

# ECON452: Week 1

Pier-André Bouchard St-Amant

January 12, 2012

## About...

1. This course is about applied time series analysis.
2. The goal is also to teach you to write meaningful empirical reports.
3. The syllabus can be found [here](#).

## A Review: OLS & GLS

1. We want to model a dependent variable ( $y$ ) as a function of independent variables ( $x_1, x_2, \dots, x_k$ ).
2. We *assume* that a linear relationship between  $y$  and the  $x$ 's is a good representation of the true relationship:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

3. The  $\beta_i$ s are the unknown multipliers that we seek to estimate and  $\epsilon$  is an error of approximation. Here only a constant and two variables are showed for simplicity.
4. What values of  $\beta_i$  does represent “best” the relationship we have assumed?

# A Review: OLS & GLS

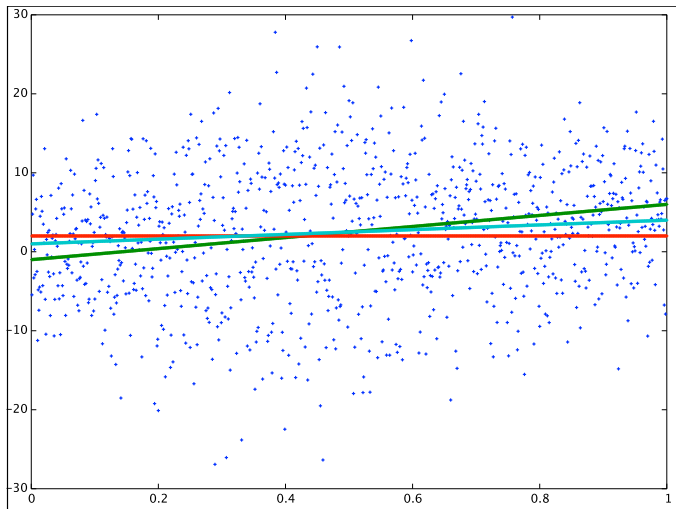


Figure: Which Line Fits Best?

# A Review: OLS & GLS

1. “Best” needs to be clarified: best according to what criterion?
2. The usual criterion is to minimize the estimation error ( $\epsilon$  close to zero).
3. “Close to zero” is not precise enough. By best, we mean to minimize the (sum of the) error terms squared.

## A Review: OLS & GLS

1. In order to do this, one must have a sample of  $n$  observations.
2. Each element  $i$  of the sample is actually a coherent observation of the variables  $y_i, x_{i1}, x_{i2}$ .
3. Thus, for each observation, we assume the linear relationship:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \epsilon_1 \quad (i=1)$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \epsilon_2 \quad (i=2)$$

$$\vdots = \vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \epsilon_n \quad (i=n)$$

# A Review: OLS & GLS

1. We then want the sum of all squared errors to be minimized. So we seek the set of  $\beta_0, \beta_1, \beta_2$  such that  $\sum_{i=1}^n \epsilon_i^2$  is minimized.
2. This is equivalent to:

$$\min_{\beta_0, \beta_1, \beta_2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

since  $\epsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}$ .

## A Review: OLS & GLS

1. This is a standard optimization problem. The function is convex, so there are unique values of  $\beta_0, \beta_1, \beta_2$  solving this problem.
2. We note this solution by adding “hats” to the betas. Thus  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$  is the solution to the problem (the “estimated betas”).
3. It obviously satisfies the first order conditions for optimization:

$$0 = \frac{\partial}{\partial \beta_j} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \quad j \in \{0, 1, 2\}$$

# A Review: OLS & GLS

1. These three first order conditions are:

$$0 = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) \quad (\hat{\beta}_0)$$

$$0 = -2 \sum_{i=1}^n x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) \quad (\hat{\beta}_1)$$

$$0 = -2 \sum_{i=1}^n x_{i2} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) \quad (\hat{\beta}_2)$$

2. All the three equations are almost identical. If we denote 1 by  $x_{i0}$ , we can summarize these three equations by:

$$0 = -2 \sum_{i=1}^n x_{ij} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) \quad \forall j$$

# A Review: OLS & GLS

1. Denote  $\hat{\epsilon}_i$  the estimated error terms. The previous equations gives us a few insights:

$$0 = -2 \sum_{i=1}^n x_{ij} \hat{\epsilon}_i \quad \forall j$$

2. For  $j = 0$ , the equation boils down to :  $\sum_i^n \hat{\epsilon}_i = 0$  (since  $x_{i0} = 1$ ). So if we divide by  $n$ , we get that the average error is zero. This means that :

$$\begin{aligned} \sum_{i=1}^n y_i &= \sum_{i=1}^n \left[ \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 + \cdots + \hat{\beta}_k x_{ik} + \hat{\epsilon}_i \right] \\ \Rightarrow \bar{y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \cdots + \hat{\beta}_k \bar{x}_k + 0 \end{aligned}$$

3. In other words, the estimated betas make no prediction error at the average of the sample.

# A Review: OLS & GLS

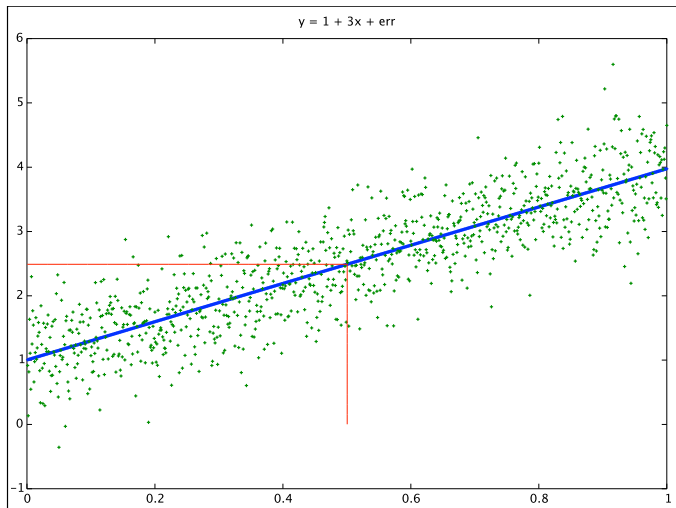


Figure: No approximation error at the sample average.

# A Review: OLS & GLS

1. For any other  $j$ , the equation can be rewritten as:

$$\begin{aligned} 0 &= \sum_{i=1}^n x_{ij} \hat{\epsilon}_i - \bar{x}_j \underbrace{\sum_{i=1}^n \hat{\epsilon}_i}_{=0} \\ &= \sum_{i=1}^n (x_{ij} - \bar{x}_j)(\hat{\epsilon}_i - 0) \quad \forall j \\ \Rightarrow 0 &= \text{cov}(x_j, \hat{\epsilon}) \end{aligned}$$

2. The sample covariance between the error term and the variable  $x_j$  (for any  $j$ ) is zero. E.g.: there is nothing left in  $\hat{\epsilon}$  that can be explained by  $x_j$ .

# A Review: OLS & GLS

1. But more can be said:

$$\begin{aligned}0 &= \sum_{i=1}^n x_{ij}(\hat{\epsilon}_i - 0) \\&= \sum_{i=1}^n x_{ij} \left( (y_i - \bar{y}) - \hat{\beta}_1(x_{i1} - \bar{x}_1) - \hat{\beta}_2(x_{i2} - \bar{x}_2) \right) \\&= \sum_{i=1}^n (x_{ij} - \bar{x}_j) \left( (y_i - \bar{y}) - \hat{\beta}_1(x_{i1} - \bar{x}_1) - \hat{\beta}_2(x_{i2} - \bar{x}_2) \right) \\ \Rightarrow \hat{\beta}_j &= \frac{\text{cov}(x_j, y) - \sum_{m \neq j} \hat{\beta}_m \text{cov}(x_j, x_m)}{\text{var}(x_j)}\end{aligned}$$

# A Review: OLS & GLS

1. This version of the equation reveals a lot:

$$\hat{\beta}_j = \frac{\text{cov}(x_j, y) - \sum_{m \neq j} \hat{\beta}_m \text{cov}(x_j, x_m)}{\text{var}(x_j)}$$

2. Denominator ( $\text{var}(x_j)^{-1}$ ): the smaller the sample variance in  $x_j$ , the bigger will be  $\beta_j$  in magnitude (precision).
3. Numerator:  $\beta_j$  is determined by the share of the covariance between  $x_j$  and  $y$  that cannot be explained by the covariance of other variables.
4. So  $\beta_j$  is pinned down by the variations in  $y$  that we know that comes for sure from  $x_j$  (other variations are discarded).

## Intuition (two variables)

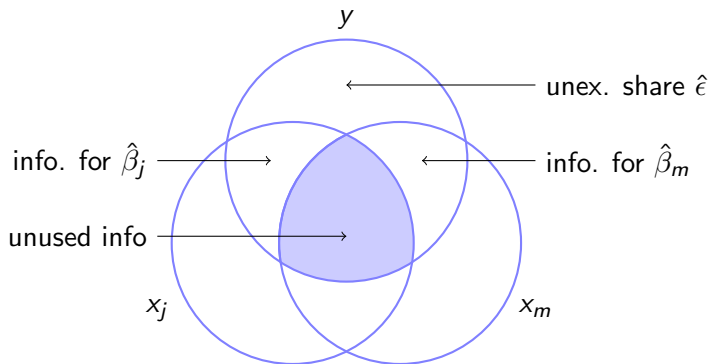


Figure: Intuitive Explanation of the OLS formula

## A Review: OLS & GLS

1. In short, the OLS estimator *projects* all the possible information of  $y$  on each variable.
2. What cannot be projected is the unexplained part  $\hat{\epsilon}$ .
3. One can clearly see with this intuitive picture the concept of *omitted variable bias* (board).

# A Review: OLS & GLS

The same analysis can be rewritten in matrix notation:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_y = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}}_{\equiv X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{\equiv \beta} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{\equiv \epsilon}$$
$$\Rightarrow y = X\beta + \epsilon$$

# A Review: OLS & GLS

1. Minimizing the sum of squares leads to  $\min_{\beta} \epsilon' \epsilon$  (where ' denotes the transpose).
2. First order conditions commands:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \hat{\beta}} (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= 2X'(y - X\hat{\beta}) \\ &\quad \text{(zero covariance condition)} \end{aligned}$$

$$\begin{aligned} \Rightarrow X'X\hat{\beta} &= X'y \\ \Rightarrow (X'X)^{-1}X'X\hat{\beta} &= (X'X)^{-1}X'y \\ \Rightarrow \hat{\beta} &= (X'X)^{-1}X'y \end{aligned}$$

3. This is the solution for estimated betas in matrix notation.

# A Review: OLS & GLS

To sum up:

1. OLS minimizes the sum of errors squared.
2. OLS makes no prediction error at the average of the sample.
3. The OLS formula is given by  $\hat{\beta} = (X'X)^{-1}X'y$ .
4. Only the variations that comes solely from variable  $j$  is used to pin down  $\hat{\beta}_j$ .

## Some Announcements

1. Homework/reports are to be dropped in the boxes on the second floor of DUN. They **close at 16:00** everyday.
2. Today, we split the group in two for laboratories. Group A will be on wednesdays, 10:00 while group B will be on thursdays, 8:30.

## A Review: OLS & GLS

1. So far, we have *not yet* specified the statistical behavior of the system. We have only specified what is the best prediction in a least square sense.
2. It would be interesting to have the statistical behavior. We could then test if coefficients are statistically meaningful, etc.
3. Most of the theory we use relies on the idea that if the sample is large enough, the estimated coefficients are normally distributed.
4. We will discuss two assumptions. The first one is covered in details (its simple). The second one is more realistic but will only be discussed.

# A Review: OLS & GLS

1. First assumption:  $X$  is known and  $\epsilon$  is NID (normally, independently distributed).
2.  $X$  is a set of variables we can control. For instance, a physicist can control every parameter in his laboratory: length, weight, etc.
3.  $\epsilon_i \sim N(0, \sigma^2)$  and every drawings are independent. This implies that  $\mathbb{E}[\epsilon_i \epsilon_j] = 0 \forall i \neq j$ .

# A Review: OLS & GLS

1. In this setup, this means that the estimated betas are also normally distributed.
2. Recall that if  $e \sim N(0, \sigma^2)$ , then  $ae + b \sim N(b, a^2\sigma^2)$ .
3. We thus have :

$$\begin{aligned} & N(0, \sigma^2) && \sim \epsilon \\ \Leftrightarrow & N(0, \sigma^2(X'X)) && \sim X'\epsilon \\ \Leftrightarrow & N(0, \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1}) && \sim (X'X)^{-1}X'\epsilon \\ \Leftrightarrow & N(\beta, \sigma^2(X'X)^{-1}) && \sim \beta + (X'X)^{-1}X'\epsilon \\ \Leftrightarrow & N(\beta, \sigma^2(X'X)^{-1}) && \sim (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon \\ \Leftrightarrow & N(\beta, \sigma^2(X'X)^{-1}) && \sim (X'X)^{-1}X'y \end{aligned}$$

## A Review: OLS & GLS

1. Hence, in this framework, the estimated betas have a normal distribution  $N(\beta, \sigma^2(X'X)^{-1})$ .
2. As long as  $X$  is controlled, this is an exact result.
3. We can then use this to perform inference.

## A Review: OLS & GLS

1. Major problem: in economy, variables contained in  $X$  are not controlled.
2.  $X$  contains usually, wages, GDP, inflation and other variables that the experimenter cannot control.
3.  $X$  has a statistical life of its own: previous assumptions are invalid!
4. We thus need other assumptions. The basic idea is that if  $n$  is “large enough”,  $\frac{X'X}{n}$  would behave as if it was a fixed matrix.
5. Large enough: this is an *asymptotic theory*: results holds exactly when  $n \rightarrow \infty$ . Otherwise, we assume we are “close enough” to have a good approximation.

## A Review: OLS & GLS

1. Asymptotic theory tells us that as  $n$  gets large  $\hat{\beta}$  will be arbitrarily close to  $\beta$ .
2. In practice this means that  $\text{var}\hat{\beta}$  is asymptotically equal to  $\sigma^2(X'X)^{-1}$ .
3. So although the theoretical framework is different, we end-up with the same type of inference in practice (and hope  $n$  is “large enough”).

## A Review: OLS & GLS

1. Note that the OLS estimator has, by design, the property that  $0 = X'\hat{\epsilon}$ . This implies in particular that  $\sum_{i=1}^n \hat{\epsilon} = 0$ .
2. If you look at the statistical assumptions we made, they imply that  $\mathbb{E}\epsilon = 0$  and  $\mathbb{E}X'\epsilon = 0$  as well.
3. This match between the conditions of the estimator and the statistical equivalent leads to unbiasedness.
4. The fact that there is no particular way to predict the error term (NID or IID) leads to efficiency. Amongst the class of unbiased linear estimator, OLS is the best one.

# GLS: An Application

1. Generalized least squares relax the idea that all error terms should bear the same weight.
2. If past errors can help predict current errors (autocorrelation), the performance of OLS can be improved by adding weight to past observations.
3. Likewise, if some observations are more likely/precise than others (heteroskedasticity), they should bear more weight as well.
4. In terms of mathematics, the problem is minimizing the weighted sum of squares:

$$\hat{\beta}_{GLS} \equiv \arg \max_{\beta_0, \dots, \beta_k} \sum_{j=1}^n \sum_{i=1}^n \epsilon_i \omega_{ij} \epsilon_j$$

5. The question is then: what is the optimal structure for the weights  $\omega_{ij}$ ?

# GLS: An Application

1. It depends a lot on the information at hand.
2. If we know the structure of the error term process,  $\Omega$  can be exact.
3. Otherwise, it must be estimated through iterated (two stage) least squares.

# GLS: An Application

1. Most of the times, surveys taken from a population are not “random” .
2. Efficiency of the data is often of concern: for instance, we might want to make sure that there is a “large enough” sample of students in a general survey about Canadians.
3. Hence, students are not taken at random, but “oversampled” relatively to their weight in the population.
4. Such oversampling is afterwards corrected by a weight, the sampling weight.
5. We can then use weighted samples to perform inference:

$$\omega_{ij} = \begin{cases} p_{ij} & \text{(weight of the observation)} \\ 0 & \text{if } i \neq j \end{cases}$$

## GLS: An Application

1. All Surveys in Statistics Canada contain weights because their sampling technique is cost-efficient (they use *strata*).
2. Let's say we want to estimate the future wage of postgraduate students in Québec: cégeps graduates vs. university graduates.
3. A very simple way to do this is to perform the following *mincer* regression:

$$\begin{aligned} \ln(\text{wage}) = & \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + \\ & \beta_3 \mathbb{1}_{\text{CÉGEP grad}} + \beta_4 \mathbb{1}_{\text{univ grad}} + \\ & \beta_5 \mathbb{1}_{\text{CÉGEP grad}} \text{age} + \beta_6 \mathbb{1}_{\text{CÉGEP grad}} \text{age}^2 \\ & \beta_7 \mathbb{1}_{\text{univ grad}} \text{age} + \beta_8 \mathbb{1}_{\text{univ grad}} \text{age}^2 \\ & + \epsilon_i \end{aligned}$$

## GLS: An Application

1. The variable  $\mathbb{1}_{\text{CÉGEP grad}}$  is the indicator function of CÉGEP graduation: it takes the value one if the observation is a college graduate and zero otherwise. The variable  $\mathbb{1}_{\text{univ grad}}$  is defined likewise.
2. Age is the age of each observation,  $\text{age}^2$  is the age square. Age is introduced as a proxy for experience. The quadratic form is to approximate diminishing returns. It is thus expected that  $\beta_1 + \beta_2 \text{age} > 0$  (increasing returns) and  $\beta_2 < 0$  (diminishing marginal returns).
3. Other terms capture the combined effect of age and experience together.
4. Such variables do not exist directly in the survey, they must be *derived* (coded) from other variables.
5. Once this is done, the regression can be performed.

# GLS: An Application

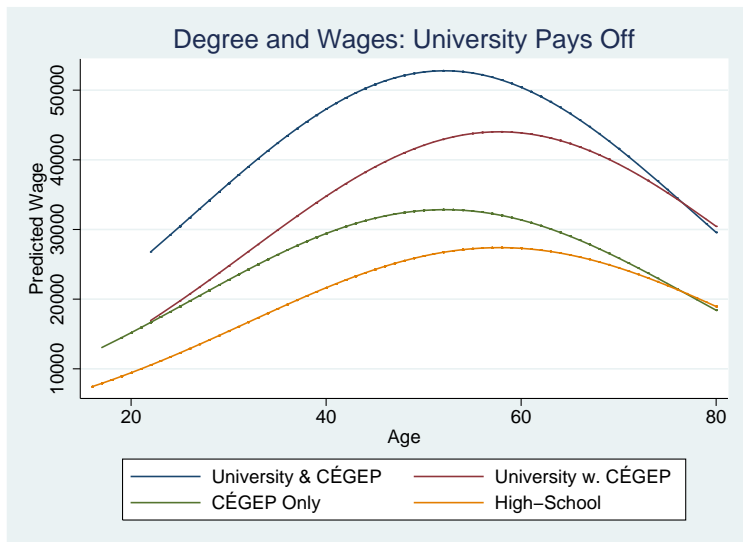


Figure: A Simple Application of GLS To Predict Wages With Degrees.

# Statistical Tests

1. Recall that estimated betas is given by  $\hat{\beta} = (X'X)^{-1}X'y$  and that the estimated variance-covariance matrix is given by :

$$\text{cov}(\hat{\beta}) = \underbrace{\frac{\hat{\epsilon}'\hat{\epsilon}}{n-k}}_{=\hat{\sigma}^2} (X'X)^{-1}$$

2. How can we use this to perform statistical inference?
3. There are two tests that we will use in practice, depending on the situation: the *Wald* ( $W$ ) and *Likelihood Ratio* ( $LR$ ) tests.
4. Asymptotically, they all yield the same answer. In finite sample, we usually have that  $W < LR$ .
5. We will first see a particular case of the Wald test, namely testing the t-test for a single parameter.

# Testing A Single Parameter

1. Assume we have estimated the regression model  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$  and found some values  $\hat{\beta}_i$  and the variance  $\text{var}(\hat{\beta}_i)$ .
2. If we know  $\sigma^2$ , that  $X$  are fixed regressors and that error terms are normal, we can form the  $z$  statistic:

$$z \equiv \frac{\hat{\beta}_i - \beta_{H_0}}{\sqrt{\text{var}(\hat{\beta}_i)}}$$

where the variance is given by the  $i$ -th element of the diagonal of  $\sigma^2(X'X)^{-1}$ .

3. We can then compare this value with the critical one of a normal distribution.
4. In practice, however,  $\sigma^2$  is not known. To account for this uncertainty, we compare against a Student (or  $t$ ) distribution.

# Testing A Single Parameter

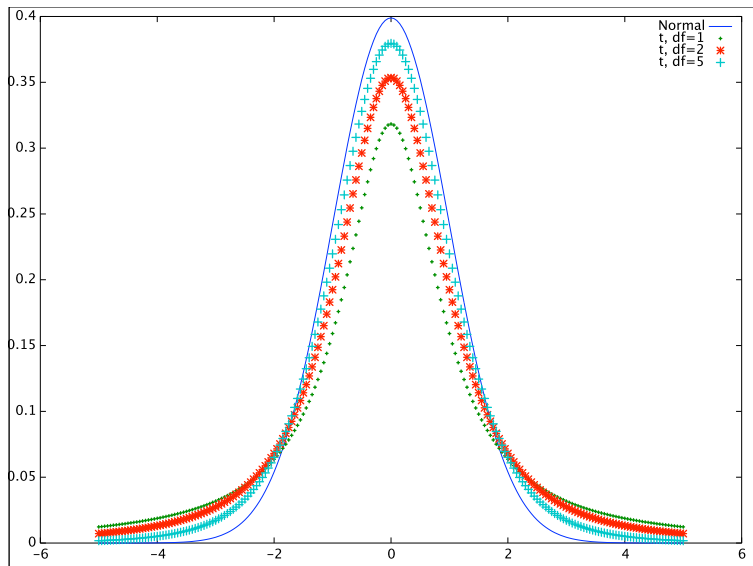


Figure: Normal vs t distribution for various degrees of freedom.

# Testing A Single Parameter

1. Both the  $t$  and normal distributions are “bell curved”.
2. The  $t$  distributions have ticker tails: they put higher probability weights away from the mean than the normal does.
3. As the degrees of freedom gets large, the  $t$  distribution gets closer and closer to the normal distribution.
4. There is no practical difference for degrees of freedom greater than 1500.
5. This is because the estimated variance gets close to the true variance with the sample size.

# Testing A Single Parameter

1. Assume we have estimated the regression model  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$  on a sample of  $n = 100$  and found the following values:

$$\hat{\beta} = \begin{bmatrix} 1.1 \\ 0.5 \\ -0.4 \end{bmatrix} \quad \hat{\sigma}^2(X'X)^{-1} = \begin{bmatrix} 0.1 & -0.01 & 0 \\ -0.01 & 0.1 & 0.01 \\ 0 & 0.01 & 0.1 \end{bmatrix}$$

2. We want to test if  $\hat{\beta}_0$  (1.1) is significantly different than one  
 $H_0 : \hat{\beta}_0 = 1$  with a level of 95% against the alternative  
 $H_1 : \hat{\beta}_0 \neq 1$ .
3. This is a two-sided test. We then form the  $t$  statistic:

$$t = \frac{1.1 - 1}{\sqrt{0.1}} = 0.31.$$

4. There are 3 estimated parameters and thus  $100 - 3 = 97$  degrees of freedom. The  $t$  statistic 0.31 is well below the 95% level of  $\approx 1.985$ . Hence, we cannot reject  $H_0$ .

# Testing A Single Parameter

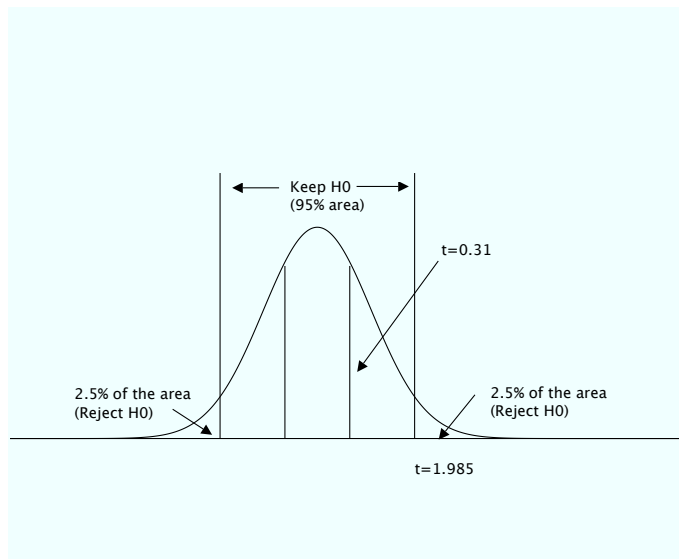


Figure: The 95% level defines a critical region (line).

## Next Week

1. We will begin laboratories wednesdays with the first half of the class and thursdays for the second half.
2. Tutorial: introduction to Stata and where to get variables.
3. Next week: we dive in time-series as well as testing more than one coefficient.