  has a Student's t distribution with T_K degrees of freedom
- Looks like a normal distribution, but has fatter tails
- Use t tables or functions in spreadsheets for this statistic

t-tests in Regressions

- Consider the statistical model

$$y_t = B_0 + B_1 x_{t1} + B_2 x_{t2} + e_t$$

- t-tests can be used to test any linear combination of the regression coefficients in this model

H_o: $\beta_1 = 0$

H_o: $\beta_1 + \beta_2 = 4$

H_o: $3\beta_1 - 7\beta_2 = 11$

- Every such t-test has $T-K$ degrees of freedom where K is the number of coefficients estimated including the intercept

One-tailed Tests

- For

$$y_t = B_0 + B_1 x_{t1} + B_2 x_{t2} + e_t$$

- Suppose we want to test the null hypothesis $H_o : \beta_2 = 0$ against the alternative $H_a : \beta_2 > 0$
- The test statistic is $t = \frac{b_2}{s.e.(b_2)}$ which is distributed t with $T - K = T - 3$ degrees of freedom
- We need to select a value for α - a significance level - for the test in order to find a critical value t_c to compare the value of the t-statistic to
- If $t > t_c$ we reject the null, else accept

One-tailed Test