

# 2016 Essex Summer School

## 3C: Pooled Time Series Cross-Sectional Analysis

Robert W. Walker, Ph. D.

Associate Professor of Quantitative Methods  
Atkinson Graduate School of Management  
Willamette University  
Salem, Oregon USA  
[rwalker@willamette.edu](mailto:rwalker@willamette.edu)

August 6, 2019

2016 Essex *SS<sup>2</sup>DA*: Pooled CSTS/TSCS

# Course Outline

- Who am I? Why am I here? What will I learn?
- How will we do this?
- Information on backgrounds
- Discussion of syllabus

# Preliminaries

Reader: You have a course reader that contains the essential articles for the course. The syllabus outlines issues from textbooks that have more detailed foundations. One can learn a great deal from walking through fine details.

Questions: Ask them. Please.

Diction: I speak (VERY) quickly. Remind me when this is excessive (ALWAYS?); I will undoubtedly forget and this doesn't help anyone (including me as I burn through slides).

## The course: Pooled Data

- Data analysis is all about variation – random variables.
- We usually want parameters that usefully describe that variation.
- Sources of variation are our key concern and their influence on our parameters.  
This comes in two classes.
  - Uncorrelated with regressors.
  - Correlated with regressors.
- Pooling admits two obvious sources of variation:  $i$  and  $t$  (however defined).
- All our concerns follow from this.

# Data Examples

- Bond ratings among U. S. states.
- US Supreme Court justices/votes
- Wars/Military conflicts
- OECD countries/global samples in political economy.
- FDI inflows/outflows, etc.
- Voter turnout across nations or US states/counties/municipalities.

# Panels and Related Data Structures

- Balanced or unbalanced panels?
- Panels or rolling cross-sections?
- Asymptotics
  - Time dominant?
  - Unit dominant?
  - Both?
- Substantive question?  
What's your substantive interest in panel data?

## Before We Start, A Review?

- Matrix Algebra

What are matrices?

Matrix multiplication/addition/etc.

Matrix Inversion?

Trace, Kronecker Product, Eigenvalues, Determinants, and the like

- Statistics – Expectations and Normal,  $\chi^2$ ,  $t$ , and  $F$

- Linear Regression

What assumptions are necessary for the OLS estimator to be unbiased?

How about the asymptotic variance to be correct?

What about the Gauss-Markov conditions?

## Why Matrices?

- A matrix is the natural structure for panel data/CSTS/TSCS.
- Individuals/households/countries form rows.
- Time points form columns.



## On Matrices

A matrix is a rectangular array of real numbers. If it has  $m$  rows and  $n$  columns, we say it is of dimension  $m \times n$ .

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

A vector  $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$  can be thought as as a matrix with  $n$  rows and one column or as a matrix with one row and  $n$  columns.

# Products

Periodicially, we will wish to make use of two types of vector products.

1. Inner (or dot) product Let  $\mathbf{u} = (u_1, \dots, u_n)$  and  $\mathbf{v} = (v_1, \dots, v_n)$  be two vectors in  $\mathbb{R}^n$ . The **Euclidean inner product** of  $\mathbf{u}$  and  $\mathbf{v}$ , written as  $\mathbf{u} \cdot \mathbf{v}$ , is the number

$$\mathbf{u} \cdot \mathbf{v} = u_1v_1 + u_2v_2 + \dots + u_nv_n$$

which can also be written  $\mathbf{u}'\mathbf{v}$

2. Outer product  $\mathbf{u}\mathbf{v}'$

## Matrix Inversion

There are two primary methods for inverting matrices. The first is often referred to as Gauss-Jordan elimination and the second is known as Cramer's rule. The former involves a series of elementary row operations undertaken on the matrix of interest and equally to  $I$  while the latter relies on a determinant and the adjoint of the matrix of interest.

Let  $A$  and  $B$  be square invertible matrices. It follows that:

- $(A^{-1})^{-1} = A$
- $(A^T)^{-1} = (A^{-1})^T$
- $AB$  is invertible and  $(AB)^{-1} = B^{-1}A^{-1}$ .

For a matrix  $A$ , the following are equivalent:

1.  $A$  is invertible.
2.  $A$  is nonsingular.
3. For all  $y$ ,  $Ax = y$  has a unique solution.
4.  $\det A \neq 0$  and  $A$  is square.

# Definiteness

Given a square matrix  $A$  and a vector  $\mathbf{x}$ , we can claim that

- $A$  is negative definite if,  $\forall \mathbf{x} \neq 0, \mathbf{x}^T A \mathbf{x} < 0$
- $A$  is positive definite if,  $\forall \mathbf{x} \neq 0, \mathbf{x}^T A \mathbf{x} > 0$
- $A$  is negative semi-definite if,  $\forall \mathbf{x} \neq 0, \mathbf{x}^T A \mathbf{x} \leq 0$
- $A$  is positive semi-definite if,  $\forall \mathbf{x} \neq 0, \mathbf{x}^T A \mathbf{x} \geq 0$

Brief diversion on principal submatrices and (leading) principal minors toward a sufficient condition for characterizing definiteness.

NB: By the way, I forgot to mention that the trace of a square matrix is the sum of its diagonal elements.

# Enough Matrices: To Statistics

We needed this to:

- Determine invertibility and the relation to definiteness.
- A matrix version of the Cauchy-Schwartz inequality sets the characteristics of variance/covariance matrices for linear regression problems (avoiding equality).
- To familiarize the idea because error matrices are nice to think on.
- Linking time series properties through the previous to invertibility is key. Especially because of block inversion



## Random Variables, Expectations, etc. etc.

- Random Variables: Real-valued function with domain: a sample space.
- Mean (Expected Value):  $E[x] = \int_x x f(x) dx$  or  $E[x] = \sum_x x p(x)$
- Variance (Spread):  $V[x] = E(x^2) - [E(x)]^2$
- Covariance:  $E[xy] = E[(x_i - [E(x)])(y_i - [E(y)])]$
- Correlation:  $\rho = \frac{E[(x_i - [E(x)])(y_i - [E(y)])]}{\sigma_x \sigma_y} = \frac{Cov(XY)}{\sqrt{\sigma_x^2} \sqrt{\sigma_y^2}}$
- Variance of Linear Combination is  $\sum_i \sum_j a_i a_j Cov(X_i X_j)$

## Some (Loosely Stated) Distributional Results

- Self-reproducing property of  $N$
- The implications of Basu's Theorem (Independence of Mean and Variance of Normals)
- $N^2 \sim \chi^2$
- $\frac{\chi_m^2}{\chi_n^2} \sim F_{m,n}$
- $t$  is given by  $\frac{N}{\sqrt{\frac{\chi^2}{v}}}$

## Normal random variables

A random variable  $X$  has a normal distribution with mean  $\mu$  and variance  $\sigma^2$  ( $\mu \in \mathbb{R}$  and  $\sigma^2 \in \mathbb{R}^{++}$ ) if  $X$  has a continuous distribution for which the probability density function (p.d.f.)  $f(x|\mu, \sigma^2)$  is as follows (for finite  $x$ ):

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right]$$

- If  $X \sim N(\mu, \sigma^2)$ , Let  $Y = aX + b$  ( $a \neq 0$ ).  $Y \sim N(a\mu + b, a^2\sigma^2)$ .
- $Z \sim N(0, 1)$  allow the percentiles for any normal:  $Z = \frac{x - \mu}{\sigma}$ .
- Sums of (independent) normals are normal.
- Sums of affine transformations of (independent) normals are normal.

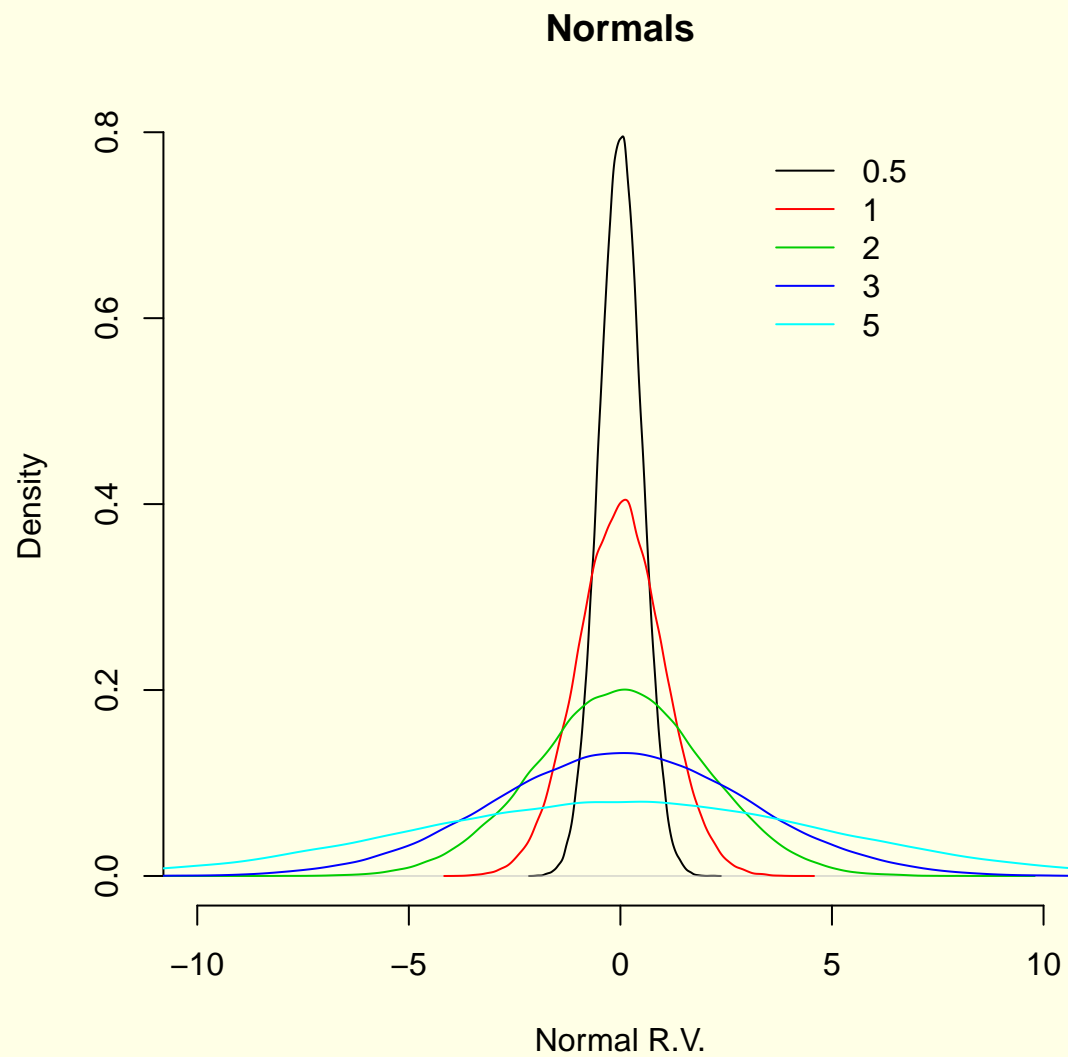


Figure 1: The Normal distribution (with varying  $\sigma^2$ )

## Basu's Theorem

Let  $X_1, X_2, \dots, X_n$  be independent, identically distributed normal random variables with mean  $\mu$  and variance  $\sigma^2$ . With respect to  $\mu$ ,

$$\hat{\mu} = \frac{\sum X_i}{n},$$

the sample mean is a complete sufficient statistic – it is informationally optimal to estimate  $\mu$ .

$$\hat{\sigma}^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1},$$

the sample variance, is an ancillary statistic – its distribution does not depend on  $\mu$ . These statistics are independent (also can be proven by Cochran's theorem). This property (that the sample mean and sample variance of the normal distribution are independent) characterizes the normal distribution; no other distribution has this property.

## $\chi^2$ random variables

If a random variable  $X$  has a  $\chi^2$  distribution with  $n$  degrees of freedom then the probability density function of  $X$  (given  $x > 0$ ) is

$$f(x) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{(n)}{2}-1} \exp\left(\frac{-x}{2}\right)$$

Two key properties:

1. If the random variables  $X_1, \dots, X_k$  are independent and if  $X_i$  has a  $\chi^2$  distribution with  $n_i$  degrees of freedom, then  $\sum_{i=1}^k X_i$  has a  $\chi^2$  distribution with  $\sum_{i=1}^k n_i$  degrees of freedom.
2. If the random variables  $X_1, \dots, X_k$  are independent and,  $\forall i : X_i \sim N(0, 1)$ , then  $\sum_{i=1}^k X_i^2$  has a  $\chi^2$  distribution with  $k$  degrees of freedom.

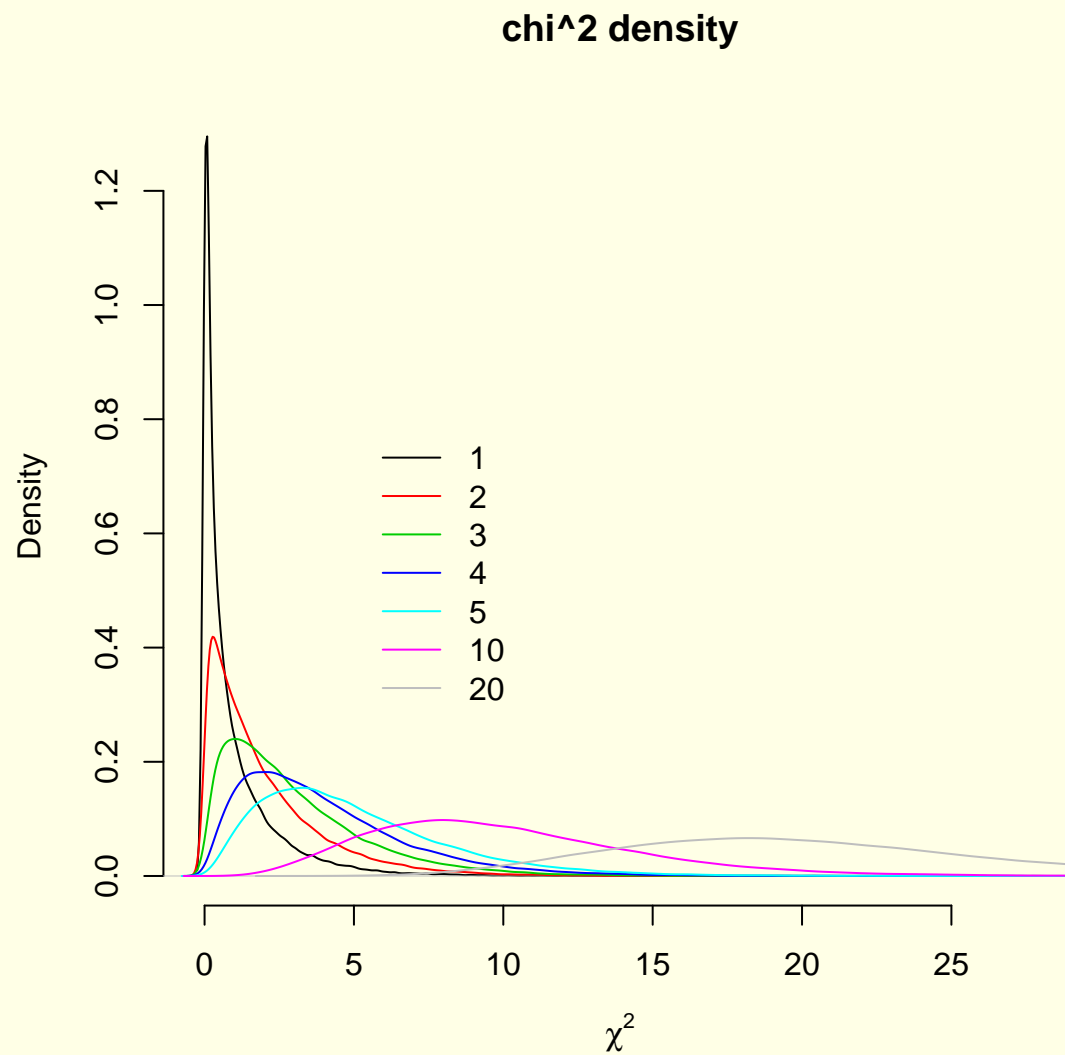


Figure 2: The  $\chi^2$  distribution

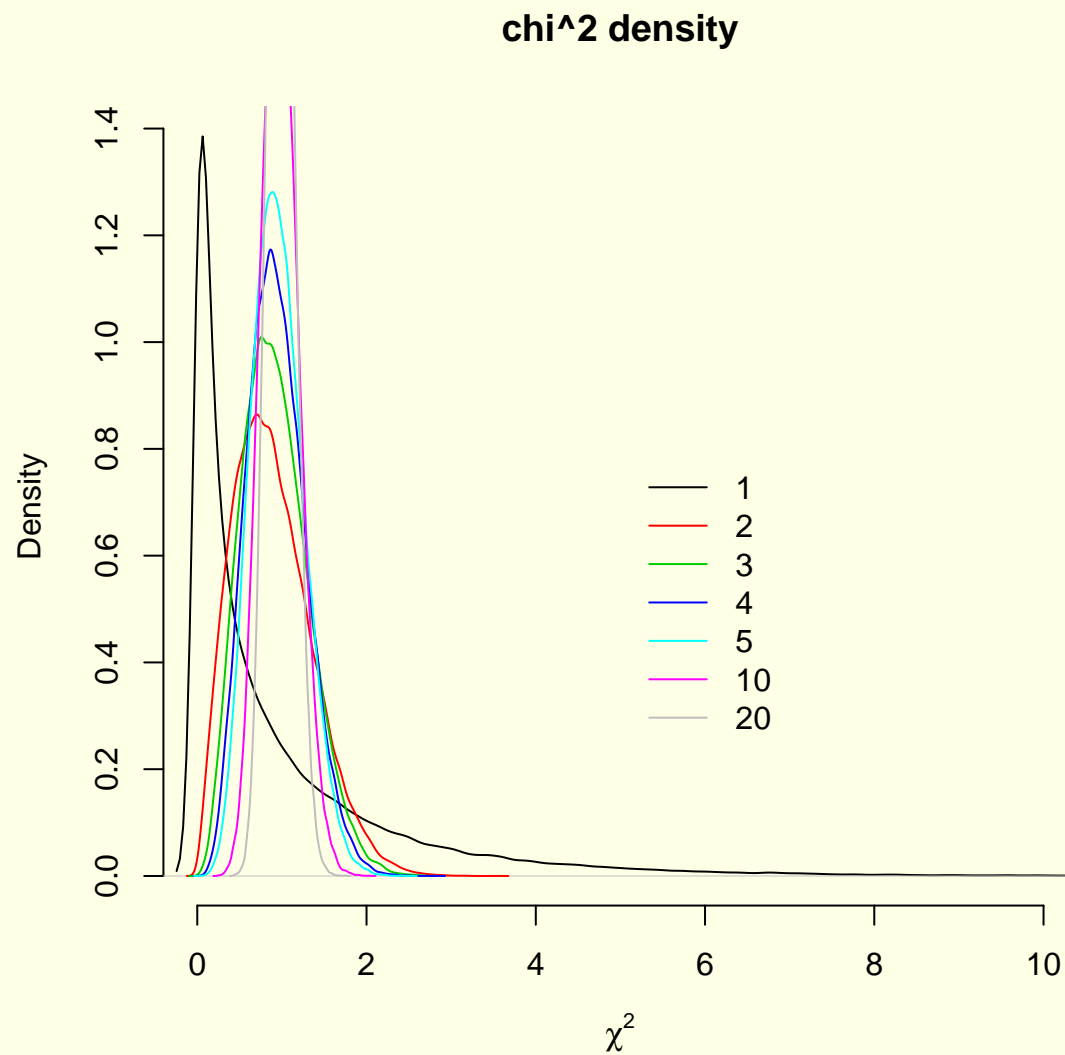


Figure 3: The  $\chi^2$  distribution (/n root)



## Student's $t$ distributed random variables

Consider two independent random variables  $Y$  and  $Z$ , such that  $Y$  has a  $\chi^2$  distribution with  $n$  degrees of freedom and  $Z$  has a standard normal distribution. If we define

$$X = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

then the distribution of  $X$  is called the  $t$  distribution with  $n$  degrees of freedom. The  $t$  has density (for finite  $x$ )

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{(n\pi)^{\frac{1}{2}}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{(n+1)}{2}}$$

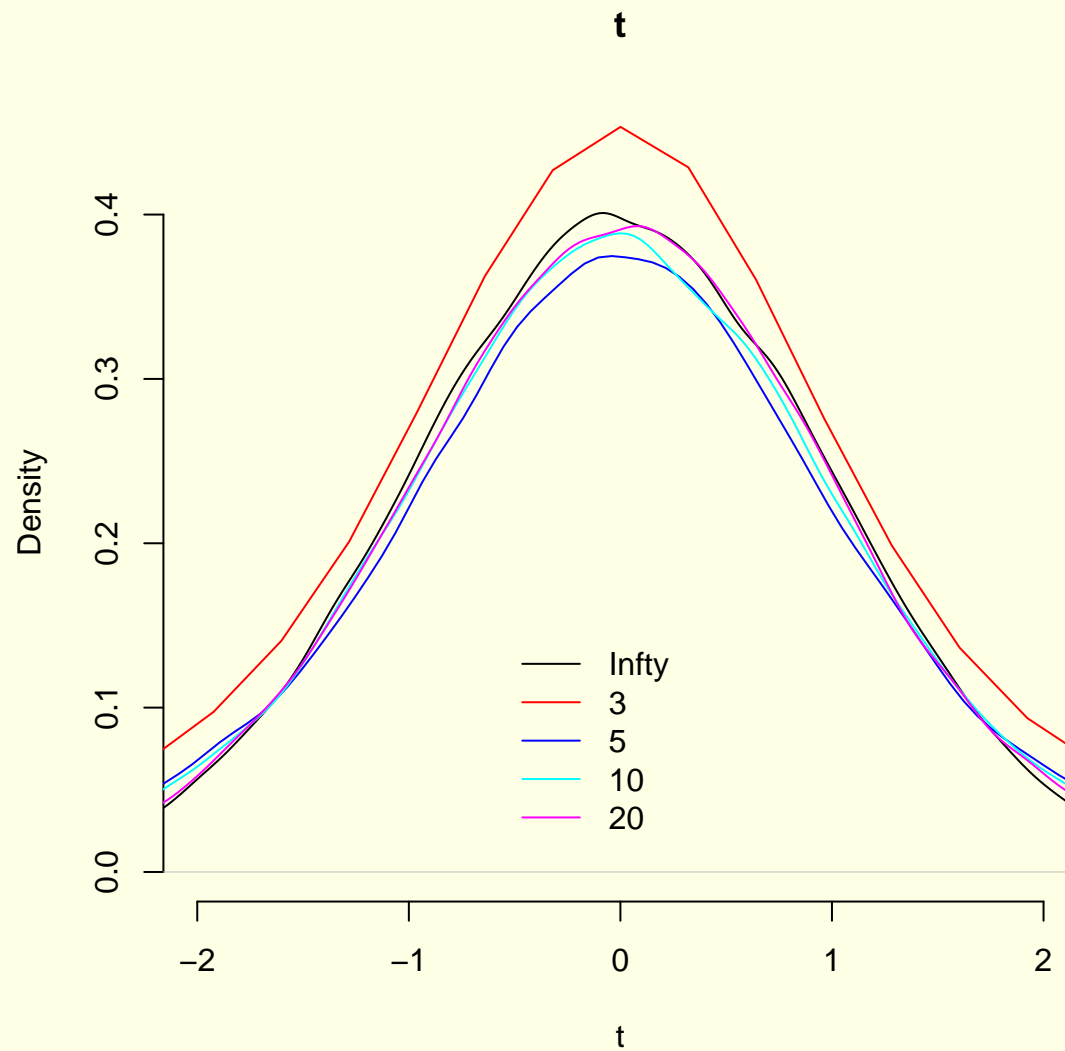


Figure 4: The  $t$  distribution

## $F$ distributed random variables (Variance-Ratio)

Consider two independent random variables  $Y$  and  $W$ , such that  $Y$  has a  $\chi^2$  distribution with  $m$  degrees of freedom and  $W$  has a  $\chi^2$  distribution with  $n$  degrees of freedom, where  $m, n \in \mathbb{R}^{++}$ . We can define a new random variable  $X$  as follows:

$$X = \frac{\frac{Y}{m}}{\frac{W}{n}} = \frac{nY}{mW}$$

then the distribution of  $X$  is called an  $F$  distribution with  $m$  and  $n$  degrees of freedom.

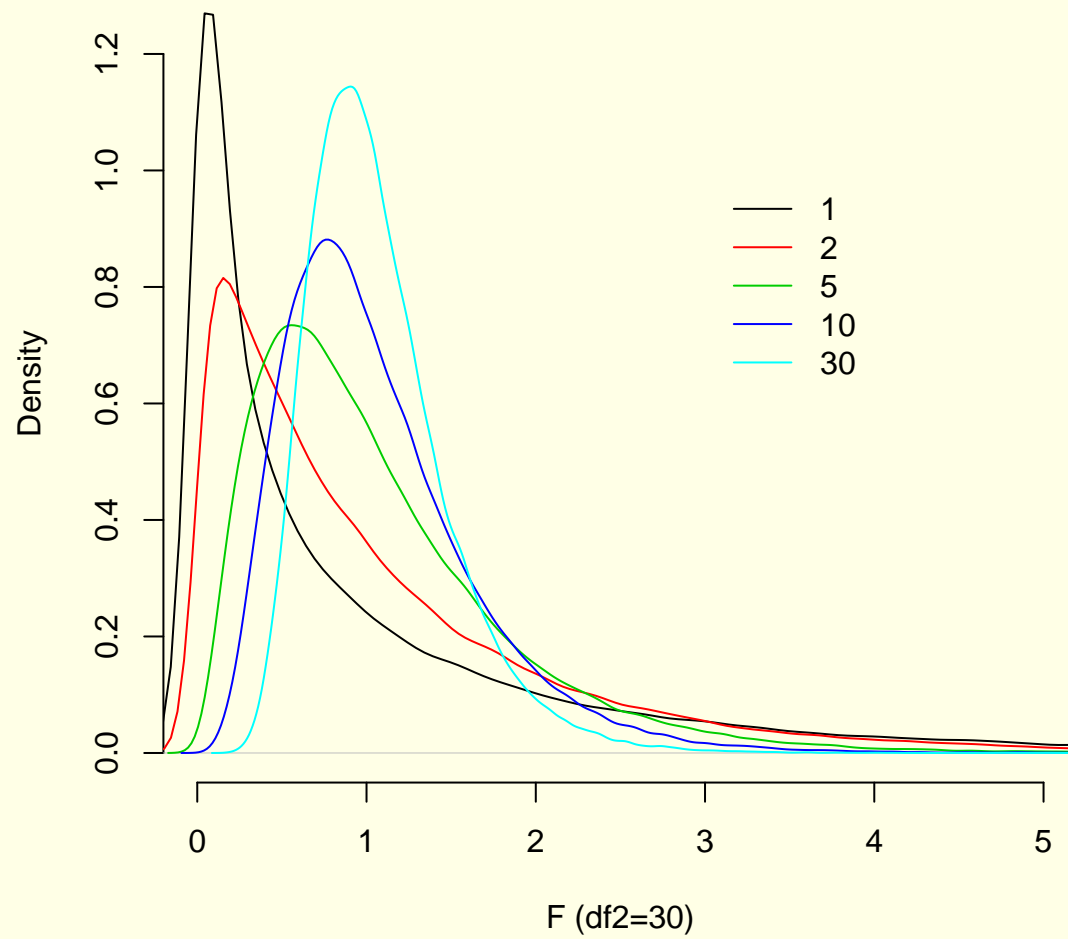


Figure 5: The  $F$  distribution: Numerator

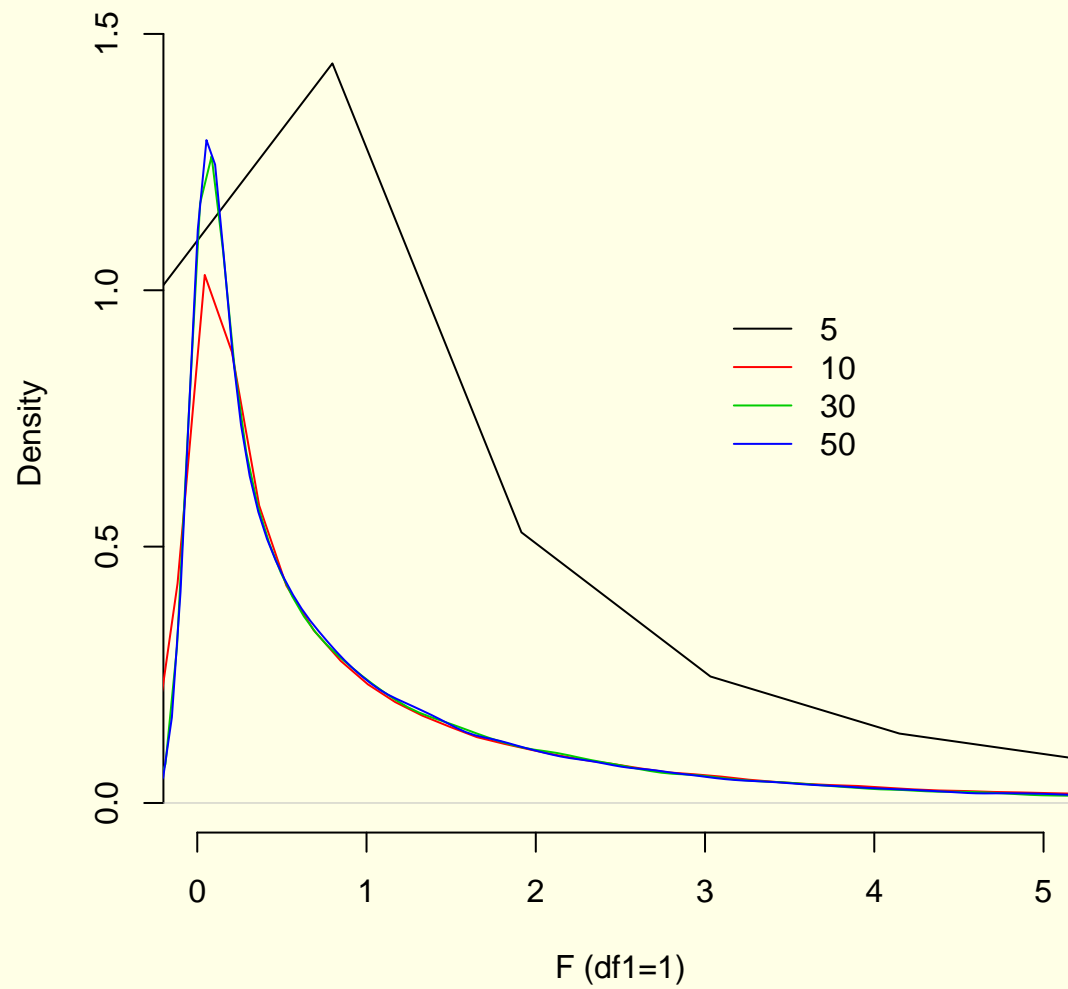


Figure 6: The  $F$  distribution: Denominator

## Distributions and Matrices

This gets us through the background to this point. We will invoke parts of this as we go from here. We have matrices and their inverses. We have some distributional results that link the normal,  $\chi^2$ ,  $t$ , and  $F$ . We have a theorem regarding separable moments and the normal. This gives us the intuition for Gauss-Markov. Nevertheless, let's begin the meat of it all: regression.

# The Orthogonal Projection

One of the features of the Ordinary Least Squares Estimator is the orthogonality (independence) of the estimation space and the error space.

- $E[\hat{e}_i \hat{y}_i] = 0$
- $E[\hat{e}_i x_i] = 0$

## OLS: Assumptions

1. Linearity:

$$y = X\beta + \epsilon$$

2. Strict Exogeneity

$$E[\epsilon|X] = 0$$

3. No [perfect] multicollinearity:

Rank of the  $N \times K$  data matrix  $X$  is  $K$  with probability 1 ( $N > K$ ).

4.  $X$  is a nonstochastic matrix.

5. Homoskedasticity

$$E[\epsilon\epsilon'] = \sigma^2 I \text{ s.t. } \sigma^2 > 0$$



## Returning to the Regression Model

Now we want to reexamine the minimization of the sum of squared errors in a matrix setting. We wish to minimize the inner product of  $\mathbf{e}$ .

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (1)$$

$$= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \quad (2)$$

Take the derivative, set it equal to zero, and solve....

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \therefore \quad (3)$$

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta} \quad (4)$$

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (5)$$

## Properties of OLS Estimators

- Unbiasedness  $E[\hat{\beta} - \beta] = 0$
- Variance  $E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$
- The Gauss-Markov Theorem – Minimum Variance Unbiased Estimator

## The first two

Need nothing about the distribution other than the two moment definitions. It is for number three that this starts to matter and, in many ways, this is directly a reflection of Basu's theorem.

## Unbiasedness

With  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$ ,  $\mathbb{E}[\hat{\beta} - \beta] = 0$  requires,

$$\mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y - \beta] = 0$$

We require an inverse already. Invoking the definition of  $y$ , we get

$$\mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon) - \beta] = 0 \quad (6)$$

$$\mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon - \beta] = 0 \quad (7)$$

Taking expectations and rearranging.

$$\beta - \beta = -\mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon] \quad (8)$$

If the latter multiple is zero, all is well.

## Variance

$\mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$  can be derived as follows.

$$\mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon - \beta)((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon - \beta)'] \quad (9)$$

$$\mathbb{E}[(\mathbf{I}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon - \beta)(\mathbf{I}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon - \beta)'] \quad (10)$$

Recognizing the zero part from before, we are left with the manageable,

$$\mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \quad (11)$$

Nifty. With nonstochastic  $\mathbf{X}$ , it's the structure of  $\epsilon\epsilon'$  and we know what that is. By assumption, we have  $\sigma^2\mathbf{I}$ . If stochastic, we need more steps to get to the same place.

Proving the Gauss-Markov theorem is not so instructive. From what we already have, we are restricted to linear estimators, we add or subtract something. So after computation, we get the OLS standard errors plus a positive semi-definite matrix. OLS always wins. From here, a natural place to go is corrections for non I. We will do plenty of that. And we will eventually need Aitken.

Beyond this, lets take up two special matrices (that will be your favorite matrices):

1. Projection Matrix **P**:  $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
2. Residual Maker **M**:  $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

which are both symmetric and idempotent ( $\mathbf{M}^2 = \mathbf{M}$ ).

## On M and P

### 1. $M$

$$M = I - X(X'X)^{-1}X' \quad (12)$$

$$My = (I - X(X'X)^{-1}X')y \quad (13)$$

$$My = Iy - X \underbrace{(X'X)^{-1}X'y}_{\hat{\beta}} \quad (14)$$

$$My = y - X\hat{\beta} \quad (15)$$

$$My = \hat{\epsilon} \quad (16)$$

## 2. $\mathbf{P}$

$$\mathbf{P} = \mathbf{I} - \mathbf{M} \quad (17)$$

$$\mathbf{P} = \mathbf{I} - (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \quad (18)$$

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (19)$$

$$\mathbf{P}\mathbf{y} = \mathbf{X} \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}_{\hat{\beta}} \quad (20)$$

$$\mathbf{P}\mathbf{y} = \mathbf{X}\hat{\beta} \quad (21)$$

$$\mathbf{P}\mathbf{y} = \hat{\mathbf{y}} \quad (22)$$



# General Linear Regression Model: Multiple Regression

$$y = X\beta + \epsilon$$

1.  $X$  is a  $n \times k$  matrix of regressors (where the first column is 1)
  2.  $\beta$  is a  $k \times 1$  matrix of unknown (partial) slope coefficients
  3.  $\epsilon$  is an (unknown) residual
- A partial slope is the analog of our simple bivariate regression except that there are multiple regressors.
  - Partial: it is the effect of  $x_k$  when all other variables are held constant.

# Regression and Inference

$$y = X\beta + \epsilon$$

1. Linear regression model:

- Same tools for testing hypotheses
- Tests work for multiple partial slopes

2.  $t$  can test the hypothesis of no relationship between  $x_k$  and  $y$ ;  $\beta_k = 0$

3.  $F$  can compare the fit of two models or the joint hypothesis that  $\beta_1 = \beta_2 = \dots = \beta_k = 0$ .

The latter  $F$  test appears in standard regression output and the probability compares a model with only a constant to the model you have estimated with

$H_0$  : Constant only.

If we just like  $F$  better than  $t$ , it happens that  $(t_\nu)^2 \sim F_{1,\nu}$ , but  $t$  has the advantage of being (potentially) one-sided.

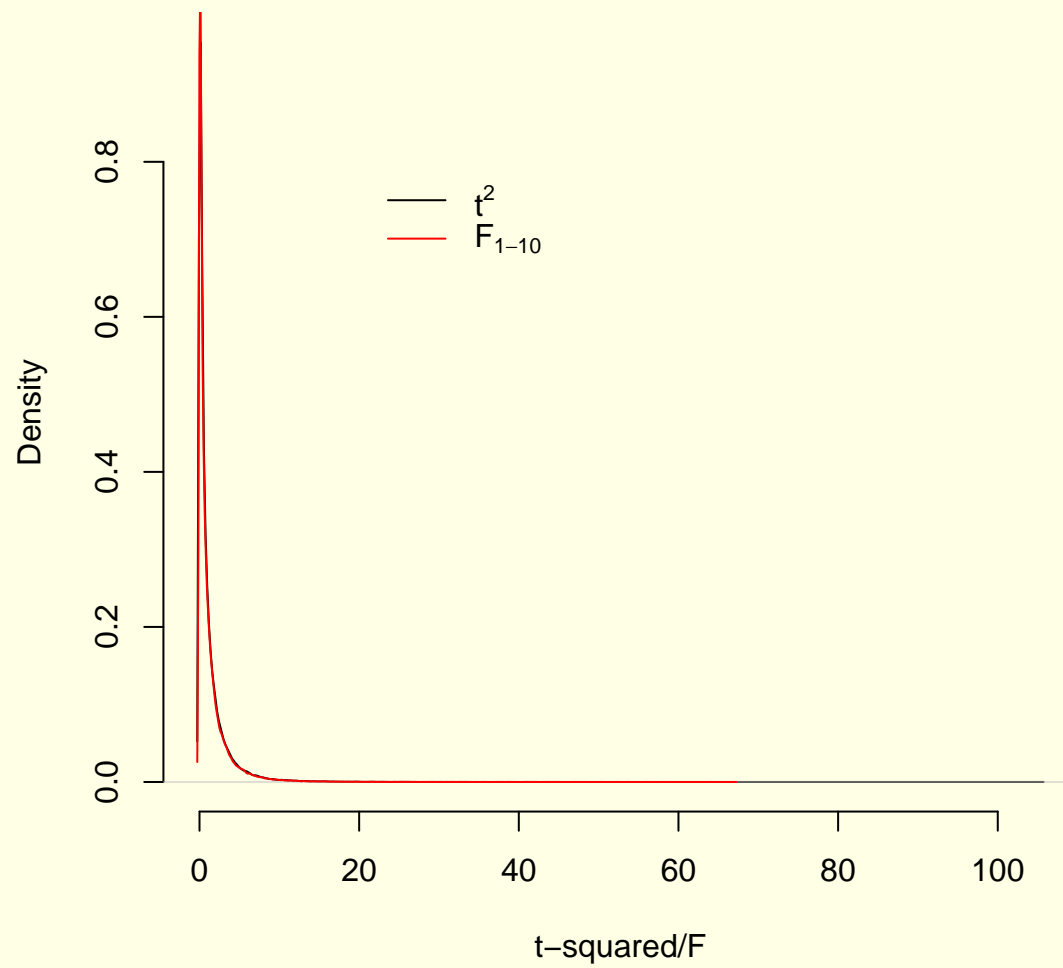


Figure 7:  $t$  and  $F$  distributions for Simple Hypotheses ( $t_{10}, F_{1,10}$ )

## Confidence Intervals

In forming confidence intervals, one must account for metrics.  $t$  is defined by a standard deviation metric and the standard deviation remains in a common metric with the parameter  $\hat{\beta}$ . Variance represents a squared metric in terms of the units measured by  $\hat{\beta}$ . As a result, we will form confidence intervals from standard deviations instead of variances.

- Prediction intervals:

- ★ A future response:

$$\hat{y}_0 \pm t_{n-p}^{(\frac{\alpha}{2})} \hat{\sigma} \sqrt{1 + x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0}$$

- ★ A mean response:

$$\hat{y}_0 \pm t_{n-p}^{(\frac{\alpha}{2})} \hat{\sigma} \sqrt{x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0}$$

- $\hat{\beta}$  Confidence Intervals

- ★ Individual:

$$\hat{\beta} \pm t_{n-p}^{(\frac{\alpha}{2})} \hat{\sigma} \sqrt{(\mathbf{X}'\mathbf{X})_{ii}^{-1}}$$

Unfortunately, unless the off-diagonal elements of the variance/covariance matrix of the estimates are zero, individual confidence intervals are/can be deceptive. A better way is to construct a simultaneous confidence interval.

- ★ Simultaneous for  $k$  regressors:

$$(\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta) \leq k \hat{\sigma}^2 F_{k, n-k}^{(\alpha)}$$

## Omitted Variable Bias

Suppose that a correct specification is

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

where  $X_1$  consists of  $k_1$  columns and  $X_2$  consists of  $k_2$  columns. Regress just  $X_1$  on  $y$  without including  $X_2$ , we can characterize  $b_1$ ,

$$b_1 = (X_1'X_1)^{-1}X_1'y \quad (23)$$

$$= (X_1'X_1)^{-1}X_1'[X_1\beta + X_2\beta_2 + \epsilon] \quad (24)$$

$$= (X_1'X_1)^{-1}X_1'X_1\beta + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\epsilon \quad (25)$$

$$= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\epsilon \quad (26)$$

$$(27)$$

Two elements worthy of consideration.

1. If  $\beta_2 = 0$  everything is fine, assuming the standard assumptions hold. The reason: we have not really misspecified the model.
2. Also, if the standard assumptions hold and  $X_1'X_2 = 0$  then the second term also vanishes (even though  $\beta_2 \neq 0$ ). If, on the other hand, neither of these conditions hold, but we estimate the regression in any case, the estimate of  $b_1$  will be biased by a factor of (defining  $P = (X_1'X_1)^{-1}X_1'X_2$ )

$$P_{X_1X_2}\beta_2$$

What is  $P_{X_1X_2}$ ?



## Structural Consistency

In specifying a regression model, we assume that its assumptions apply equally well to all the observations in our sample. They may not. Fortunately, we can test claims of “structural stability” using techniques that we already have encountered.

$H_0$  : Structural stability.

1. Estimate a linear regression assuming away the structural instability. Save the Residual Sum of Squares, call it  $S_1$ .
2. Estimate whatever regressions you believe to be implied by the hypothesis of structural instability and obtain their combined Residual Sum of Squares. Call it  $S_4$ .
3. Subtract the RSS obtained from step 2 ( $S_4$ ) from the RSS obtained in step 1 ( $S_1$ ). Call it  $S_5$ .

4.

$$F_{(k, n_1 + n_2 - 2k)} = \frac{S_5/k}{S_4/(n_1 + n_2 - 2k)}$$

## Revisiting $\sigma^2 I$

Now we can move on to considering the properties of the residuals and their conformity with assumptions we have made about their properties.

- Homoscedasticity
- Normality
  - ★ Jarque-Bera test [for regression, should be  $n = N - k$ ] (sum , detail)

$$JB = \frac{n}{6} \left( S^2 + \frac{(K - 3)^2}{4} \right)$$

★

# Comparing Regression Models

Two types of models, in general

- Nested models
- Nonnested models

In layman's terms, nested models arise when one model is a special case of the other. For example,  $y = \beta_0 + X_1\beta_1 + \epsilon$  is nested in  $y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon$  using the restriction that  $\beta_2 = 0$ . If models are nested, usual techniques can be used. If not, we must turn to alternative tools. Technically, there is probably an intermediate class that would be appropriately named "overlapping". Practically, "overlapping" have some nested elements and some nonnested elements. Almost always, we will need the nonnested tools for these.

# Influence Diagnostics

- dfbeta
- Cook's Distance
- Added-variable plots
- RESET

## A Natural Conception of Panel Data as an Array

Consider  $i \in N$  units observed at  $t \in T$  points in time. The normal cross-sectional data structure will use variables as columns and  $i$  as rows. Panel data then adds the complication of a third dimension,  $t$ . If we were to take this third dimension and transform the array to two dimensions, we would end up with an  $(N \times T)$  by  $K$  matrix of covariates and (for a single outcome) an  $NT$  vector.

# To Pool or Not to Pool?

- Virtues of Panel Data
  - ★ More accurate inference.  
Variety in asymptotics.
  - ★ Control over complexity (Regressors and parameters).
  - ★ Required to isolate short-run and long-run effects simultaneously. Policy and reaction functions?
  - ★ The number of observations grows.  
More data can't really provide less information, with all the Berk/Freedman caveats. Provable in straightforward fashion with important implications.
  - ★ Explicit characterizations of within- and between- variation.
  - ★ Simplification of computation for complex problems
    - Non-stationary time series
    - Measurement error.
    - Computational tricks (dynamic tobit)

- Examining Pooling Assumptions

- ★ Data (What is an outlier in this setting?)

- \*  $D_M(x) = \sqrt{(x - \mu)' \Sigma^{-1} (x - \mu)}$ :

- Mahalanobis distance is a generalization of Euclidean distance ( $\Sigma^{-1} = \Sigma = I$ ) with an explicit covariance matrix. Ali Hadi's work on multivariate outliers uses something similar with reordering to maximize a two subset Mahalanobis distance. aside...

- \* Jackknife summary statistics on one- or two-dimensions.

- \* The real worry seems to be classes/clusters/groups that are different/distinct.

- ★ Models (Stability of a model? and Influence)

- \* Chow test: F-test on pooled against split sample regression. Perhaps iterated Chow tests using combinatoric algorithms over sizes.

- \* Changepoint models, Regimes and Regime Switching and Mixtures



## Convenience Samples and the Like

- Hsiao isolates many of the central issues in panel data from the view of an econometrician. The argument is a bit broader in the sense of repetition.
- Berk and Freedman isolate important issues of particular relevance to the types of structures we will look at.
  - ★ Convenience samples are a fact of social scientific life.
  - ★ Treating the data as a population obviates inference.
  - ★ As-if and imaginary sampling mechanisms. Uncertainty gets really hard.
  - ★ Imaginary populations and imaginary sampling designs: a fiction?
  - ★ What we do with them is our responsibility, but we should be fair.
- Getting more data gives us the ability, but also the need, to do much more.

# The Dimensions of TSCS/CSTS and Summary

- Presence of a time dimensions gives us a natural ordering.
- Space is not irrelevant under the same circumstances as time – nominal indices are irrelevant on some level. Defining space is hard. Ex. targeting of Foreign Direct Investment and defining proximity.
- ANOVA is informative in this two-dimensional setting.
- A part of any good data analysis is summary and characterization. The same is true here; let's look at some examples of summary in panel data settings.

## Basic xt commands

In Stata's language, xt is the way that one naturally refers to CSTS/TSCS data. Consider  $NT$  observations on some random variable  $y_{it}$  where  $i \in N$  and  $t \in T$ . The TSCS/CSTS commands almost always have this prefix.

- xtset: Declaring xt data
- xtodes: Describing xt data structure
- xtsum: Summarizing xt data
- xttab: Summarizing categorical xt data.
- xttrans: Transition matrix for xt data.

## A Primitive Question

Given two-dimensional data, how should we break it down? The most common method is unit-averages; we break each unit's time series on each element into deviations from their own mean. This is called the within transform. The between portion represents deviations between the unit's mean and the overall mean. Stationarity considerations are generically implicit. We will break this up later.

## Some Useful Variances and Notation

- W(ithin) for unit  $i^1$ :

$$W_i = \sum_{t=1}^T (x_{it} - \bar{x}_i)^2$$

- B(etween):

$$B_T = \sum_{i=1}^N (\bar{x}_i - \bar{x})^2$$

- T(otal):

$$T = \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x})^2$$

---

<sup>1</sup>Thus the total within variance would be a summary over all  $i \in N$

## Day 2: Stationarity and ARIMA

## Three Standard Time-Serial Structures [ARIMA]

AutoRegressive Integrated Moving Average (ARIMA) structures characterize most time series of interest (virtually all with the integration of their seasonal counterparts). In general, we write

- Autoregression [AR(p)]:

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \cdots + \rho_p e_{t-p} + v_t$$

- Moving Average [MA(q)]:

$$e_t = v_t + \theta_1 v_{t-1} + \theta_2 v_{t-2} + \cdots + \theta_q v_{t-q}$$

- Autoregression and Moving Average [ARMA(p, q)]:

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \cdots + \rho_p e_{t-p} + v_t + \theta_1 v_{t-1} + \theta_2 v_{t-2} + \cdots + \theta_q v_{t-q}$$

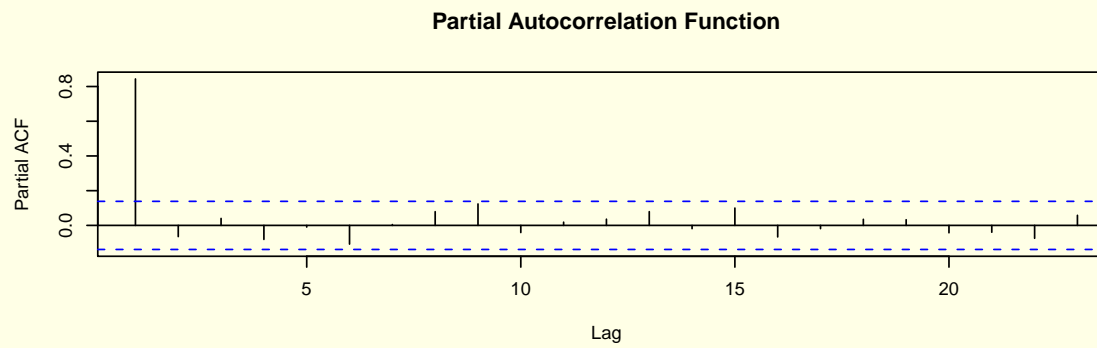
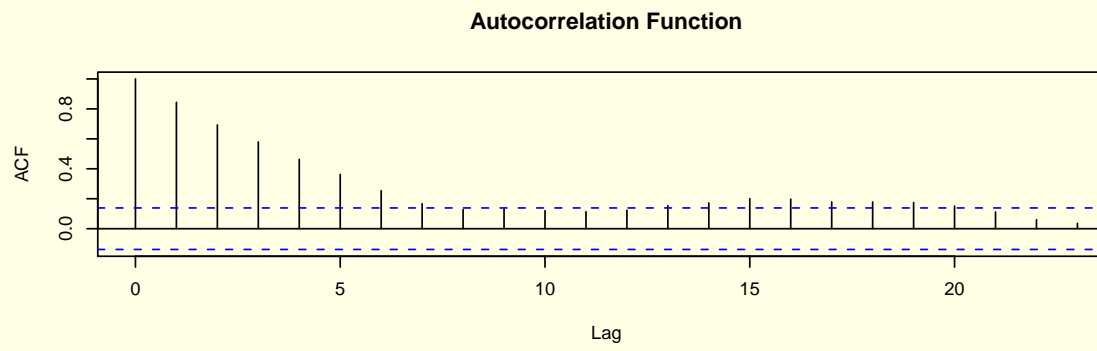
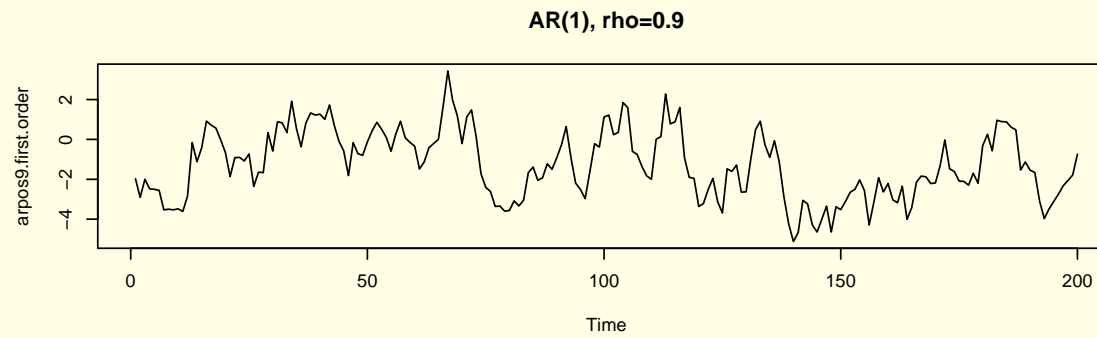
Before, I mentioned two relevant autocorrelations and discussed them briefly:

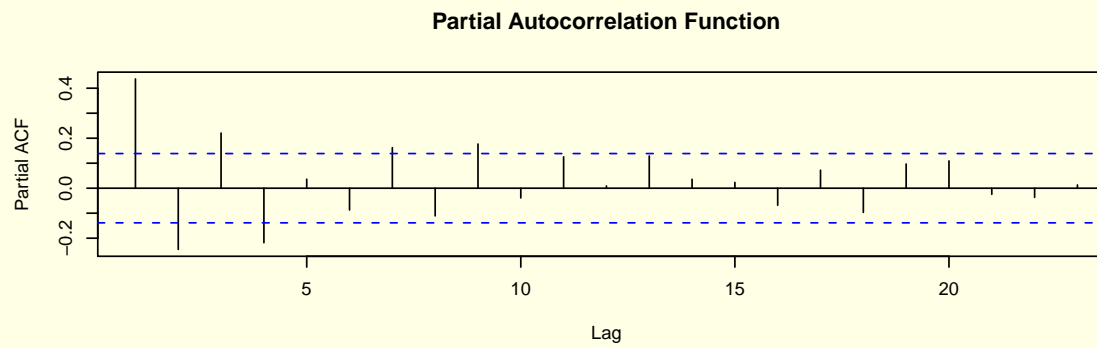
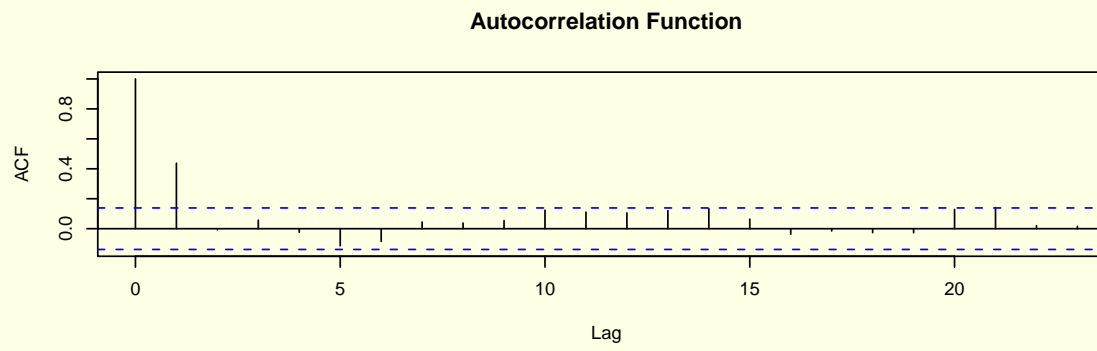
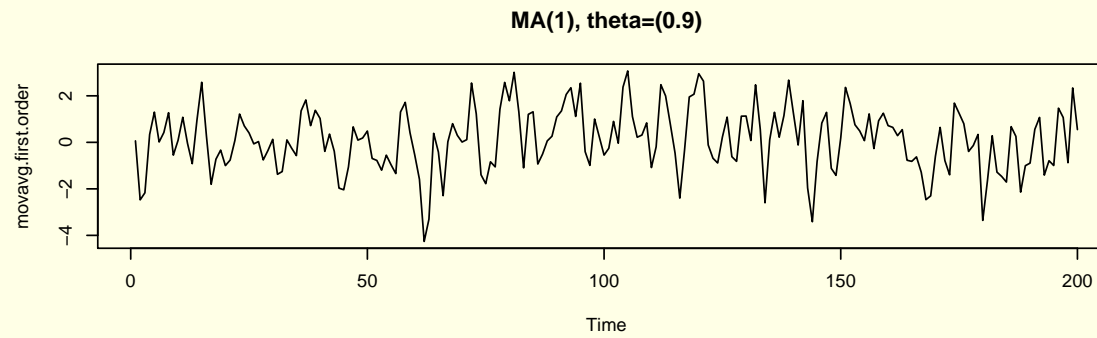
1. Autocorrelation:  $\rho_s = \frac{\sum_{t=s+1}^T (y_t - \bar{y})(y_{t-s} - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2}$

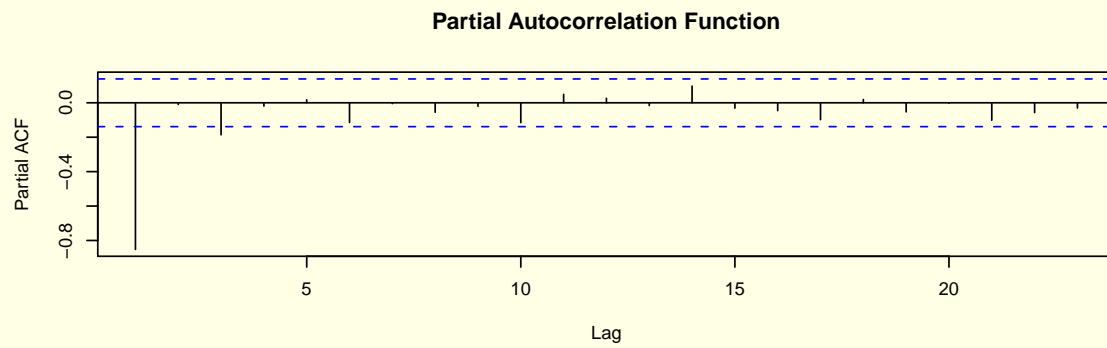
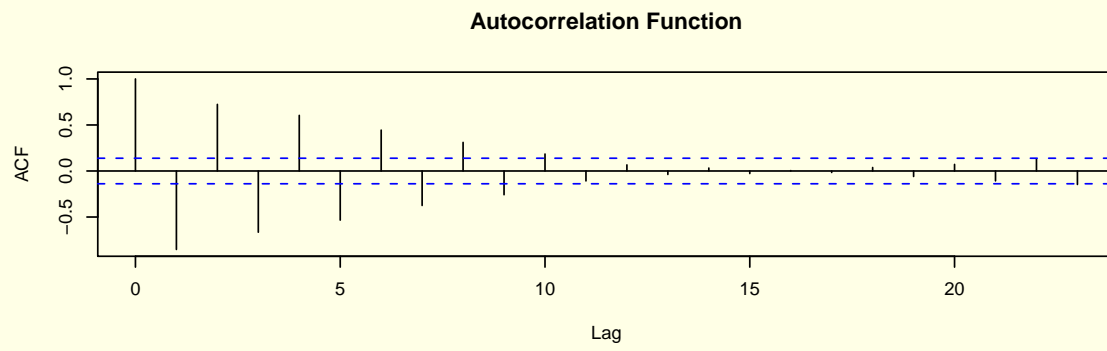
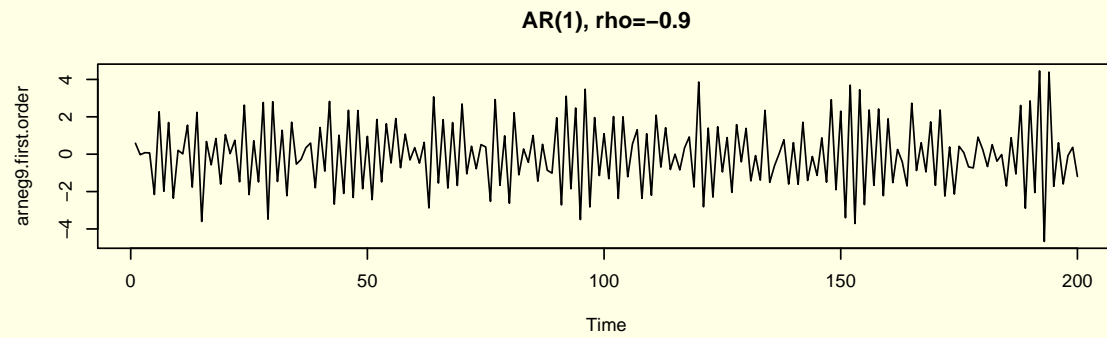
2. Partial Autocorrelation  $\phi_s = \frac{\rho_s - \sum_{j=1}^{s-1} \phi_{s-1,j} \rho_{s-j}}{1 - \sum_{j=1}^{s-1} \phi_{s-1,j} \rho_j}$

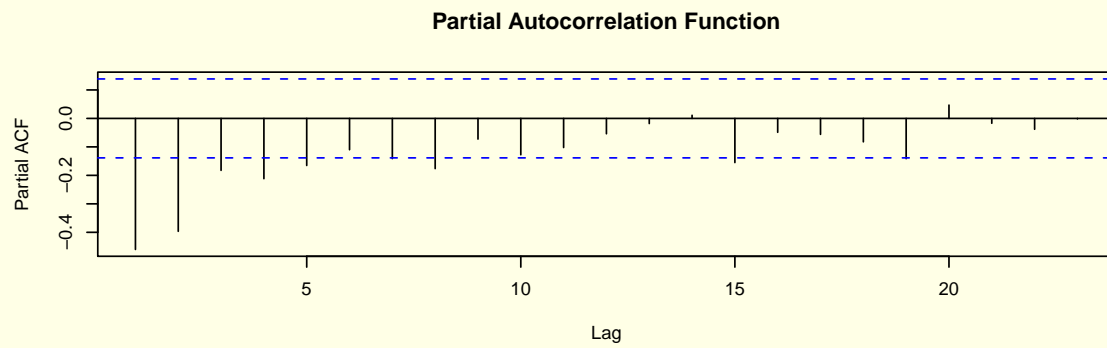
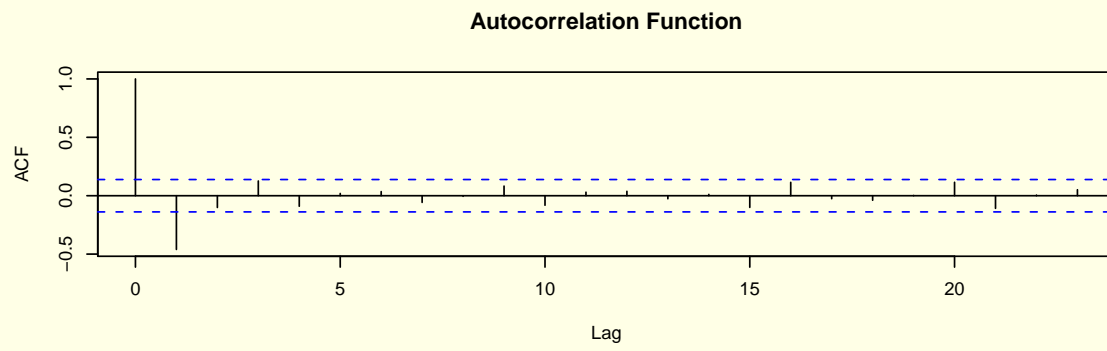
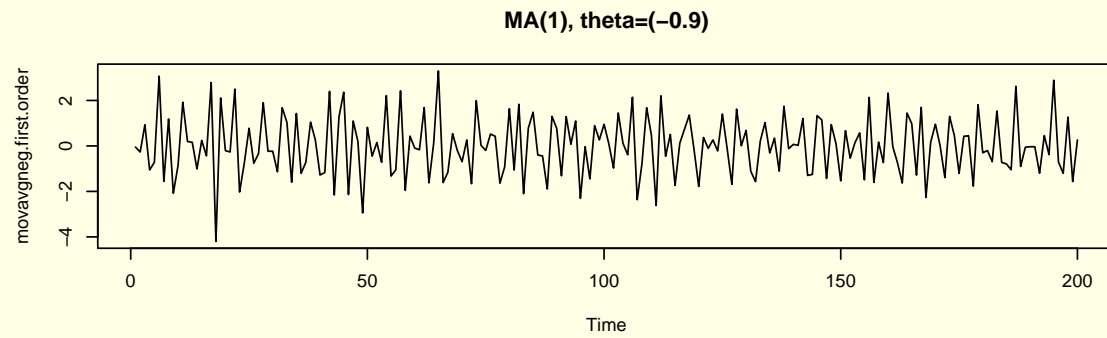
In ARIMA modeling, these are two critical components as each process has a characteristic signature. An autoregressive process typically exhibits geometric decay in the autocorrelation function and spikes in the partial; moving average processes exhibit the reverse. Nonstationary series decay very slowly (the  $I$  in ARIMA).

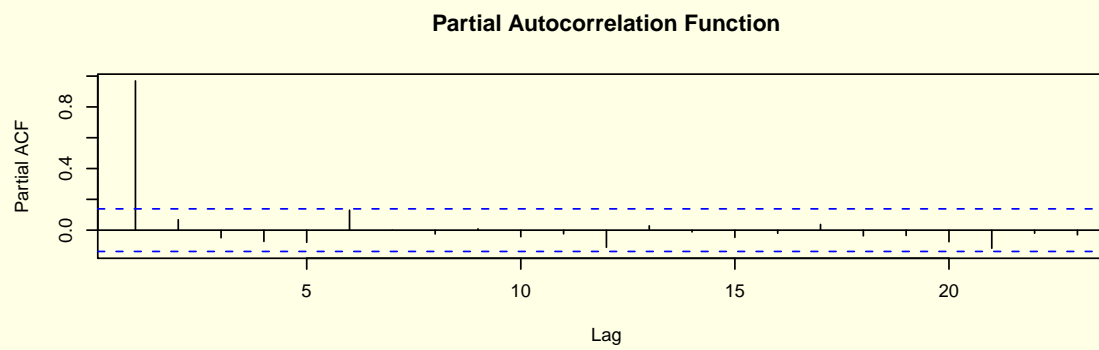
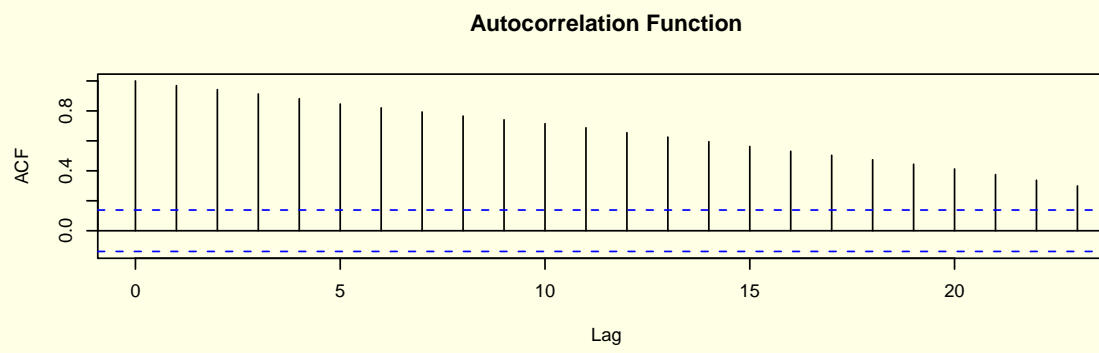
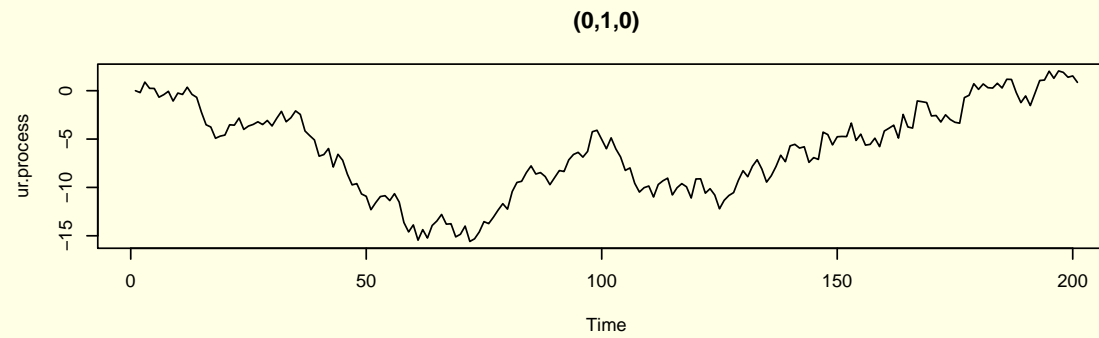












## Stata

Though these plots were generated in *R*, we could do the same thing in Stata. For a quick summary with a little graphic, have a look at the `corrgram`. For (pretty) plots, Stata has two commands to recreate this, `ac` and `pac`. The former generates the autocorrelations while the latter creates the partial autocorrelations. We will have a go at this in the lab.

## TSCS and Time Series

- Common structure restrictions may be difficult to deal with and limit our ability to gain much from combining individual time series.
- Most will be pretty simple structures.
- Mixed orders of integration present special problems.

## Diagnosing Serial Correlation Individually

If we can reject a range of pathologies, we can justify inference rationally?

- First question is the integrity of the estimand; does the conditional mean make sense?
- Unit root tests come in a host of forms with nulls of a unit root and nulls of stationarity. The processes have different implications. Unfortunately, in TSCS/CSTS settings, tests are pretty unreliable. That said,
  - ★ Levin and Lin: `levinlin` with  $H_0 : I(1)$ .
  - ★ Im, Pesaran, and Shin: `ipshin` with  $H_0 : I(1)$ .
  - ★ KPSS: `kpss` with  $H_0 : I(0)$ .
  - ★ Fisher: `xtfisher` works with unbalanced panels
  - ★ Simple `xtreg` with lagged  $y$ , if  $\beta_{y_{t-1}} \approx 1$  then there is a worry.



- Given this:
  - ★ Plots (Every structure has different theoretical ACF/PACF)
  - ★ Durbin-Watson  $d$  and Durbin's  $h$  with endogenous variables
  - ★ Dickey-Fuller tests and many others.  $\Delta y_t = \rho y_{t-1} + \theta_L \Delta y_{t-L} + \lambda_t + u_t$
  - ★ Breusch-Godfrey test and the like (Fit regression, isolate residuals, regress residual on  $X$  and lags of residual,  $nR^2 \sim \chi_p^2$ ).
  
- The above alongside:
  - (1) is the temporal process common or distinct? and
  - (2) if distinct, how and why?

```
* These code are designed to accomplish the following three objectives:
* 1) An introduction to simulating autoregressive processes in STATA
* 2) An introduction to simulating moving averages in STATA
* 3) An introduction to STATA's ARIMA diagnostic procedures
* Code comes from Guy Whitten, Texas A&M University
* Relating this to course slides, rhos are his phis, theta is still theta

* We'll start with some housekeeping and the creation of a time variable:
```

```

clear
set seed 12344
set obs 100
egen time=fill(0 1 2)
tsset time

*****
* Step 1) An introduction to simulating autoregressive processes in STATA*
*****

* First we need to create a set of well-behaved disturbance terms:
drawnorm r1
drawnorm r2
drawnorm r3
drawnorm r4
drawnorm r5
drawnorm r6

* Here is a unit root process.   $y_{\{t\}} = y_{\{t-1\}} + e_{\{t\}}$ 
* y has perfect memory.
gen y1ur=.
replace y1ur=r1 if time==0
replace y1ur=y1ur[_n-1]+r1 if time>0

* Now let's create an AR1 process with phi equal to .7
gen y1=.

```

```

replace y1=r1 if time==0
replace y1=.7*L.y1+r1 if time>0

twoway line y1 time, yline(0)

* We can create an AR2 or higher-order AR processes by adding additional terms.
* Here is an AR2 process with phi1=.4 and phi2=.3
gen y2=.
replace y2=r2 if time==0
replace y2=.4*L.y2+r2 if time==1
replace y2=.4*L.y2+.3*L2.y2+r2 if time>1

* Let's take a look at this:
twoway line y2 time, yline(0)

*****
* Step 2) An introduction to simulating moving averages in STATA*
*****

* Moving averages are created by theta coefficients on the lagged error process.
* So an MA1 process with theta=.4 would be created by:

gen y3=.7*L.r3+r3

twoway line y3 time, yline(0)

```

\* An MA2 processes with theta1=.3 and theta2=.1 would be created by:

```
gen y4=.3*L.r4+.1*L2.r4+r4
```

```
twoway line y4 time, yline(0)
```

\*\*\*\*\*

\* Step 3) An introduction to STATA's ARIMA diagnostic procedures\*

\*\*\*\*\*

\* The basic diagnostic suite for autocorrelations in STATA is contained in the

\* "corrgram" command. For instance, if we use this command with y1 (an AR1

\* process with phi=7):

```
corrgram y1
```

\* The "ac" command (help and options for this command can be found under "corrgram")

\* plots a graph of autocorrelations with confidence intervals:

```
ac y1
```

\* The "pac" command (help and options for this command can also be found under

\* "corrgram") plots partial autocorrelations with confidence intervals:

```
pac y1
```

```
ac y2
```

```
pac y2
```

```
ac y3  
pac y3
```

```
* Now let's run an ARIMA model of y1 using our pdq=(1,0,0)  
arima y1, arima(1,0,0)
```

```
* saving the residuals:  
predict y1_res, resid  
corrgram y1_res
```

```
arima y2, arima(2,0,0)
```

```
* Now it's your turn:
```

- \* 1) Try to create some different types of AR processes
- \* 2) Try to create some different types of MA processes
- \* 3) Run some autocorrelation diagnostics on the series that you create

## Panel Unit Root Testing in Stata

As of Stata 11, a battery of panel unit-root tests have emerged. There are many and they operate under differing sets of assumptions.

- Levin-Lin-Chu (`xtunitroot llc`): trend nocons (unit specific) demean (within transform) lags. Under (crucial) cross-sectional independence, the test is an advancement on the generic Dickey-Fuller theory that allows the lag lengths to vary by cross-sections. The test relies on specifying a kernel (beyond our purposes) and a lag length (upper bound). The test statistic has a standard normal basis with asymptotics in  $\frac{\sqrt{N_T}}{T}$  ( $T$  grows faster than  $N$ ). The test is of either all series containing unit roots ( $H_0$ ) or all stationary; this is a limitation. It is recommended for moderate to large  $T$  and  $N$ .

1. Perform separate ADF regressions:

$$\Delta y_{it} = \rho_i \Delta y_{i,t-1} + \sum_{L=1}^{p_i} \theta_{iL} \Delta y_{i,t=L} + \alpha_{mi} d_{mt} + \epsilon_{it}$$

with  $d_{mt}$  as the vector of deterministic variables (none, drift, drift and trend). Select a max  $L$  and use  $t$  on  $\hat{\theta}_{iL}$  to attempt to simplify. Then use  $\Delta y_{it} = \Delta y_{i,t-L}$  and  $d_{mt}$  for residuals

- Harris-Tzavalis (xtunitroot ht): trend nocons (unit specific) demean (within transform) altt (small sample adjust) Similar to the previous, they show that  $T \rightarrow \infty$  faster than  $N$  (rather than  $T$  fixed) leads to size distortions.
- Breitung (xtunitroot breitung): trend nocons (unit specific) demean (within transform) robust (CSD) lags.  
Similar to LLC with a common statistic across all  $i$ .

- Im, Pesaran, Shin (xtunitroot ips): trend demean (within transform) lags. They free  $\rho$  to be  $\rho_i$  and average individual unit root statistics. The null is that all contain unit roots while the alternative specifies at least some to be stationary. The test relies on sequential asymptotics (first T, then N). Better in small samples than LLC, but note the differences in the alternatives.
- Fisher type tests (xtunitroot fisher): dfuller pperron demean lags.
- Hadri (LM) (xtunitroot hadri): trend demean robust

All but the last are null hypothesis unit-root tests. Most assume balance but the fisher and IPS versions can work for unbalanced panels.



## Stationarity Issues

- Essence of stationarity is threefold: means, variances, and crosses are not time-dependent.
- There is a quite famous spurious regressions result in econometrics that owes to the statistician Yule in 1926.
- Basically, the regression of  $I(1)$  series on one another has non- $\alpha$  rejection rates.
- Applied to panels, a mix of orders of integration will give  $t$  statistics non- $t$  properties.
- In the end, I suspect the best advice is to partition data on the basis of likely orders of integration and proceed from there.

## The Managable Autocorrelation Structure

$$\Phi = \sigma^2 \Psi = \sigma_e^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{T-2} \\ \rho_2 & \rho_1 & 1 & \dots & \rho_{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{T-1} & \rho_{T-2} & \rho_{T-3} & \dots & 1 \end{pmatrix}$$

given that  $e_t = \rho e_{t-1} + v_t$ . A Toeplitz form....

This allows us to calculate the variance of  $e$  using results from basic statistics, i.e.  $Var(e_t) = \rho^2 Var(e_{t-1}) + Var(v)$ . If the variance is stationary, we can rewrite,

$$\sigma_e^2 = \frac{\sigma_v^2}{1 - \rho^2}$$

## Three Standard Time-Serial Structures [ARIMA]

AutoRegressive Integrated Moving Average (ARIMA) structures characterize most time series of interest (virtually all with the integration of their seasonal counterparts). In general, we write

- Autoregression [AR(q)]:

$$e_t = \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + v_t$$

- Moving Average [MA(p)]:

$$e_t = v_t + \alpha_1 v_{t-1} + \alpha_2 v_{t-2} + \cdots + \alpha_p v_{t-p}$$

- Autoregression and Moving Average [ARMA(q, p)]:

$$e_t = \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + v_t + \alpha_1 v_{t-1} + \alpha_2 v_{t-2} + \cdots + \alpha_p v_{t-p}$$

# Autocorrelation

When discussing heteroscedasticity, we notice that the off-diagonal elements are all zeroes. This is the assumption of no correlation among [somehow] adjacent elements. The somehow takes two forms: (1) spatial and (2) temporal. Just as before where time-induced heteroscedasticity simply involved interchanging  $N$  and  $T$  and  $i$  and  $t$ ; the same idea prevails here.

Two relevant autocorrelations:

1. Autocorrelation: 
$$\rho_s = \frac{\sum_{t=s+1}^T (y_t - \bar{y})(y_{t-s} - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2}$$

2. Partial Autocorrelation 
$$\phi_s = \frac{\rho_s - \sum_{j=1}^{s-1} \phi_{s-1,j} \rho_{s-j}}{1 - \sum_{j=1}^{s-1} \phi_{s-1,j} \rho_j}$$