Linear regression model: the basics

BUFN 758N

Prof. Skoulakis

Regression

- Regression is the single most important tool at the econometrician's disposal
- What is regression analysis?
- It is concerned with the description and evaluation of the relationship between a variable typically called the dependent variable, and one or more other variables, typically called the independent or explanatory variables.

Notation

- We denote the dependent variable by y and the independent variables by $x_1, x_2, \ldots, x_k$ where k is the number of independent variables
- Alternative terminology for the y and x variables

y	x
dependent variable	independent variables
regressand	regressors
effect variable	causal variables
explained variable	explanatory variables

- Note that there can be many explanatory variables (x), but we will focus on the case where there is only one x variable to start with.

Simple Linear Regression

- We first focus on the case where y depends on only one x variable
- Strictly speaking, there are two regressors: a constant (typically set equal to 1) and the x variable
- Examples of the kind of relationship that may be of interest include
  - Variation of average asset returns with their level of market risk
  - Determining the relationship between stock prices and dividends

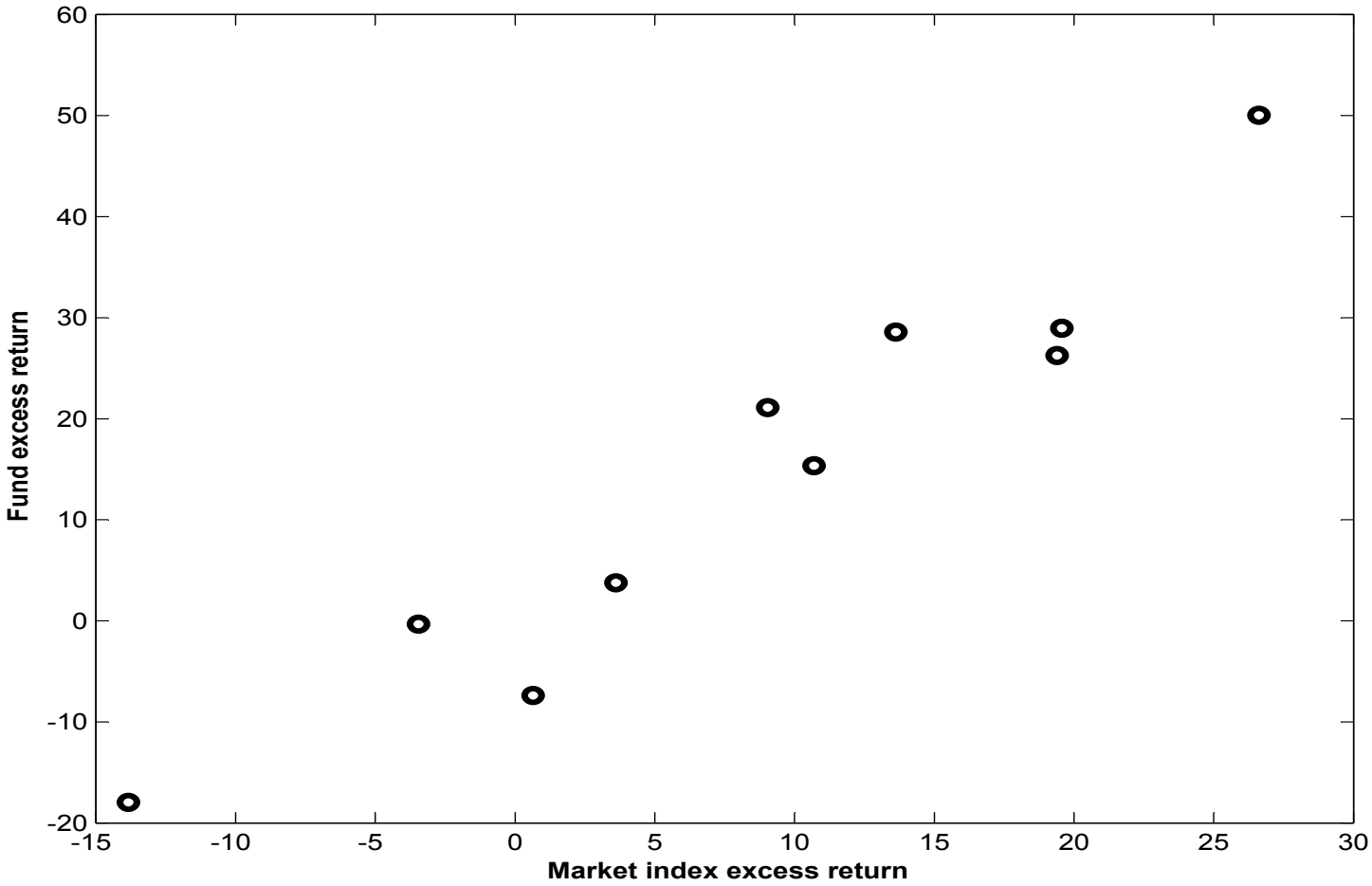
Simple Linear Regression: An Example

- Suppose that we have the following data on the excess returns on a fund manager's portfolio (fund "XYZ") together with the excess returns on a market index:

Year	1	2	3	4	5	6	7	8	9	10
XYZ	15.36	-0.33	21.10	-17.97	28.97	28.57	26.25	50.03	-7.39	3.75
Market	10.70	-3.46	9.05	-13.83	19.57	13.62	19.40	26.62	0.64	3.61

- Intuition suggests that the *beta* on this fund is positive, and hence we would like to find whether there appears to be a relationship between market index excess return (x) and the fund XYZ excess return (y) given the available data.

Graph (Scatter Plot)



Finding the Line of Best Fit

- We could use the equation for a straight line

$$y = a + bx$$

  to obtain the line that best "fits" the data
- However, this equation ($y = a + bx$) is completely deterministic. That is, it does not allow for randomness.
- This is not realistic, as we do not expect such relationships, especially in economics and finance.
- Hence, we add a random disturbance term, denoted by u, to the above equation to obtain

$$y_t = \alpha + \beta x_t + u_t, \quad t = 1, 2, \ldots, 10$$

Why do we include a Random Disturbance term

The disturbance term can capture a number of features

- It is practically impossible to identify all of the determinants of y
- There may be errors in the measurement of y that cannot be modeled
- There may be random exogenous influences on y which we cannot model

Estimating the Regression Parameters

- Determining the line of best fit boils down to determining the regression parameters (or coefficients) α and β
- The standard terminology for this task is *estimation*
- Estimation is the process of using the available data to obtain "best guesses" for the unknown underlying parameters
- Obtaining a "best guess" for an unknown parameter requires a criterion
- In the linear regression setting, loosely speaking, the criterion is to minimize the distances from the data points to the fitted line (so that the line fits the data as closely as possible)

Ordinary Least Squares

- The most common method used to fit a line to the data is known as ordinary least squares (OLS)
- As the terminology suggests, we take the square of the distance from each data point to the line, add all the square distances, and then select the parameter estimates to minimize the sum of squared distances
- Notation
  - y_t : actual data point at time t
  - $\hat{y}_t$: fitted value from the regression line at time t
  - $\hat{u}_t$: residual $y_t - \hat{y}_t$ at time t

Ordinary Least Squares (cont'd)

- The residual $\hat{u}_t = y_t - \hat{y}_t$ can be thought of as an estimate of the true unknown disturbance u_t
- The residual $\hat{u}_t = y_t - \hat{y}_t$ is the distance between the data point and the fitted straight line at time t

How OLS works

- The OLS criterion for obtaining estimates for α and β is to minimize the sum of residuals

$$\hat{u}_1^2 + \cdots + \hat{u}_T^2 = \sum_{t=1}^{T} \hat{u}_t^2 = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2$$

- But the fitted values are given by $\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t$ and so the sum of residuals is a function of the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ which can be expressed as

$$S(\hat{\alpha}, \hat{\beta}) = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{T} \left(y_t - \hat{\alpha} - \hat{\beta} x_t \right)^2$$

Deriving the OLS Estimators

- The OLS estimators can be derived in a number of different ways
- One standard way is to solve the corresponding first order conditions (FOC). In other words, we set the derivatives of $S(\hat{\alpha}, \hat{\beta}) = \sum_{t=1}^{T} \left(y_t - \hat{\alpha} - \hat{\beta} x_t \right)^2$ with respect to $\hat{\alpha}$ and $\hat{\beta}$ equal to zero, and solve the two-equation system
- The two equations are

$$\frac{\partial S(\hat{\alpha}, \hat{\beta})}{\partial \hat{\alpha}} = -2 \sum_{t=1}^{T} (y_t - \hat{\alpha} - \hat{\beta} x_t) = 0 \quad (A)$$

$$\frac{\partial S(\hat{\alpha}, \hat{\beta})}{\partial \hat{\beta}} = -2 \sum_{t=1}^{T} x_t (y_t - \hat{\alpha} - \hat{\beta} x_t) = 0 \quad (B)$$

Deriving the OLS Estimators (cont'd)

- Define the sample averages $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$ and $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$
- From (A) we obtain $\sum_{t=1}^{T} y_t - T\hat{\alpha} - \hat{\beta}\sum_{t=1}^{T} x_t = 0 \Leftrightarrow$ $T\bar{y} - T\hat{\alpha} - \hat{\beta}\bar{x} = 0 \Leftrightarrow \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
- Substitution of the above expression for $\hat{\alpha}$ in equation (B) yields

$$\sum_{t=1}^{T} x_t \left(y_t - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_t\right) = 0$$

$$\Leftrightarrow \quad \sum_{t=1}^{T} x_t(y_t - \bar{y}) - \hat{\beta}\sum_{t=1}^{T} x_t(x_t - \bar{x}) = 0$$

$$\Leftrightarrow \quad \hat{\beta} = \frac{\sum_{t=1}^{T} x_t(y_t - \bar{y})}{\sum_{t=1}^{T} x_t(x_t - \bar{x})} = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

Deriving the OLS Estimators (cont'd)

- An alternative expression for $\hat{\beta}$ is

$$\hat{\beta} = \frac{\sum_{t=1}^{T} x_t y_t - T\bar{x}\bar{y}}{\sum_{t=1}^{T} x_t^2 - T\bar{x}^2}$$

- To summarize the OLS estimates for α and β are given by

$$\hat{\beta} = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2} = \frac{\sum_{t=1}^{T} x_t y_t - T\bar{x}\bar{y}}{\sum_{t=1}^{T} x_t^2 - T\bar{x}^2}$$

and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Example continued

- For the data at hand, we have $\bar{y} = 14.83$, $\bar{x} = 8.59$, $\sum_{t=1}^{10} x_t y_t = 3410.62$, $\sum_{t=1}^{10} x_t^2 = 2066.37$, and so

$$\hat{\beta} = \frac{3410.62 - 10 \cdot 8.59 \cdot 14.83}{2066.37 - 10 \cdot 8.59^2} = \frac{2136.72}{1328.49} = 1.61$$

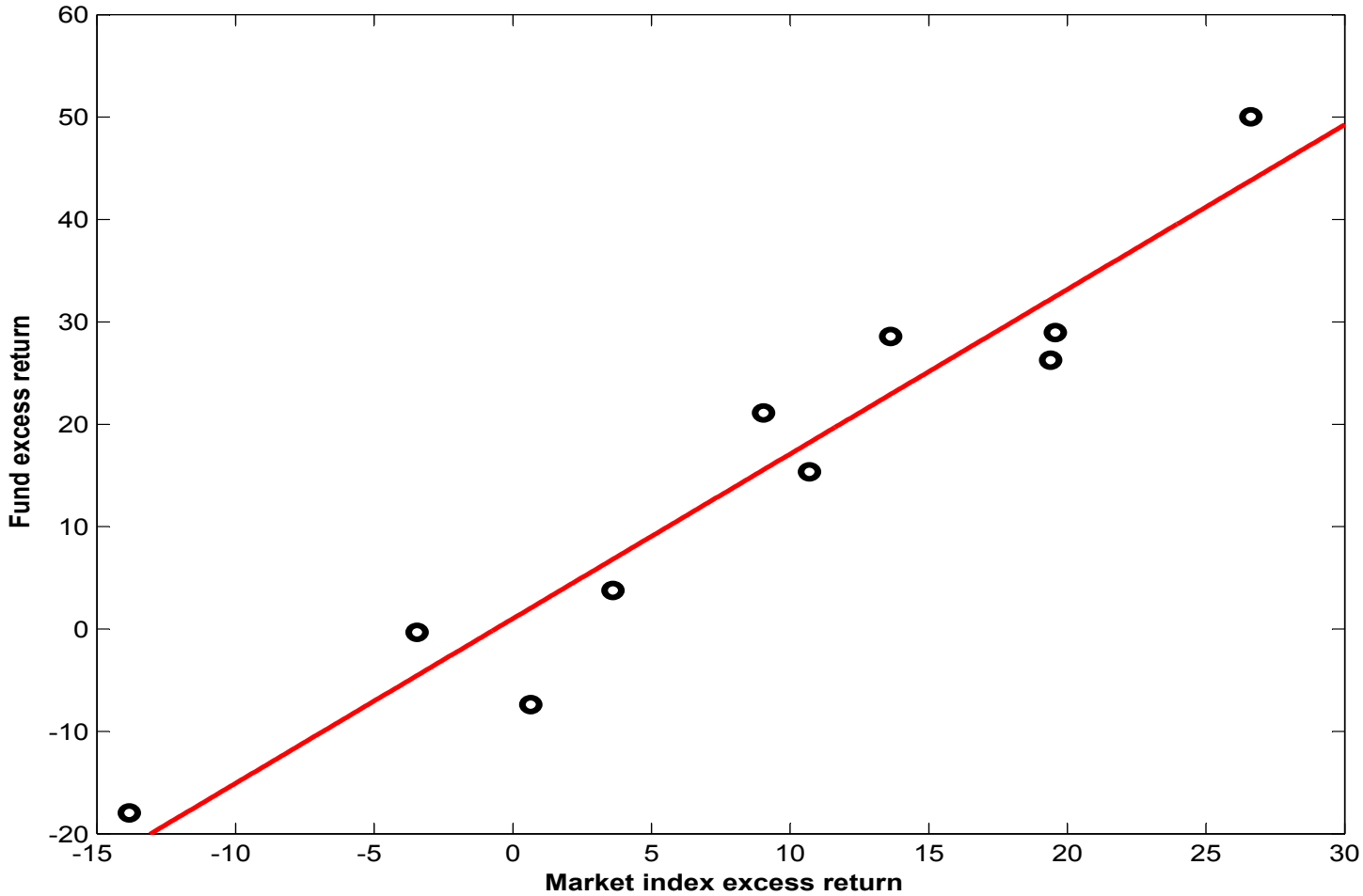
and

$$\hat{\alpha} = 14.83 - 1.61 \cdot 8.59 = 1.01$$

- The estimated fitted line is

$$\hat{y}_t = 1.01 + 1.61 x_t$$

Example continued: fitted regression line



Example continued

- Question: If an analyst tells that she expects the market to yield an excess return of 15% next year, what would you expect the return on fund XYZ to be?
- Answer: We can say that the expected value of y is $\hat{\alpha} + \hat{\beta}x$ and plug $x = 15\%$ into the equation to obtain the expected value for y:

$$\hat{y} = 1.01 + 1.61 \cdot 15 = 25.16\%$$

Population and Sample Regression Functions

- The population regression function (PRF) is a description of the model that is thought to generate the actual data and the true relationship between the variables y and x.
- In the present context the PRF is determined and characterized by the true values of α and β and is given by

$$y_t = \alpha + \beta x_t + u_t$$

- The sample regression function (SRF) is given by

$$\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t$$

- We use the SRF to make inferences about the PRF
- We would also like to know how "good" the estimates of α and β are.

Linearity

- One crucial assumption in the preceding analysis is the linearity of the regression function in the parameters α and β. Note that the model does not have to be linear in the variables y and x
- Certain models can be transformed to linear ones by suitable manipulation
- For example, the exponential regression model

$$Y_t = e^{\alpha} X_t^{\beta} e^{u_t} \Leftrightarrow \log(Y_t) = \alpha + \beta \log(X_t) + u_t$$

Letting $y_t = \log(Y_t)$ and $x_t = \log(X_t)$ we can write the model in the usual linear form

$$y_t = \alpha + \beta x_t + u_t$$

Linear and Nonlinear Models

- In the exponential regression model above, the coefficients can be interpreted as elasticities.
- Similarly, if theory suggests that y and x should be inversely related, the regression model could be

$$y_t = \alpha + \frac{\beta}{x_t} + u_t$$

  which can be estimated by OLS by substituting $z_t = \frac{1}{x_t}$
- However, some models are intrinsically nonlinear and cannot be transformed into linear ones, e.g.,

$$y_t = \alpha + x_t^{\beta} + u_t$$

Assumptions Underlying the Linear Regression Model

- Note that no assumptions about the data were explicitly made so far.
- In particular, the regressors x can be either deterministic (fixed) or stochastic.
- However, one assumption is required for the OLS estimators to be well-defined: the denominator $\sum_{t=1}^{T}(x_t - \bar{x})^2$ in the formula for $\hat{\beta}$ should be positive.
- This condition is equivalent to assuming that no all x's are equal. In other words, we need some variability in x for the OLS estimators to be well-defined.
- How would the scatter plot look like if all the x's were equal?

Assumptions Underlying the Linear Regression Model

- To further study the parameter estimators $\hat{\alpha}$ and $\hat{\beta}$, we need to make additional assumptions about the data generating process.
- In particular, we need to make assumptions about how the random disturbances u_t are generated.
- First, we assume that the regressors x_t are non-stochastic (fixed). We will relax this assumption in the sequel.
- The following assumptions about the shocks u_t are made:
  - A1. The shocks u_t have zero mean: $\mathbb{E}[u_t] = 0$.
  - A2. The shocks u_t are homoscedastic: $\mathbb{V}[u_t] = \sigma^2 < \infty$.
  - A3. The shocks u_t are uncorrelated: $\mathbb{C}[u_t, u_s] = 0$, for $t \neq s$.
  - A4. The shocks u_t are (jointly) normally distributed: $u_t \sim N(0, \sigma^2)$.

Linearity of the OLS estimators

- The OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are linear in the y's. Specifically,

$$\hat{\beta} = \sum_{t=1}^{T} w_t y_t$$

where

$$w_t = \frac{x_t - \bar{x}}{\sum_{t=1}^{T}(x_t - \bar{x})^2}.$$

Moreover,

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{1}{T}\sum_{t=1}^{T} y_t - \bar{x}\sum_{t=1}^{T} w_t y_t = \sum_{t=1}^{T} q_t y_t$$

where

$$q_t = \frac{1}{T} - \bar{x} w_t.$$

Unbiasedness of the OLS estimators

- Under assumption A.1, the OLS estimators are unbiased, i.e., $\mathbb{E}[\hat{\alpha}] = \alpha$ and $\mathbb{E}[\hat{\beta}] = \beta$.
- Note that $\sum_{t=1}^{T} w_t = 0$ and $\sum_{t=1}^{T} w_t x_t = 1$.
- $\hat{\beta}$ is an unbiased estimator of β

$$
\begin{aligned}
& \hat{\beta} = \sum_{t=1}^{T} w_t y_t = \sum_{t=1}^{T} w_t(\alpha + \beta x_t + u_t) \\
\Rightarrow \quad & \mathbb{E}[\hat{\beta}] = \sum_{t=1}^{T} w_t(\alpha + \beta x_t) \\
\Rightarrow \quad & \mathbb{E}[\hat{\beta}] = \alpha \sum_{t=1}^{T} w_t + \beta \sum_{t=1}^{T} w_t x_t \\
\Rightarrow \quad & \mathbb{E}[\hat{\beta}] = \alpha \cdot 0 + \beta \cdot 1 = \beta.
\end{aligned}
$$

Unbiasedness of the OLS estimators (cont'd)

- Note that $\bar{y} = \alpha + \beta\bar{x} + \bar{u}$ and so $\mathbb{E}[\bar{y}] = \alpha + \beta\bar{x}$.
- $\hat{\alpha}$ is an unbiased estimator of α

$$\begin{aligned} & \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \\ \Rightarrow \quad & \mathbb{E}[\hat{\alpha}] = \mathbb{E}[\bar{y}] - \mathbb{E}[\hat{\beta}]\bar{x} \\ \Rightarrow \quad & \mathbb{E}[\hat{\alpha}] = (\alpha + \beta\bar{x}) - \beta\bar{x} = \alpha. \end{aligned}$$

Variance of the OLS estimators

- Under assumptions A.1, A.2, and A.3 the variances of $\hat{\alpha}$ and $\hat{\beta}$ are $\mathbb{V}[\hat{\alpha}] = \left(\sum_{t=1}^{T} q_t^2\right)\sigma^2$ and $\mathbb{V}[\hat{\beta}] = \left(\sum_{t=1}^{T} w_t^2\right)\sigma^2$.
- Since $w_t = \frac{x_t - \bar{x}}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$, we have

$$\sum_{t=1}^{T} w_t^2 = \frac{1}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

- Since $q_t = \frac{1}{T} - \bar{x}w_t$, we have $q_t^2 = \frac{1}{T^2} + \bar{x}^2 w_t^2 - 2\frac{1}{T}\bar{x}w_t$. Since $\sum_{t=1}^{T} w_t = 0$ we have

$$\begin{aligned}\sum_{t=1}^{T} q_t^2 &= \frac{1}{T} + \bar{x}^2 \frac{1}{\sum_{t=1}^{T}(x_t - \bar{x})^2} \\ &= \frac{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2 + \bar{x}^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2} = \frac{\frac{1}{T}\sum_{t=1}^{T} x_t^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}\end{aligned}$$

The OLS estimators are BLUE

- The OLS estimator $\hat{\beta}$ ($\hat{\alpha}$) has the smallest variance among the linear unbiased estimators of β (α). The OLS estimators are Best Linear Unbiased Estimators (BLUE).

- Specifically, if $\tilde{\beta} = \sum_{t=1}^{T} \phi_t y_t$ is an unbiased estimator of β, then $\mathbb{V}[\tilde{\beta}] \geq \mathbb{V}[\hat{\beta}]$ and $\tilde{\alpha} = \sum_{t=1}^{T} \lambda_t y_t$ is an unbiased estimator of α, then $\mathbb{V}[\tilde{\alpha}] \geq \mathbb{V}[\hat{\alpha}]$.

The OLS estimators are BLUE (cont'd)

- To see why this is the case, note that $\sum_{t=1}^{T} \phi_t = 0$, $\sum_{t=1}^{T} \phi_t x_t = 1$, $\sum_{t=1}^{T} \lambda_t = 1$, and $\sum_{t=1}^{T} \lambda_t x_t = 0$. Hence, if $\delta_t = \phi_t - w_t$ and $\zeta_t = \lambda_t - w_t$, we have $\sum_{t=1}^{T} \delta_t = \sum_{t=1}^{T} \delta_t x_t = 0$ and $\sum_{t=1}^{T} \zeta_t = \sum_{t=1}^{T} \zeta_t x_t = 0$. It follows that $\sum_{t=1}^{T} \delta_t w_t = 0$ and so $\sum_{t=1}^{T} \phi_t^2 = \sum_{t=1}^{T} w_t^2 + \sum_{t=1}^{T} \delta_t^2 \geq \sum_{t=1}^{T} w_t^2$ implying $\mathbb{V}[\tilde{\beta}] = \left(\sum_{t=1}^{T} \phi_t^2\right) \sigma^2 \geq \left(\sum_{t=1}^{T} w_t^2\right) \sigma^2 = \mathbb{V}[\hat{\beta}]$. Similarly, $\sum_{t=1}^{T} \zeta_t w_t = 0$ and so $\mathbb{V}[\tilde{\alpha}] = \left(\sum_{t=1}^{T} \lambda_t^2\right) \sigma^2 \geq \left(\sum_{t=1}^{T} q_t^2\right) \sigma^2 = \mathbb{V}[\hat{\alpha}]$.

Precision of the OLS estimators and Standard Errors

- A measure of reliability or precision is desirable for the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$.
- For instance, we would like to be able to answer questions like (i) Is α different from zero?, and (ii) Is β different from 1?, with certain degree of confidence.
- The standard measure of precision is the variance or equivalently the standard deviation (the square root of the variance).
- The variance (standard deviation) is unknown and has to be estimated. An estimate of the standard deviation is called the *standard error*.

Precision of the OLS estimators and Standard Errors (cont'd)

Recall that the variances of the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are

$$\mathbb{V}[\hat{\alpha}] = \frac{\frac{1}{T}\sum_{t=1}^{T} x_t^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2} \cdot \sigma^2, \quad \mathbb{V}[\hat{\beta}] = \frac{1}{\sum_{t=1}^{T}(x_t - \bar{x})^2} \cdot \sigma^2$$

- Hence, to obtain estimates of $\mathbb{V}[\hat{\alpha}]$ and $\mathbb{V}[\hat{\beta}]$, we need an estimate of $\sigma^2 = \mathbb{E}[u_t^2] = \mathbb{V}[u_t]$.
- We could estimate σ^2 by the average $\frac{1}{T}\sum_{t=1}^{T} u_t^2$, but this is not feasible since the (true) disturbances are not observable.

Precision of the OLS estimators and Standard Errors (cont'd)

- Instead, we can use the residuals $\hat{u}_t$, to obtain the estimator $\frac{1}{T}\sum_{t=1}^{T}\hat{u}_t^2$. However, it turns out that this estimator is biased.
- An unbiased estimator of σ^2 is

$$s^2 = \frac{1}{T-2}\sum_{t=1}^{T}\hat{u}_t^2.$$

- Hence, the standard errors of $\hat{\alpha}$ and $\hat{\beta}$ are

$$\mathrm{SE}[\hat{\alpha}] = s \cdot \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}x_t^2}}{\sqrt{\sum_{t=1}^{T}(x_t-\bar{x})^2}}, \quad \mathrm{SE}[\hat{\beta}] = s \cdot \frac{1}{\sqrt{\sum_{t=1}^{T}(x_t-\bar{x})^2}}.$$

Some Comments on Standard Errors

- Both $\mathrm{SE}[\hat{\alpha}]$ and $\mathrm{SE}[\hat{\beta}]$ depend on s and they do so linearly. The lower the shock variance estimate s, the more precise the OLS estimators.
- The quantity $\sum_{t=1}^{T}(x_t - \bar{x})^2$ appears in the denominator in the expression for both $\mathrm{SE}[\hat{\alpha}]$ and $\mathrm{SE}[\hat{\beta}]$. The higher the dispersion of the x variable, the more precise the OLS estimators. This is intuitive, since this quantity appears in the denominator of the estimator $\hat{\beta}$. For small values of $\sum_{t=1}^{T}(x_t - \bar{x})^2$ the estimator $\hat{\beta}$ becomes unstable, and the same holds for $\hat{\alpha}$ since $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$. Recall that in the extreme case in which this quantity is zero, the OLS estimators are not even well-defined.

Some Comments on Standard Errors (cont'd)

- As the sample size T increases, the standard errors become smaller and the OLS estimators become more precise. To see this, observe that

$$\mathbb{V}[\hat{\alpha}] = \frac{1}{T} \cdot \sigma^2 \cdot \frac{\frac{1}{T}\sum_{t=1}^{T} x_t^2}{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

$$\mathbb{V}[\hat{\beta}] = \frac{1}{T} \cdot \sigma^2 \cdot \frac{1}{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

and that $\frac{1}{T}\sum_{t=1}^{T} x_t^2$, $\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2$ converges to the x population second moment and variance, respectively, as $T \to \infty$.

Some Comments on Standard Errors (cont'd)

- Both $\mathrm{SE}[\hat{\alpha}]$ and $\mathrm{SE}[\hat{\beta}]$ depend on s and they do so linearly. The lower the shock variance estimate s, the more precise the OLS estimators.
- The quantity $\sum_{t=1}^{T} x_t^2$ appears in the numerator of the expression for $\mathrm{SE}[\hat{\alpha}]$. The larger $\sum_{t=1}^{T} x_t^2$ is, the further away the data are from the y-axis, and the harder it is to estimate the intercept α.

Linear regression model: Distribution results

BUFN 758N

Prof. Skoulakis

Distribution of the OLS estimators

Under assumptions A.1, A.2, A.3, and A.4, the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are normally distributed. This can be seen as follows.

Under assumption A.4, the u_t's are jointly normally distributed and, since the x_t's are assumed to be fixed, the y_t's are also jointly normally distributed.

Moreover, both $\hat{\alpha}$ and $\hat{\beta}$ are linear combinations of the y_t's: $\hat{\alpha} = \sum_{t=1}^{T} q_t y_t$ and $\hat{\beta} = \sum_{t=1}^{T} w_t y_t$. It follows that both $\hat{\alpha}$ and $\hat{\beta}$ are normally distributed. Hence,

$$\hat{\alpha} \sim N(\alpha, \mathbb{V}[\hat{\alpha}]), \quad \hat{\beta} \sim N(\beta, \mathbb{V}[\hat{\beta}]).$$

Distribution of the OLS estimators (cont'd)

The so-called standardized versions of $\hat{\alpha}$ and $\hat{\beta}$ follow standard normal distributions

$$\frac{\hat{\alpha} - \alpha}{\sqrt{\mathbb{V}[\hat{\alpha}]}} \sim N(0,1), \quad \frac{\hat{\beta} - \beta}{\sqrt{\mathbb{V}[\hat{\beta}]}} \sim N(0,1).$$

Distribution of the OLS estimators (cont'd)

However, the above formulas are not operational since the variances $\mathbb{V}[\hat{\alpha}]$ and $\mathbb{V}[\hat{\beta}]$ are unknown since σ^2 is unknown. In practice, they are estimated using $s^2 = \frac{1}{T-2}\sum_{t=1}^{T}\hat{u}_t^2$ as estimate of σ^2. The resulting standardized versions follow a t distribution with $T-2$ degrees of freedom:

$$\frac{\hat{\alpha}-\alpha}{\mathrm{SE}[\hat{\alpha}]} \sim t_{T-2}, \quad \frac{\hat{\beta}-\beta}{\mathrm{SE}[\hat{\beta}]} \sim t_{T-2}.$$

The χ^2 and t distributions

The χ^2 distribution with n degrees of freedom (where n is a positive integer) ia defined as follows. Let $Z_1, \ldots, Z_n$ be n *independent* $N(0,1)$ random variables. Then, $W = Z_1^2 + \cdots + Z_n^2$ follows a χ^2 distribution with n degrees of freedom. Notation: $W \sim \chi_n^2$. Generalization: one can define the χ^2 distribution with ν degrees of freedom where ν is any positive real number.

The t distribution is defined as follows. Let Z be a $N(0,1)$ random variable and W be a χ^2 random variable with ν degrees of freedom such that Z and W are independent. Then, the ratio $\frac{Z}{\sqrt{W/\nu}}$ follows a t distribution with ν degrees of freedom.

Standardized OLS estimators follow a t distribution

Recall that $\mathrm{SE}[\hat{\alpha}] = \sqrt{\frac{s^2}{\sigma^2} \cdot \mathbb{V}[\hat{\alpha}]}$ and so

$$\frac{\hat{\alpha} - \alpha}{\mathrm{SE}[\hat{\alpha}]} = \frac{\frac{\hat{\alpha}-\alpha}{\sqrt{\mathbb{V}[\hat{\alpha}]}}}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{Z}{\sqrt{W/(T-2)}} \sim t_{T-2}$$

where $Z = \frac{\hat{\alpha}-\alpha}{\sqrt{\mathbb{V}[\hat{\alpha}]}} \sim N(0,1)$ and

$W = (T-2)\frac{s^2}{\sigma^2} = \frac{\sum_{t=1}^{T} \hat{u}_t^2}{\sigma^2} \sim \chi^2(T-2)$.

Stochastic regressors x_t

- The previous results were derived under the assumption that the regressors x_t are fixed (non-stochastic).
- This assumption, however, is very restrictive for most empirical applications.
- How do the results change if the regressors are allowed to be stochastic?
- The OLS estimators remain unbiased under a suitable assumption (independence between the shocks and the regressors)
- More can be said about the *asymptotic* behavior of the OLS estimators, i.e., when the sample size is large.

Unbiasedness of the OLS estimators with stochastic regressors x_t

- Under the assumptions (i) $\mathbb{E}[u_t] = 0$ and (ii) u_t is independent from the regressors $x_1, x_2, \ldots, x_T$, the OLS estimators are unbiased.

- Recall that $\hat{\alpha} = \sum_{t=1}^{T} q_t y_t = \alpha + \sum_{t=1}^{T} q_t u_t$ and $\hat{\beta} = \sum_{t=1}^{T} w_t y_t = \beta + \sum_{t=1}^{T} w_t u_t$. Moreover the weights q_t and w_t are functions of the regressors $x_1, x_2, \ldots, x_T$, which implies that $\mathbb{E}[q_t u_t] = \mathbb{E}[q_t]\,\mathbb{E}[u_t] = 0$ and so $\sum_{t=1}^{T} \mathbb{E}[q_t u_t] = 0$ which, in turn, implies $\mathbb{E}[\hat{\alpha}] = \alpha$. Similarly, $\mathbb{E}[\hat{\beta}] = \beta$.

Some background on asymptotics: Convergence of random variables

- We need to define two notions of convergence of random variables: *in probability* and *in distribution*.
- **Convergence in probability**: A sequence $\{X_n : n = 1, 2, \ldots\}$ of random variables is said to converge to a random variable X in probability if, for any $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left[|X_n - X| \geq \varepsilon\right] = 0$. Notation: $X_n \xrightarrow{p} X$, as $n \to \infty$.

Some background on asymptotics: Convergence of random variables

- **Convergence in distribution**: Let $\{X_n : n = 1, 2, \ldots\}$ be a sequence of random variables and let F_n denote the cumulative distribution function (CDF) of X_n, i.e., $F_n(x) = \mathbb{P}[X_n \leq x]$. Further let X be a random variable with CDF denoted by F, i.e., $F(x) = \mathbb{P}[X \leq x]$. The sequence $\{X_n\}$ is said to converge to X in distribution if $\lim_{n\to\infty} F_n(x) = F(x)$ for all x at which F is continuous. Notation: $X_n \xrightarrow{d} X$, as $n \to \infty$.

Convergence of random variables: some useful properties

- If $X_n \xrightarrow{p} X$ then $X_n \xrightarrow{d} X$.
- If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$ then
  - $aX_n + bY_n \xrightarrow{p} aX + bY$, where a and b are constants.
  - $X_n Y_n \xrightarrow{p} XY$.
  - $X_n / Y_n \xrightarrow{p} X/Y$ under the assumption $Y \neq 0$.
- If $X_n \xrightarrow{d} X$ and $Y_n - X_n \xrightarrow{p} 0$ then $Y_n \xrightarrow{d} X$.
- If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a constant, then
  - $X_n \pm Y_n \xrightarrow{d} X \pm c$.
  - $X_n Y_n \xrightarrow{d} cX$.
  - If $c = 0$, i.e., $Y_n \xrightarrow{p} 0$, then $X_n Y_n \xrightarrow{p} 0$.

Convergence of estimators

- The above two notions of convergence are useful for describing the large-sample behavior of estimators: consistency and asymptotic normality.
- **Consistency**. Let $\hat{\theta}_T$ be an estimator of a parameter θ, based on a sample of size T. The estimator $\hat{\theta}_T$ is called consistent if it converges to θ in probability, i.e., $\hat{\theta}_T \xrightarrow{p} \theta$, as $T \to \infty$. The property of consistency is important as it guarantees that the estimator gets closer to the parameter of interest as the sample size increases.

Convergence of estimators

- **Asymptotic normality**. Under suitable assumptions, many estimators (or standardized versions of them) turn out to converge to a normal random variable in distribution, i.e., they are asymptotically normal. The property of asymptotic normality is important because it allows us to make statements about the distribution of estimators, when the sample size is large. It is used extensively in hypothesis testing.

Some background on asymptotics: LLN and CLT

- There are two fundamental results in asymptotic theory: the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT).
- **Law of Large Numbers**: Let $\{X_n : n = 1, 2, \ldots\}$ be a sequence of independent and identically distributed (i.i.d.) random variables with finite mean μ. Then the sample average $\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$ converges to μ in probability, i.e., $\bar{X}_n \xrightarrow{p} \mu$.
- **Central Limit Theorem**: Let $\{X_n : n = 1, 2, \ldots\}$ be a sequence of i.i.d. random variables with finite mean μ and finite variance σ^2. Then the standardized sample average $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ converges to a standard normal variable in distribution, i.e., $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1)$.

Consistency and asymptotic normality of the OLS estimators

- Assume that (i) the sequence (x_t, u_t) is i.i.d. over time with finite mean and variance (ii) the shock u_t has zero mean, i.e., $\mathbb{E}[u_t] = 0$ and (iii) the regressor x_t and the shock u_t are uncorrelated, i.e., $\mathbb{E}[x_t u_t] = 0$.
- Under these assumptions, the OLS estimators are consistent and asymptotically normal.

Consistency of the OLS estimators

- Recall that $\hat{\beta} = \beta + \sum_{t=1}^{T} w_t u_t$, where $w_t = \frac{x_t - \bar{x}}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$. The estimator $\hat{\beta}$ is consistent if $\sum_{t=1}^{T} w_t u_t \xrightarrow{p} 0$. But

$$\sum_{t=1}^{T} w_t u_t = \frac{\frac{1}{T}\sum_{t=1}^{T} x_t u_t - \bar{x}\bar{u}}{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2} \xrightarrow{p} \frac{\mathbb{E}[x_t u_t] - \mathbb{E}[x_t]\mathbb{E}[u_t]}{\mathbb{V}[x_t]} = 0.$$

- Since $y_t = \alpha + \beta x_t + u_t$, we have $\mathbb{E}[y_t] = \alpha + \beta \mathbb{E}[x_t]$. By the LLN, $\bar{y} \xrightarrow{p} \mathbb{E}[y_t]$ and $\bar{x} \xrightarrow{p} \mathbb{E}[x_t]$. Hence, $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \xrightarrow{p} \mathbb{E}[y_t] - \beta \mathbb{E}[x_t] = \alpha$, i.e., $\hat{\alpha}$ is a consistent estimator of α.

Asymptotic normality of $\hat{\beta}$

The slope estimator $\hat{\beta}$ is asymptotically normal with mean β and variance $\frac{1}{T} \cdot \frac{\mathbb{E}[(x_t - \mathbb{E}[x_t])^2 u_t^2]}{(\mathbb{V}[x_t])^2}$, i.e.,

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \frac{\mathbb{E}[(x_t - \mathbb{E}[x_t])^2 u_t^2]}{(\mathbb{V}[x_t])^2}\right).$$

If x_t and u_t are independent, then

$$\mathbb{E}[(x_t - \mathbb{E}[x_t])^2 u_t^2] = \mathbb{E}[(x_t - \mathbb{E}[x_t])^2] \cdot \mathbb{E}[u_t^2] = \sigma^2 \cdot \mathbb{V}[x_t],$$

In this case, the asymptotic variance of $\hat{\beta}$ simplifies to $\frac{1}{T} \cdot \frac{1}{\mathbb{V}[x_t]} \cdot \sigma^2$ (compare this to the formula for the variance of $\hat{\beta}$ in the case on non-stochastic regressors).

Asymptotic normality of $\hat{\alpha}$

The intercept estimator $\hat{\alpha}$ is asymptotically normal with mean α and variance $\frac{1}{T}\frac{\mathbb{E}[(x_t-\mathbb{E}[x_t])^2 u_t^2]}{(\mathbb{V}[x_t])^2}$, i.e.,

$$\sqrt{T}(\hat{\alpha}-\alpha) \xrightarrow{d} N\left(0, \frac{\mathbb{E}[(\mathbb{E}[x_t^2]-\mathbb{E}[x_t]\cdot x_t)^2 \cdot u_t^2]}{(\mathbb{V}[x_t])^2}\right).$$

If x_t and u_t are independent, then

$$\mathbb{E}[(\mathbb{E}[x_t^2]-\mathbb{E}[x_t]\cdot x_t)^2 \cdot u_t^2] = \sigma^2 \cdot \mathbb{E}[x_t^2] \cdot \mathbb{V}[x_t].$$

In this case, the asymptotic variance of $\hat{\alpha}$ simplifies to $\frac{1}{T} \cdot \frac{\mathbb{E}[x_t^2]}{\mathbb{V}[x_t]} \cdot \sigma^2$ (compare this to the formula for the variance of $\hat{\alpha}$ in the case on non-stochastic regressors).

Asymptotic standard errors of the OLS estimators

Under the independence assumption, the asymptotic variances of $\hat{\alpha}$ and $\hat{\beta}$ are given by $\frac{1}{T} \cdot \frac{\mathbb{E}[x_t^2]}{\mathbb{V}[x_t]} \cdot \sigma^2$ and $\frac{1}{T} \cdot \frac{1}{\mathbb{V}[x_t]} \cdot \sigma^2$, respectively. In practice, they are estimated by their sample analogues. The resulting standard errors are

$$\mathrm{SE}[\hat{\alpha}] = s \cdot \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T} x_t^2}}{\sqrt{\sum_{t=1}^{T}(x_t - \bar{x})^2}}, \quad \mathrm{SE}[\hat{\beta}] = s \cdot \frac{1}{\sqrt{\sum_{t=1}^{T}(x_t - \bar{x})^2}}.$$

where now the (consistent) estimate s^2 of the disturbance variance σ^2 is given by $s^2 = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_t^2$.
Note that the standard error formulas are the same as in the case of non-stochastic regressors. However, here they are justified in asymptotic terms.

Asymptotic normality of the standardized OLS estimators

- It follows that under the assumption of independence between x_t and u_t we have

$$\frac{\hat{\alpha}-\alpha}{\mathrm{SE}[\hat{\alpha}]} \xrightarrow{d} N(0,1), \quad \frac{\hat{\beta}-\beta}{\mathrm{SE}[\hat{\beta}]} \xrightarrow{d} N(0,1),$$

where $\mathrm{SE}[\hat{\alpha}] = s \cdot \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T} x_t^2}}{\sqrt{\sum_{t=1}^{T}(x_t-\bar{x})^2}}$ and $\mathrm{SE}[\hat{\beta}] = s \cdot \frac{1}{\sqrt{\sum_{t=1}^{T}(x_t-\bar{x})^2}}$.

- Recall that a similar exact result (involving the t distribution) holds for the case of non-stochastic regressors and normally distributed shocks.

Linear regression model: Inference

BUFN 758N

Prof. Skoulakis

Statistical Inference

- The goal is to make inferences about the true (but unknown) regression parameter values. We achieve this goal by using the results on the distribution of the OLS parameter estimates developed so far.
- Example: Consider running a CAPM regression on monthly data. Let $r_{Q,t}$ be the excess return on stock Q and $r_{M,t}$ be the excess return on the S&P 500 index (market proxy). The regression is

$$r_{Q,t} = \alpha_Q + \beta_Q r_{M,t} + u_t$$

- Questions of interest: (i) Is $\alpha_Q = 0$? (ii) Is $\beta_Q > 1$?
- Suppose that we obtain the following results: $\hat{\alpha}_Q = 0.5$ with $\mathrm{SE}[\hat{\alpha}_Q] = 0.3$ and $\hat{\beta}_Q = 1.3$ and $\mathrm{SE}[\hat{\beta}_Q] = 0.6$. How do we use these results to answer the above questions?

Hypothesis Testing: Some Concepts

- A question such as "Is $\alpha_Q = 0$?" is formally referred to as a hypothesis and the procedure used to answer the question is referred to as hypothesis testing.
- There will be two hypotheses: the null hypothesis (denoted by H_0) and the alternative hypothesis (denoted by H_1)
- The null hypothesis is the question or statement of primary interest. The alternative hypothesis represents the remaining outcomes of interest.
- Suppose that, in the previous example, we wish to test the hypothesis that α_Q is equal to 0. Then

$$H_0 : \alpha_Q = 0$$
$$H_1 : \alpha_Q \neq 0$$

This type of test is referred to as a two-sided test.

Hypothesis Testing: Some Concepts

- Occasionally, we may be interested in (or may expect) the parameter of interest to be less than (or greater than) the benchmark value of primary interest as opposed to just different. For example, suppose that, in the example above, we wish to test $\beta_Q = 1$ but have some prior information that β_Q might be larger than 1. In such a situation, we consider the one-sided tests.

$$H_0 : \beta_Q = 1$$
$$H_1 : \beta_Q > 1$$

- Another one-sided hypothesis test would be

$$H_0 : \beta_Q = 1$$
$$H_1 : \beta_Q < 1$$

- There are two standard ways to conduct a hypothesis test: (i) *the test of significance* approach and (ii) *the confidence interval* approach.

The Test of Significance Approach

- Suppose that, in the above example, we wish to test the null hypothesis $H_0 : \alpha_Q = 0$ against the alternative hypothesis $H_1 : \alpha_Q \neq 0$.
- In words, we wish to determine whether the alpha associated with stock Q is zero.
- It is important to understand that such a question cannot be answered with absolute confidence. It can only be answered in probabilistic terms, i.e., with certain degree of confidence.
- Our best guess about the true, but unknown, α_Q is the OLS estimate $\hat{\alpha}_Q$. Hence, to figure out whether α_Q is zero, it makes sense to examine whether $\hat{\alpha}_Q$ is close to zero and reject the null hypothesis $H_0 : \alpha_Q = 0$ if $\hat{\alpha}_Q$ is "too large" (in absolute value).

The Test of Significance Approach

- How large is "too large"? This should be evaluated in likelihood terms and depends on the distribution of the estimator $\hat{\alpha}_Q$.
- Following the above logic, we would reject $H_0 : \alpha_Q = 0$ if the observed (or realized) value of $\hat{\alpha}_Q$ lies in the tails of the distribution of $\hat{\alpha}_Q$, i.e., beyond certain thresholds.
- It is more convenient to work with the standardized version of $\hat{\alpha}_Q$, i.e., the so-called *test statistic*

$$\frac{\hat{\alpha} - \alpha^*}{\mathrm{SE}[\hat{\alpha}]}.$$

where α^* is the parameter value of interest under H_0, i.e., $\alpha^* = 0$.

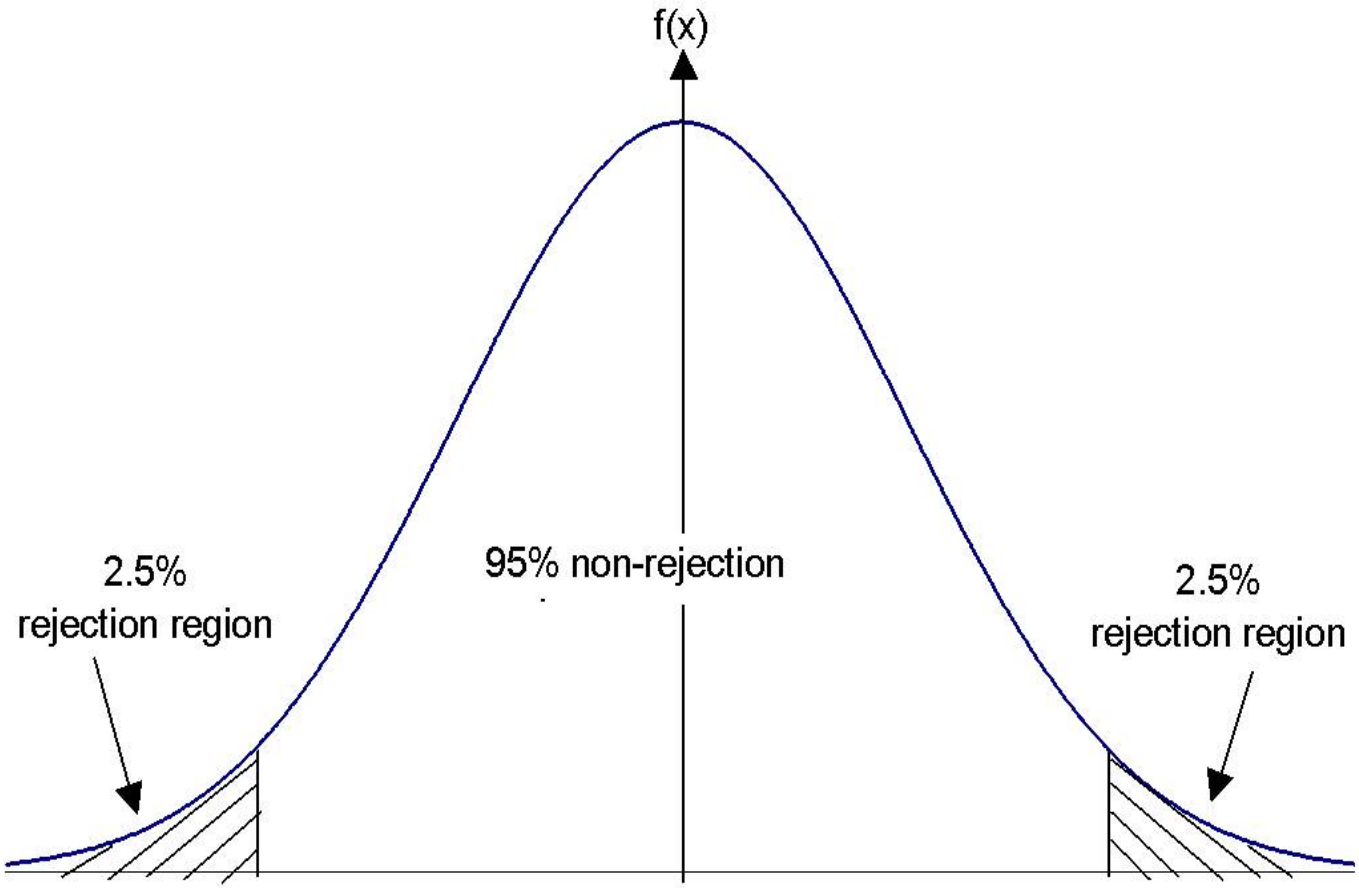
The Test of Significance Approach

- We would reject $H_0 : \alpha_Q = 0$ if the realized value of the test statistic is extreme (low or high).
- The test statistic follows a t distribution if the regressors are non-stochastic and the shock are normally distributed. If the regressors are stochastic, the test statistic is asymptotically normal under appropriate conditions.
- Hence, by using the standardized version (test statistic) we only need to focus on two distributions.

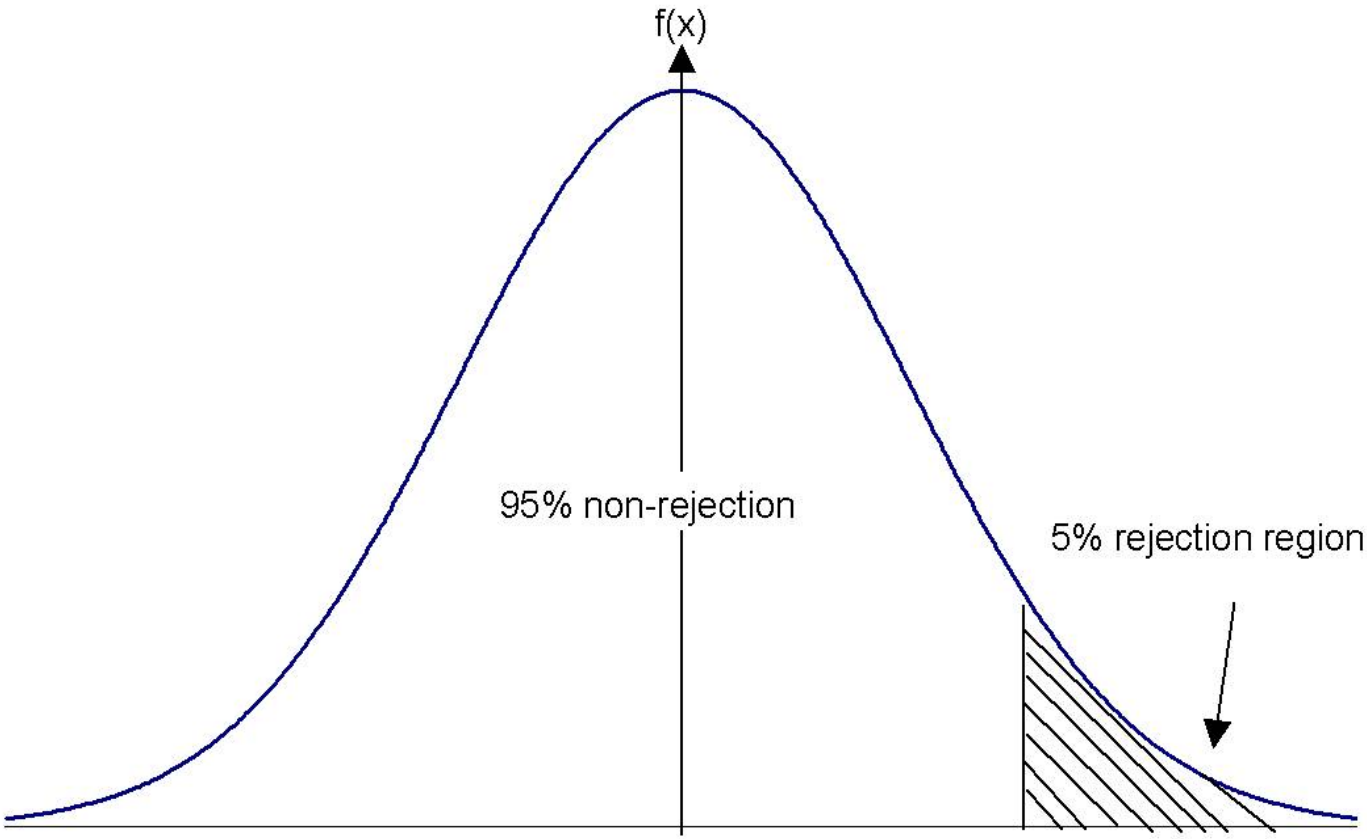
The Test of Significance Approach

- How do we determine the thresholds? They are determined by making sure that, if the null hypothesis is true, we reject with small probability. This probability is called the *level of significance* or the *size of the test*, and is denoted by α. It is typically set equal to 5% but the values 10% and 1% are also used.
- The thresholds define a rejection region (associated with probability 5%) and a non-rejection region (associated with probability 95%).

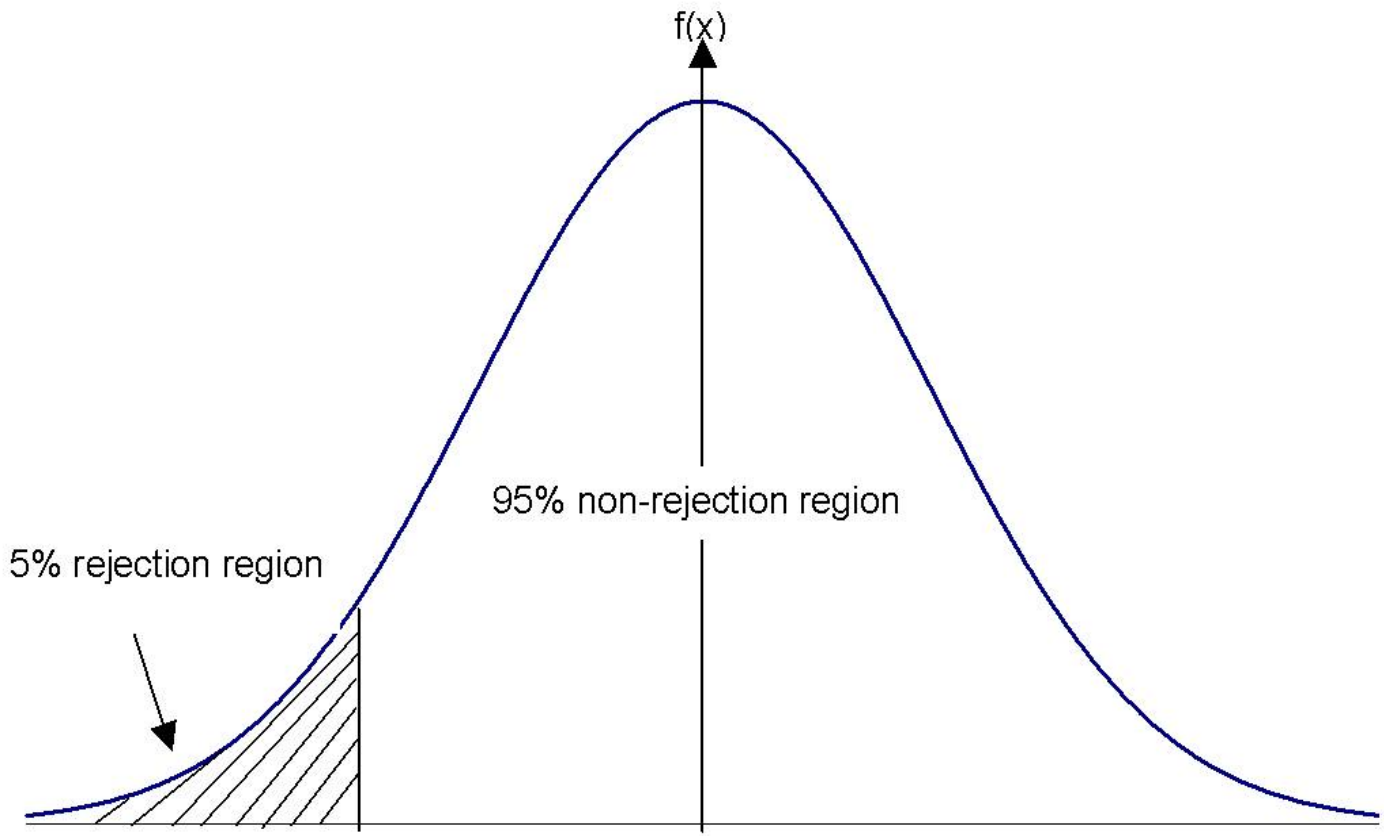
Rejection region for two-sided test



Rejection region for one-sided test (upper tail)



Rejection region for one-sided test (lower tail)



Rejection Regions

- Suppose that the sample size T is large and the hypothesis testing will be based on the normal distribution. Denote by z_p the $100(1-p)\%$ percentile of the standard normal distribution, i.e., $\mathbb{P}[Z \geq z_p] = p$ where $Z \sim N(0,1)$.
- For testing $H_0 : \beta = \beta^*$ VS $H_1 : \beta \neq \beta^*$, the rejection region is $\frac{\hat{\beta}-\beta^*}{\mathrm{SE}[\hat{\beta}]} \leq -z_{\alpha/2}$ or $\frac{\hat{\beta}-\beta^*}{\mathrm{SE}[\hat{\beta}]} \geq z_{\alpha/2}$.
- For testing $H_0 : \beta = \beta^*$ VS $H_1 : \beta > \beta^*$, the rejection region is $\frac{\hat{\beta}-\beta^*}{\mathrm{SE}[\hat{\beta}]} \geq z_{\alpha}$.
- For testing $H_0 : \beta = \beta^*$ VS $H_1 : \beta < \beta^*$, the rejection region is $\frac{\hat{\beta}-\beta^*}{\mathrm{SE}[\hat{\beta}]} \leq -z_{\alpha}$.

The Test of Significance Approach: Drawing Conclusions

- Use the data to obtain parameter estimates and standard errors and compute the test statistic.
- Use the appropriate table (from a normal or a t distribution) to obtain the critical value (or values) with which to compare the test statistic.
- Finally perform the test. If the test statistic lies in the rejection region then reject the null hypothesis H_0; otherwise, do not reject H_0.

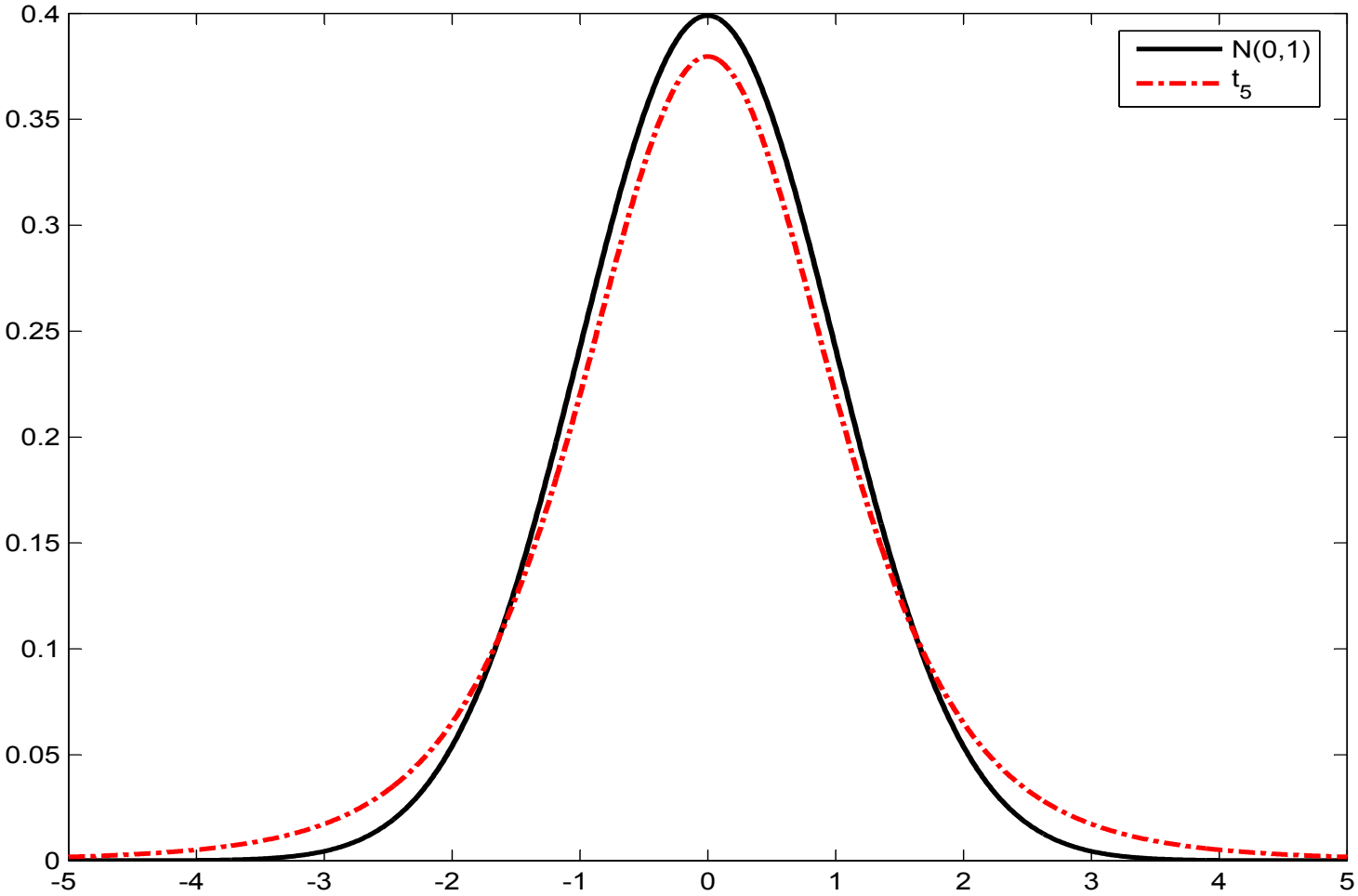
The Normal and the t distribution

- Both the standard normal and the t distribution are relevant for the hypothesis tests in the regression setting. Recall that test statistics typically follow one of the two distributions.
- Both distributions are symmetric but the t distribution has fatter tails, especially when the number of degrees of freedom ν is small.
- As ν increases, the t distribution approaches the normal distribution. In mathematical terms, $t_\nu \xrightarrow{d} N(0,1)$, as $\nu \to \infty$.

Comparing the t and the Normal distributions

Tail prob.	$\nu = 5$	$\nu = 10$	$\nu = 20$	$\nu = 30$	$\nu = 40$	$\nu = 50$	$N(0,1)$
0.1	1.48	1.37	1.33	1.31	1.30	1.30	1.28
0.05	2.02	1.81	1.72	1.70	1.68	1.68	1.64
0.025	2.57	2.23	2.09	2.04	2.02	2.01	1.96
0.01	3.36	2.76	2.53	2.46	2.42	2.40	2.33
0.005	4.03	3.17	2.85	2.75	2.70	2.68	2.58

Comparing the t and the Normal distributions



The Confidence Interval Approach

A confidence interval is a random interval that contains the parameter if interest with certain probability. Most commonly, confidence interval are two-sided but one-sided ones can also be constructed. Consider the case in which $\frac{\hat{\beta}-\beta}{\text{SE}[\hat{\beta}]} \sim t_{T-2}$ and let q be such that $\mathbb{P}[t_{T-2} \geq q] = 0.025$. Then, $\mathbb{P}\left[-q \leq \frac{\hat{\beta}-\beta}{\text{SE}[\hat{\beta}]} \leq q\right] = 0.95$ which is equivalent to

$$\mathbb{P}\left[\beta \in \left(\hat{\beta} - q \cdot \text{SE}[\hat{\beta}], \hat{\beta} + q \cdot \text{SE}[\hat{\beta}]\right)\right] = 0.95.$$

Hence, $\left(\hat{\beta} - q \cdot \text{SE}[\hat{\beta}], \hat{\beta} + q \cdot \text{SE}[\hat{\beta}]\right)$ is a 95% confidence interval.

The Confidence Interval Approach

Now suppose we would like to test the null hypothesis $H_0 : \beta = \beta^*$ against the two-sided alternative. It turns out that if β^* does not belong to the confidence interval above, then one should reject H_0. To see why this is the case, note that the rejection region for this two-sided test with level of significance equal to 5% is

$$\frac{\hat{\beta} - \beta^*}{\text{SE}[\hat{\beta}]} \leq -q \quad \text{or} \quad \frac{\hat{\beta} - \beta^*}{\text{SE}[\hat{\beta}]} \geq q.$$

This is equivalent to $\beta^* \geq \hat{\beta} + q \cdot \text{SE}[\hat{\beta}]$ or $\beta^* \leq \hat{\beta} - q \cdot \text{SE}[\hat{\beta}]$, which, in turn, is equivalent to $\beta^* \notin \left(\hat{\beta} - q \cdot \text{SE}[\hat{\beta}], \hat{\beta} + q \cdot \text{SE}[\hat{\beta}]\right)$, i.e., β^* does not belong to the 95% confidence interval.

Multiple linear regression model

BUFN 758N

Prof. Skoulakis

Generalization to the Multiple Linear Regression

- Linear model with one regressor

$$y_t = \alpha + \beta x_t + u_t, \quad t = 1, \ldots, T$$

- What if the dependent variable (y) depends on multiple independent (explanatory, x) variables?
- Example: stock returns might depend on a number of characteristics such as size and book-to-market.
- Incorporating just one independent variable is not sufficient in this case - more than one x variables are required. The generalization from the simple earlier model to a model with multiple regressors is conceptually straightforward but requires some additional notation.

Multiple Linear Regression: Notation

- Multiple regressors: $x_1, x_2, \ldots, x_k$ (k regressors)
- The regression equation now reads

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + u_t = \boldsymbol{\beta}' \mathbf{x}_t + u_t$$

  where $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \cdots & \beta_k \end{bmatrix}'$ and $\mathbf{x}_t = \begin{bmatrix} x_{1t} & \cdots & x_{kt} \end{bmatrix}'$.
- If the first regressor x_1 is equal to the constant 1, then the coefficient β_1 is called the intercept. The coefficients $\beta_2, \ldots, \beta_k$ are called the slope coefficients.
- The simple linear model obtains when $k = 2$ and the first regressor is the constant 1 (the notation correspondence is $\alpha \leftrightarrow \beta_1$ and $\beta \leftrightarrow \beta_2$).

Different ways of Representing the Multiple Linear Regression

- The full system of equations is expressed as

$$\begin{aligned} y_1 &= \beta_1 x_{11} + \cdots + \beta_k x_{k1} + u_1 = \boldsymbol{\beta}' \mathbf{x}_1 + u_1 \\ y_2 &= \beta_1 x_{12} + \cdots + \beta_k x_{k2} + u_2 = \boldsymbol{\beta}' \mathbf{x}_2 + u_2 \\ \vdots & \qquad\qquad \vdots \qquad\qquad \vdots \\ y_T &= \beta_1 x_{1T} + \cdots + \beta_k x_{kT} + u_T = \boldsymbol{\beta}' \mathbf{x}_T + u_T \end{aligned}$$

- In matrix form the full system reads

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where $\mathbf{y}$ is a $T \times 1$ vector, $\mathbf{X}$ is a $T \times k$ matrix, $\boldsymbol{\beta}$ is a $k \times 1$ vector, and $\mathbf{u}$ is a $T \times 1$ vector.

Multiple Linear Regression: Matrix Representation

- Matrix form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1T} & x_{2T} & \cdots & x_{kT} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_T' \end{bmatrix},$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \ \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}$$

- Note that the vectors/matrices $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \mathbf{u}$ are conformable, i.e., the dimensions match.

Obtaining Parameter Estimates in the Multiple Regression Model

- In the simple linear model, the estimates of α and β are obtained by minimizing the residual sum of squares (RSS)
- Using vector notation, the residuals are represented by

$$\hat{\mathbf{u}} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_T \end{bmatrix} = \begin{bmatrix} y_1 - \hat{\boldsymbol{\beta}}' \mathbf{x}_1 \\ y_2 - \hat{\boldsymbol{\beta}}' \mathbf{x}_2 \\ \vdots \\ y_T - \hat{\boldsymbol{\beta}}' \mathbf{x}_T \end{bmatrix}$$

- The RSS is given by $\hat{\mathbf{u}}'\hat{\mathbf{u}} = \hat{u}_1^2 + \cdots + \hat{u}_T^2 = \sum_{t=1}^{T} \hat{u}_t^2$

The Multiple Regression OLS Estimator

It can be shown that the OLS estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

The assumption that the $T \times k$ regressor matrix $\mathbf{X}$ has full rank is necessary for the OLS estimator to be well-defined. Under this assumption the $k \times k$ matrix $\mathbf{X}'\mathbf{X}$ has full rank equal to k and is invertible.

An alternative expression for the OLS estimator $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left(\sum_{t=1}^{T} \mathbf{x}_t y_t \right).$$

Non-stochastic Regressors

We first study the properties of the OLS estimator under the assumption that the regressors are non-stochastic (fixed).

As in the case of a single regressor, the following assumptions about the shocks u_t are made:

- A1. The shocks u_t have zero mean: $\mathbb{E}[u_t] = 0$.
- A2. The shocks u_t are homoscedastic: $\mathbb{V}[u_t] = \sigma^2 < \infty$.
- A3. The shocks u_t are uncorrelated: $\mathbb{COV}[u_t, u_s] = 0$, for $t \neq s$.
- A4. The shocks u_t are (jointly) normally distributed: $u_t \sim N(0, \sigma^2)$.

Unbiasedness of the OLS Estimator

Under Assumption A1, we have $\mathbb{E}[u_t] = 0$ and so

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \Rightarrow \mathbb{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}.$$

Since $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, we have

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Hence, the OLS estimator $\hat{\boldsymbol{\beta}}$ is unbiased.

Variance of the OLS Estimator

Under Assumption A1, we have $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ and since $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned}
\mathbb{V}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\
&= \mathbb{E}[\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right)\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right)'] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{u}\mathbf{u}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}$$

Under Assumptions A2 and A3 (that the shocks u_t are uncorrelated and homoscedastic with variance equal to σ^2), we have that $\mathbb{E}[\mathbf{u}\mathbf{u}'] = \sigma^2\mathbf{I}_T$, where $\mathbf{I}_T$ is the $T \times T$ identity matrix. It follows that, under Assumptions A1, A2, and A3, we have

$$\mathbb{V}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Standard Errors of the OLS Estimator

- Since $\mathbb{V}[u_t] = \mathbb{E}[u_t^2] = \sigma^2$, a natural estimator of σ^2 is $\frac{1}{T}\sum_{t=1}^{T}\hat{u}_t^2 = \frac{1}{T}\hat{\mathbf{u}}'\hat{\mathbf{u}}$.
- Note that $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}_X\mathbf{y}$ where $\mathbf{M}_X = \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
- The matrix $\mathbf{M}_X$ has the property $\mathbf{M}_X\mathbf{X} = \mathbf{O}_{T\times k}$, and so $\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{y}'\mathbf{M}_X\mathbf{y} = (\mathbf{X}\boldsymbol{\beta} + \mathbf{u})'\mathbf{M}_X(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \mathbf{u}'\mathbf{M}_X\mathbf{u}$
- $\mathbf{M}_X$ is symmetric and idempotent ($\mathbf{M}_X = \mathbf{M}_X\mathbf{M}_X$) with $\mathrm{tr}(\mathbf{M}_X) = \mathrm{tr}(\mathbf{I}_T) - \mathrm{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \mathrm{tr}(\mathbf{I}_T) - \mathrm{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) = \mathrm{tr}(\mathbf{I}_T) - \mathrm{tr}(I_k) = T - k$ and so $\mathbb{E}[\mathbf{u}'\mathbf{M}_X\mathbf{u}] = \mathbb{E}\left[\mathrm{tr}\left(\mathbf{u}'\mathbf{M}_X\mathbf{u}\right)\right] = \mathbb{E}\left[\mathrm{tr}\left(\mathbf{M}_X\mathbf{u}\mathbf{u}'\right)\right] = \mathrm{tr}\left(\mathbb{E}\left[\mathbf{M}_X\mathbf{u}\mathbf{u}'\right]\right) = \mathrm{tr}\left(\mathbf{M}_X\mathbb{E}\left[\mathbf{u}\mathbf{u}'\right]\right) = \mathrm{tr}\left(\mathbf{M}_X\left(\sigma^2\mathbf{I}_T\right)\right) = \sigma^2\mathrm{tr}\left(\mathbf{M}_X\right) = (T-k)\sigma^2$.

Standard Errors of the OLS Estimator

- Hence, the unbiased estimator of σ^2 is

$$s^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{T-k}.$$

- The covariance matrix of the OLS estimator $\hat{\boldsymbol{\beta}}$ is estimated by

$$s^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{T-k}(\mathbf{X}'\mathbf{X})^{-1}.$$

- The k-th diagonal element of $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is the estimator of the variance of $\hat{\beta}_k$.

- The square root of the k-th diagonal element of $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is the standard error of $\hat{\beta}_k$.

Distribution of the OLS Estimator

Under Assumptions A1, A2, A3, and A4 (joint normality of the shocks), the OLS estimator is normally distributed.

Recall that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$. Since $\mathbf{u} \sim N(\mathbf{0}_T, \sigma^2\mathbf{I}_T)$, we have

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \sim N(\mathbf{0}_T, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

and so

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

Distribution of the OLS Estimator (cont'd)

- Let $\mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1}$ and denote by v_{ii} its i-th diagonal element. Then,

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{v_{ii}}} \sim N(0,1).$$

- In practice, we estimate σ^2 by $s^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{T-k} = \frac{1}{T-k}\sum_{t=1}^{T}\hat{u}_t^2$ in which case we have

$$\frac{\hat{\beta}_i - \beta_i}{s\sqrt{v_{ii}}} \sim t_{T-k}.$$

- One can use the above result to test hypotheses of the form $H_0 : \beta_i = \beta_i^*$, where β_i^* is a scalar constant.

Testing multiple (joint) hypotheses: the F-test

- In empirical applications, one is frequently interested in testing hypotheses that involve multiple parameters.
- Suppose that the model is $y_t = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u_t$ $(k = 3)$. An example of joint hypothesis is that all slopes are equal to zero, i.e., $H_0 : \beta_2 = \beta_3 = 0$.
- The number of restrictions involved in the null hypothesis is important. In the example above, the number of restrictions is 2. For the null hypothesis $H_0 : \beta_2 = \beta_3$ there is only one restriction.
- Formally, we consider (linear) null hypotheses of the form $H_0 : \mathbf{A}\beta = \mathbf{c}$, where $\mathbf{A}$ is an $r \times k$ matrix with full rank equal to $r \leq k$, and $\mathbf{c}$ is an $r \times 1$ vector. This means that there are exactly r restrictions imposed by the null hypothesis.

Quadratic Forms of Normally Distributed Vectors

- Let $\mathbf{x} = [x_1 \ \cdots \ x_n]'$ be an $n \times 1$ vector and $\mathbf{W} = (w_{ij})_{i,j=1,\ldots,n}$ be a symmetric $n \times n$ matrix. The quantity

$$\mathbf{x}'\mathbf{W}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i x_j$$

  is called a quadratic form in $\mathbf{x}$ with weighting matrix $\mathbf{W}$.

- **Fact 1**. Let $\mathbf{x}$ be an $n \times 1$ normally distributed vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, i.e. $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, the quadratic form $q = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ follows a chi-square distribution with n degrees of freedom, i.e., $q \sim \chi^2(n)$.

Quadratic Forms of Normally Distributed Vectors

- **Fact 2**. Let $\mathbf{u}$ be an $n \times 1$ normally distributed vector with mean $\mathbf{0}_n$ and identity covariance matrix $\mathbf{I}_n$, i.e., $\mathbf{u} \sim N(\mathbf{0}_n, \mathbf{I}_n)$. Let $\mathbf{M}$ be a nonrandom $n \times n$ symmetric and idempotent matrix with $\mathrm{rank}(\mathbf{M}) = r \leq n$. Then, the quadratic form $q = \mathbf{u}'\mathbf{M}\mathbf{u}$ follows a chi-square distribution with r degrees of freedom, i.e., $q \sim \chi^2(r)$.

- **Fact 3**. Let $\mathbf{u}$ be an $n \times 1$ normally distributed vector with mean $\mathbf{0}_n$ and identity covariance matrix $\mathbf{I}_n$, i.e., $\mathbf{u} \sim N(\mathbf{0}_n, \mathbf{I}_n)$. Let $\mathbf{M}$ be a nonrandom $n \times n$ idempotent matrix with $\mathrm{rank}(\mathbf{M}) = r \leq n$, and $\mathbf{L}$ be a nonrandom $m \times n$ matrix such that $\mathbf{LM} = \mathbf{O}_{m \times n}$. Then, the random vectors $\mathbf{x} = \mathbf{Mu}$ and $\mathbf{y} = \mathbf{Lu}$ are independently distributed.

Definition of the F distribution

Let W_1 and W_2 be two independent chi-square random variables with m_1 and m_2 degrees of freedom, i.e., $W_1 \sim \chi^2(m_1)$ and $W_2 \sim \chi^2(m_2)$. Then, $U = \frac{W_1/m_1}{W_2/m_2}$ follows the (Snedecor) F distribution with m_1 and m_2 degrees of freedom.

Fact: If $U \sim F(m, n)$ then, as $n \to \infty$, $mU \xrightarrow{d} \chi^2(m)$.

To see this, note that we have $U = \frac{W_1/m}{W_2/n}$ with $W_1 \sim \chi^2(m)$ and $W_2 \sim \chi^2(n)$ and so $mU = \frac{W_1}{W_2/n}$. But W_2 can be written as $Z_1^2 + \cdots + Z_n^2$ where $Z_1, \ldots, Z_n$ are i.i.d. standard normal random variables. By LLN, $\frac{Z_1^2+\cdots+Z_n^2}{n} \xrightarrow{p} \mathbb{E}[Z^2] = 1$, as $n \to \infty$. It follows that $W_2/n \xrightarrow{p} 1$, as $n \to \infty$ and $mU \xrightarrow{d} W_1 \sim \chi^2(m)$.

The F test

Suppose we wish to test the hypothesis $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ where the $r \times k$ matrix $\mathbf{A}$ has rank equal to r. Under Assumptions A1-A4, the OLS estimator is normally distributed: $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. Hence, under the null hypothesis $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$, we have

$$\mathbf{A}\hat{\boldsymbol{\beta}} \sim N(\mathbf{c}, \sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')$$

and so

$$(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})' \left[\sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\right]^{-1} (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}) \sim \chi^2(r)$$

since $\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'$ is an $r \times r$ invertible (full rank) matrix.

The F test (cont'd)

However, this result cannot be used in this form since σ^2 is unknown and, in practice, is estimated by $s^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{T-k}$. It turns out that $(T-k)\frac{s^2}{\sigma^2}$ follows a chi-square distribution with $T-k$ degrees of freedom and is independent of $\hat{\boldsymbol{\beta}}$. Hence, according to the definition of the F distribution, the ratio

$$\frac{(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})' \left[\sigma^2 \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\right]^{-1} (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})/r}{\left[(T-k)\frac{s^2}{\sigma^2}\right]/(T-k)}$$

follows an $F(r, T-k)$ distribution. This ratio equals

$$\frac{1}{r}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})' \left[s^2 \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\right]^{-1} (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})$$

and is the test statistic for testing the null hypothesis $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$.

Stochastic regressors

- The results on the OLS estimators extend to the case of the multiple linear regression model.
- **Unbiasedness**. If (i) the shocks u_t have zero mean, i.e., $\mathbb{E}[u_t] = 0$ and (ii) the shock vector $\mathbf{u}$ and the regressor matrix $\mathbf{X}$ are independent, then $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. To see this, note that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ and so

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}.$$

Due to independence, $\mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right] = \mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbb{E}\left[\mathbf{u}\right]$ $= \mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{0}_T = \mathbf{0}_k$ and so $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

Consistency and Asymptotic Normality of the OLS Estimator

- **Consistency and asymptotic normality**. Assume that (i) the sequence $(\mathbf{x}_t, u_t)$ is i.i.d. over time with finite mean and variance (ii) the shock u_t has zero mean, i.e., $\mathbb{E}[u_t] = 0$ and (iii) the regressor $\mathbf{x}_t$ and the shock u_t are uncorrelated, i.e., $\mathbb{E}[\mathbf{x}_t u_t] = \mathbf{0}_k$. Then, the OLS estimator $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal.

  Recall that $\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right)^{-1} \left(\sum_{t=1}^{T} \mathbf{x}_t y_t\right)$ and use $y_t = \mathbf{x}_t' \boldsymbol{\beta} + u_t$ to obtain $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t u_t\right)$ from which, by the LLN, it follows $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta} + \left(\mathbb{E}[\mathbf{x}_t \mathbf{x}_t']\right)^{-1} \mathbb{E}[\mathbf{x}_t u_t]$ $= \boldsymbol{\beta} + \left(\mathbb{E}[\mathbf{x}_t \mathbf{x}_t']\right)^{-1} \mathbf{0}_k = \boldsymbol{\beta}$.

Consistency and Asymptotic Normality of the OLS Estimator

Since $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t u_t\right)$, we have $\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{x}_t u_t\right)$. By the LLN, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' \xrightarrow{p} \mathbb{E}[\mathbf{x}_t\mathbf{x}_t'],$$

and by the CLT, we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{x}_t u_t \xrightarrow{d} N(\mathbf{0}_k, \mathbb{E}[u_t^2(\mathbf{x}_t\mathbf{x}_t')]).$$

Consistency and Asymptotic Normality of the OLS Estimator

Hence,

$$\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N\left(\mathbf{0}_k, (\mathbb{E}[\mathbf{x}_t\mathbf{x}_t'])^{-1} \cdot \mathbb{E}[u_t^2(\mathbf{x}_t\mathbf{x}_t')] \cdot (\mathbb{E}[\mathbf{x}_t\mathbf{x}_t'])^{-1}\right).$$

If $\mathbf{x}_t$ and u_t are independent, then $\mathbb{E}[u_t^2(\mathbf{x}_t\mathbf{x}_t')] = \mathbb{E}[u_t^2]\mathbb{E}[\mathbf{x}_t\mathbf{x}_t'] = \sigma^2\mathbb{E}[\mathbf{x}_t\mathbf{x}_t']$, and so

$$\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N\left(\mathbf{0}_k, \sigma^2(\mathbb{E}[\mathbf{x}_t\mathbf{x}_t'])^{-1}\right).$$

Standard Errors of the OLS Estimator

- The asymptotic covariance matrix $\sigma^2(\mathbb{E}[\mathbf{x}_t\mathbf{x}_t'])^{-1}$, is estimated by the consistent estimator $s^2\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}$, where $s^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{T} = \frac{\sum_{t=1}^{T}\hat{u}_t^2}{T}$.
- The asymptotic standard error of $\hat{\beta}_i$ is equal to s times the square root of the i-th diagonal element of the matrix $\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}$.
- Using the fact $\frac{\hat{\beta}_i-\beta_i}{\mathrm{SE}[\hat{\beta}_i]} \xrightarrow{d} N(0,1)$, we can conduct hypothesis testing for null hypotheses of the form $H_0 : \beta_i = \beta_i^*$.

Stochastic Regressors: Testing Joint Hypotheses

- One can also test joint hypotheses of the form $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$, where $\mathbf{A}$ is an $r \times k$ matrix with full rank equal to $r \leq k$, and $\mathbf{c}$ is an $r \times 1$ vector.
- Under the assumption that $\mathbf{x}_t$ and u_t are independent, we have $\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N\left(\mathbf{0}_k, \sigma^2(\mathbb{E}[\mathbf{x}_t\mathbf{x}_t'])^{-1}\right)$. Under $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$, we have $\sqrt{T}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}) \xrightarrow{d} N\left(\mathbf{0}_k, \sigma^2\mathbf{A}(\mathbb{E}[\mathbf{x}_t\mathbf{x}_t'])^{-1}\mathbf{A}'\right)$ and so

$$T(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})' \left[\sigma^2\mathbf{A}(\mathbb{E}[\mathbf{x}_t\mathbf{x}_t'])^{-1}\mathbf{A}'\right]^{-1} (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}) \xrightarrow{d} \chi^2(r).$$

Stochastic Regressors: Testing Joint Hypotheses

- To obtain an operational version of the above test, we consistently estimate σ^2 by $s^2 = \frac{\sum_{t=1}^{T} \hat{u}_t^2}{T}$ and $\mathbb{E}[\mathbf{x}_t \mathbf{x}_t']$ by $\frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'$.
- The resulting test is

$$\frac{T(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})' \left[\mathbf{A}(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t')^{-1} \mathbf{A}'\right]^{-1} (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})}{\sum_{t=1}^{T} \hat{u}_t^2} \xrightarrow{d} \chi^2(r).$$

- One then rejects the null hypothesis $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ at the α level of significance if the above test statistic is greater than $q_\alpha(r)$ where $\mathbb{P}[\chi^2(r) \geq q_\alpha(r)] = \alpha$.

Data Mining

- Data mining is the process of examining many series for statistical relationships with no theoretical justification
- For instance, suppose that we regress the dependent variable on each of twenty explanatory variables, that are randomly and independently selected. Then, on average one slope coefficient will be found significant at 5%
- If data mining occurs, the true significance level will be greater that the nominal significance level

R^2: A Goodness of Fit Statistic

- It is desirable to obtain some measure of how well the regression model describes the data
- There are *goodness of fit* statistics for this purpose, i.e., measuring how well the sample regression function fits the data
- The most commonly used goodness of fit statistic is the so-called R^2, which is the square of the correlation coefficient between the actual data y and the fitted data $\hat{y}$:

$$R^2 = \text{Corr}\left(\mathbf{y}, \hat{\mathbf{y}}\right)^2 = \frac{\left[\sum_{t=1}^{T}\left(y_t - \bar{y}\right)\left(\hat{y}_t - \bar{\hat{y}}\right)\right]^2}{\left[\sum_{t=1}^{T}\left(y_t - \bar{y}\right)^2\right]\left[\sum_{t=1}^{T}\left(\hat{y}_t - \bar{\hat{y}}\right)^2\right]}$$

R^2: A Goodness of Fit Statistic (cont'd)

- An alternative expression for R^2 is

$$R^2 = \frac{\text{ESS}}{\text{TSS}}$$

where the total sum of squares is $\text{TSS} = \sum_{t=1}^{T} (y_t - \bar{y})^2$ and the explained sum of squares is $\text{ESS} = \sum_{t=1}^{T} (\hat{y}_t - \bar{y})^2$.

- Moreover, TSS is decomposed as follows

$$\begin{aligned} \text{TSS} &= \text{ESS} + \text{RSS} \\ \sum_{t=1}^{T} (y_t - \bar{y})^2 &= \sum_{t=1}^{T} (\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \end{aligned}$$

R^2: A Goodness of Fit Statistic (cont'd)

- Since $\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$, we can write

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

- The lower the residual sum of squares RSS the better the fit
- R^2 always lies between 0 and 1
- The two extreme scenarios are
  - $\mathrm{RSS} = \mathrm{TSS} \Leftrightarrow \mathrm{ESS} = 0 \Leftrightarrow R^2 = 0$
  - $\mathrm{ESS} = \mathrm{TSS} \Leftrightarrow \mathrm{RSS} = 0 \Leftrightarrow R^2 = 1$

Issues with R^2 as a Goodness of Fit Measure

- R^2 is a measure of *linear* correlation. If the y variable is transformed nonlinearly, R^2 will change
- R^2 does not fall if more regressors are added to the regression
- A modification of R^2 called the adjusted R^2 accounts for the number of explanatory variables

Adjusted R^2

- The adjusted R^2, denoted by $\bar{R}^2$, is defined by

$$\bar{R}^2 = 1 - \frac{\mathrm{RSS}/(T-k)}{\mathrm{TSS}/(T-1)} = 1 - \frac{T-1}{T-k}\left(1-R^2\right).$$

- It follows from the above definition that if an additional regressor is included in the model, k increases and unless R^2 increases enough, the adjusted R^2 will actually fall.
- Example: assume the sample size is $T = 240$ and that a model with two regressors ($k = 2$) yields an $R^2 = 3.5\%$. The corresponding adjusted R^2 is 3.09%. If two more regressors are added (so that $k = 4$) and the R^2 increases to 4%, the adjusted R^2 falls to 2.78%.
- If the R^2 is very small $\left(R^2 < \frac{k-1}{T-1}\right)$, the adjusted R^2 can be negative.

Tests of Non-nested Hypotheses

- So far we have focused on hypothesis tests in the context of nested models

- But what if we wish to compare two models that are not nested?

$$\text{Model 1} \quad : \quad y_t = \gamma_1 + \gamma_2 x_{2t} + u_t$$
$$\text{Model 2} \quad : \quad y_t = \delta_1 + \delta_2 x_{3t} + v_t$$

- We could use R^2, but that can be problematic if the number of explanatory variables differs across the two models. In such a case, it is better to use adjusted R^2 that takes into account the number of explanatory variables.

Tests of Non-nested Hypotheses (cont'd)

- An alternative approach is to use an encompassing test, based on examination of the hybrid model

  Model 3 : $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + w_t$

- There are four possible outcomes when Model 3 is estimated
  - A. β_2 is statistically significant but β_3 is not
  - B. β_3 is statistically significant but β_2 is not
  - C. β_2 and β_3 are both statistically significant
  - D. neither β_2 nor β_3 are statistically significant
- The encompassing test is inconclusive in the cases C and D
- Other issues with the encompassing approach
  - The hybrid model might be economically meaningless
  - The regressors x_2 and x_3 might be highly correlated

Linear regression: assumptions and diagnostics

BUFN 758N

Prof. Skoulakis

Linear Regression Assumptions

- Zero-mean shocks: $\mathbb{E}[u_t] = 0$
- Uncorrelated shocks: $\mathbb{E}[u_t u_s] = 0,\ t \neq s$
- Homoscedastic shocks: $\mathbb{E}[u_t^2] = \sigma^2,\ t = 1, \ldots, T$
- Normally distributed shocks: $u_t \sim N(0, \sigma^2)$

Zero-mean disturbances assumption: $\mathbb{E}[u_t] = 0$

- This is a "technical" assumption that can be taken care of by including an intercept in the regression.
- What if a particular application requires no intercept? This would force the regression through the origin.
- It is good practice to include the intercept and test whether it is zero.
- If an intercept is not included in the regression, one should be cautious.

Simple Linear Regression with no Intercept

- If we do not force u_t to have zero mean, then this is equivalent to including an intercept in the regression. What happens if we do not use an intercept and require $\mathbb{E}[u_t] = 0$?
- Consider the model: $y_t = \beta x_t + u_t, \;\; t = 1, \ldots, T$.
- The OLS estimator of β is $\hat{\beta} = \frac{\sum_{t=1}^{T} x_t y_t}{\sum_{t=1}^{T} x_t^2}$.
- The requirement $\mathbb{E}[u_t] = 0$ is equivalent to $\mathbb{E}[y_t] = \beta \mathbb{E}[x_t]$. In the case of non-stochastic regressors, this guarantees unbiasedness of $\hat{\beta}$.
- In the case of stochastic regressors, consistency of $\hat{\beta}$ requires the condition $\mathbb{E}[x_t u_t] = 0 \Leftrightarrow \mathbb{E}[x_t y_t] = \beta \mathbb{E}[x_t^2]$.

Simple Linear Regression with no Intercept

- If $\mathbb{E}[y_t] = \mathbb{E}[x_t] = 0$ then there would be no problem. But this case is rare and hard to detect.
- If we do not include an intercept, the slope estimates could be severely biased.
- The interpretation of R^2 is not the same anymore. The sample average of the fitted values, i.e., $\bar{\hat{y}}$, is not equal to the sample average of the actual y values. Indeed, $\bar{\hat{y}} = \hat{\beta}\bar{x} \neq \bar{y}$.
- If one defines R^2 by $R^2 = 1 - \frac{RSS}{TSS}$ where $RSS = \sum_{t=1}^{T}(y_t - \hat{y}_t)^2$ and $TSS = \sum_{t=1}^{T}(y_t - \bar{y})^2$ (as before), then it is possible that R^2 is negative. The interpretation in this case is that the sample average $\bar{y}$ explains more of the variation in y than the regressors.

Testing Uncorrelatedness

- One of the assumptions in the linear regression model is that the shocks are uncorrelated.
- To examine whether this assumption is violated, the first natural thing to do is to examine the autocorrelation of the shocks.
- Since the true shocks are unobserved, we work with the regression residuals $\hat{u}_t = y_t - \hat{y}_t$.
- The first order autocorrelation of a variable, say z, is the correlation of z_t with the lagged value z_{t-1}.

Testing Uncorrelatedness

- The first order autocorrelation is given by $\rho = \frac{\mathbb{C}[z_t, z_{t-1}]}{\mathbb{V}[z_t]}$.
- In the case when $\mathbb{E}[z_t] = 0$, the first order autocorrelation is estimated using the model: $z_t = \rho z_{t-1} + \varepsilon_t$.
- Using data $z_1, \ldots, z_T$, the OLS estimate of ρ then is

$$\hat{\rho}_{\text{OLS}} = \frac{\sum_{t=2}^{T} z_{t-1} z_t}{\sum_{t=2}^{T} z_{t-1}^2}.$$

- A slightly modified estimator (*sample* first order autocorrelation) is

$$\hat{\rho} = \frac{\sum_{t=2}^{T} z_{t-1} z_t}{\sum_{t=1}^{T} z_t^2}.$$

- The two estimators of ρ are consistent estimators of ρ and asymptotically equivalent.

The Durbin-Watson Test for Uncorrelatedness

- The Durbin-Watson statistic is defined by

$$DW = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T} \hat{u}_t^2}$$

where $\hat{u}_1, \ldots, \hat{u}_T$ are the regression residuals, i.e.,

$$\hat{u}_t = y_t - \hat{y}_t = y_t - (\hat{\alpha} + \hat{\boldsymbol{\beta}}' x_t).$$

- The DW statistic relates to the sample first order correlation of $\hat{u}_t$, i.e., $\tilde{\rho} = \frac{\sum_{t=2}^{T} \hat{u}_{t-1}\hat{u}_t}{\sum_{t=1}^{T} \hat{u}_t^2}$, as follows

$$DW = 2(1 - \tilde{\rho}) - \frac{\hat{u}_1^2 + \hat{u}_T^2}{\sum_{t=1}^{T} \hat{u}_t^2}$$

- For large T, we have $DW \approx 2(1 - \tilde{\rho})$.

The Durbin-Watson Test for Uncorrelatedness

- The sample first order autocorrelation of the $\hat{u}_t$, i.e., $\tilde{\rho}$, is an estimator of the (population) first order correlation of the u_t, i.e., $\rho = \frac{\mathbb{E}[u_t u_{t-1}]}{\mathbb{E}[u_t^2]}$.
- Actually, $\tilde{\rho} = \frac{\sum_{t=2}^{T} \hat{u}_{t-1}\hat{u}_t}{\sum_{t=1}^{T} \hat{u}_t^2}$ is a consistent estimator of ρ.
- Using this fact, we obtain the the DW statistic converges in probability to $2(1-\rho)$.
- Hence, values of *DW* close to 2 indicate uncorrelatedness, values of *DW* significantly less than 2 (and towards 0) indicate positive autocorrelation, values of *DW* significantly larger than 2 (and towards 4) indicate negative autocorrelation.

The Durbin-Watson Test for Uncorrelatedness

- The exact distribution of the DW statistic depends of the regressor matrix $\mathbf{X}$.
- However, the distribution of the DW lies between the distributions of two other statistics DW_L (lower bound) and DW_U (upper bound) that depend only on T and k. It follows that the critical value (for a given size) for DW must lie to the right of that of DW_L and to the left of that of DW_U.

The Durbin-Watson Test for Uncorrelatedness

- Rejection and non-rejection regions for the DW test:
  - if $DW < D_L^*$ then we reject H_0 (positive correlation),
  - $DW > 4 - D_L^*$ then we reject H_0 (negative correlation),
  - if $D_U^* < DW < 4 - D_U^*$ then we do not reject H_0,
  - if $D_L^* \leq DW \leq D_U^*$ then no conclusion is drawn,
  - if $4 - D_U^* \leq DW \leq 4 - D_L^*$ then no conclusion is drawn.
- The DW statistic rejects the null hypothesis with probability less than the nominal size α and does not reject the null hypothesis with probability less than $1 - \alpha$.

The Breusch-Godfrey Test for Uncorrelatedness

- The limitation of the DW test is that it focuses on first order autocorrelation.
- The Breusch-Godfrey test allows examination of the relationship between $\hat{u}_t$ and several of its lagged values at the same time.
- It is based on the auxiliary regression

$$\begin{aligned} \hat{u}_t &= \gamma_1 + \gamma_2 x_{2t} + \cdots + \gamma_k x_{kt} \\ &\quad + \phi_1 \hat{u}_{t-1} + \cdots + \phi_m \hat{u}_{t-m} + \epsilon_t \end{aligned}$$

- We can test for the null hypothesis $H_0 : \phi_1 = \cdots = \phi_m = 0$ using the statistic $(T - m)R^2$ where R^2 is obtained from the auxiliary regression and asymptotically follows a $\chi^2(m)$ distribution.
- Selecting an appropriate value for the maximum lag value m remains an issue.

The Box-Pierce Test for Uncorrelatedness

- The Box-Pierce test is based on the sample autocorrelations (of several orders) of the residuals.
- The j-th order sample autocorrelation of the residuals is given by

$$\hat{\rho}_j = \frac{\sum_{t=j+1}^{T} \hat{u}_t \hat{u}_{t-j}}{\sum_{t=1}^{T} \hat{u}_t^2}.$$

- The Box-Pierce Q statistic is given by

$$Q = T \sum_{j=1}^{m} \hat{\rho}_j^2.$$

- The Q statistic asymptotically follows a $\chi^2(m)$ distribution.

The Ljung-Box Test for Uncorrelatedness

- The Ljung-Box Q' statistic is a refinement of Q.
- The Ljung-Box Q' statistic is given by

$$Q' = T(T+2)\sum_{j=1}^{m}\frac{\hat{\rho}_j^2}{T-j}.$$

- The Q' statistic asymptotically follows a $\chi^2(m)$ distribution.
- The Ljung-Box Q' statistic is preferred to the Box-Pierce Q statistic since it has better small sample properties.
- The choice of an appropriate value for the maximum lag value m is still an issue.

Shock Heteroscedasticity

- One of the assumptions in the linear regression model is that the shocks are homoscedastic.
- If the shocks are heteroscedastic, then the OLS estimator is unbiased but no longer Best Linear Unbiased Estimator, i.e., it is not efficient anymore.
- More importantly, the OLS standard errors are not valid anymore.
- There are several methods for detecting and dealing with shock heteroscedasticity.
- Financial data are often characterized by certain types of heteroscedasticity, such as (Generalized) Autoregressive Conditional Heteroscedasticity (ARCH, GARCH).

Digression: Tests for joint slope significance

- Consider the model $y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t$ and the null hypothesis that all slope coefficients are equal to 0, i.e., $H_0 : \beta_2 = \cdots = \beta_k = 0$.

- The F statistic for H_0 is $F = \frac{TSS - RSS}{RSS} \times \frac{T-k}{k-1}$ which follows an $F(k-1, T-k)$ distribution in the case of fixed regressors and normally distributed shocks.

- The F statistic relates to R^2 as follows:

$$F = \frac{TSS - RSS}{RSS} \times \frac{T-k}{k-1} = \frac{R^2}{1-R^2} \times \frac{T-k}{k-1}$$

Digression: Tests for joint slope significance

- Solving for R^2, we get $TR^2 = \frac{T}{(T-k)+(k-1)F} \times (k-1)F$.
- As $T \to \infty$, $(k-1)F \xrightarrow{d} \chi^2(k-1)$ and so $\frac{T}{(T-k)+(k-1)F} \xrightarrow{d} 1$ which implies $\frac{T}{(T-k)+(k-1)F} \xrightarrow{p} 1$. It then follows

$$TR^2 = \frac{T}{(T-k)+(k-1)F} \times (k-1)F \xrightarrow{d} \chi^2(k-1).$$

- So, an asymptotic test for the null hypothesis H_0 can be based on the test statistic TR^2 that asymptotically follows a $\chi^2(k-1)$ distribution.

White's Test for Heteroscedasticity

- Example: $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$. Estimate the model using standard OLS and obtain the residuals $\hat{u}_t$.
- Then run the auxiliary regression

$$\hat{u}_t^2 = \gamma_1 + \gamma_2 x_{2t} + \gamma_3 x_{3t} + \gamma_4 x_{2t}^2 + \gamma_5 x_{3t}^2 + \gamma_6 x_{2t} x_{3t} + \epsilon_t$$

  and test for the null hypothesis $H_0 : \gamma_2 = \cdots = \gamma_6 = 0$.
- Denote the number of regressors in the auxiliary regression by m (including the constant). In the above example, $m = 6$.
- We can use (i) the F statistic $F = \frac{TSS - RSS}{RSS} \times \frac{T-m}{m-1}$ which follows a $F(m-1, T-m)$ distribution or (ii) the statistic TR^2 which asymptotically follows a $\chi^2(m-1)$ distribution.
- Caution: both of these statistics should be derived from the auxiliary regression.

Dealing with Heteroscedasticity: White standard errors

- Consider the multiple linear regression model $y_t = \boldsymbol{\beta}' \mathbf{x}_t + u_t$ and assume that the regressors are fixed.
- Heteroscedasticity: $\mathbb{E}[u_t^2] = \sigma_t^2,\ t = 1, \ldots, T$
- The covariance matrix of the shock vector $\mathbf{u}$ is

$$\mathbb{E}[\mathbf{u}\mathbf{u}'] = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_{T-1}^2 & 0 \\ 0 & 0 & \cdots & 0 & \sigma_T^2 \end{bmatrix} \equiv \boldsymbol{\Sigma}$$

- The OLS estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}'\mathbf{Y}$

Dealing with Heteroscedasticity: White standard errors

- Even under heteroscedasticity, the OLS estimator $\hat{\boldsymbol{\beta}}$ is unbiased: $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. Since $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}'\mathbf{u}$, the variance of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned}
\mathbb{V}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\
&= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}$$

- Note that $\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X} = \sum_{t=1}^{T} \sigma_t^2(\mathbf{x}_t\mathbf{x}_t')$ which can be estimated consistently by $\sum_{t=1}^{T} \hat{u}_t^2(\mathbf{x}_t\mathbf{x}_t')$ (under regularity conditions).
- $\mathbb{V}[\hat{\boldsymbol{\beta}}]$ is then estimated by $(\mathbf{X}'\mathbf{X})^{-1}\left[\sum_{t=1}^{T} \hat{u}_t^2(\mathbf{x}_t\mathbf{x}_t')\right](\mathbf{X}'\mathbf{X})^{-1}$. Variations of the White standard error estimator have also been developed with better small sample properties.
- Note that uncorrelatedness is still required.

Testing Normality

- One of the assumptions in the linear regression model is that the shocks are normally distributed.
- This assumption is not so crucial if we have a sufficiently large sample, as asymptotic theory can be invoked.
- The standardized third and fourth moments (about the mean μ) of the distribution of a random variable X are known as the skewness (ξ) and kurtosis (κ):

$$\xi = \frac{\mathbb{E}\left[(X-\mu)^3\right]}{\left(\mathbb{E}\left[(X-\mu)^2\right]\right)^{\frac{3}{2}}}, \quad \kappa = \frac{\mathbb{E}\left[(X-\mu)^4\right]}{\left(\mathbb{E}\left[(X-\mu)^2\right]\right)^2}$$

- The skewness measures the extent to which a distribution is not symmetric about its mean value, while the kurtosis measures the fatness of the tails of the distribution.

Testing Normality

- A normal distribution is symmetric (zero skewness) and has kurtosis equal to 3. The excess kurtosis is defined as $\kappa - 3$ (equal to zero for the normal distribution).
- A test of normality can be constructed by examining whether the skewness and the excess kurtosis of a distribution are jointly zero.
- Let $X_1, \ldots, X_T$ be a sample of size T. Then, the sample skewness and kurtosis are

$$\hat{\xi} = \frac{\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})^3}{\left(\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})^2\right)^{\frac{3}{2}}}, \quad \hat{\kappa} = \frac{\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})^4}{\left(\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})^2\right)^{2}}.$$

Testing Normality - The Jarque-Bera test

- The Jarque-Bera test is given by

$$JB = T\left[\frac{\hat{\xi}^2}{6} + \frac{(\hat{\kappa} - 3)^2}{24}\right]$$

- Under the null hypothesis of a normal distribution, the JB statistic asymptotically follows a $\chi^2(2)$ distribution.
- In the context of OLS regression, $\hat{\xi}$ and $\hat{\kappa}$ are computed using the regression residuals.
- Caution: for small sample sizes the test is not accurate (based on the asymptotic cut-off values) and rejects the null hypothesis too often (i.e., it has *actual* size higher than the *nominal* size).