2022 Essex Summer School
3K: Dynamics and Heterogeneity

Robert W. Walker, Ph. D.
Associate Professor of Quantitative Methods
Atkinson Graduate School of Management
Willamette University
Salem, Oregon USA
rwalker@willamette.edu

July 30, 2022

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: We have a course textbook and essential articles for the course online. The syllabus outlines issues from textbooks that have more detailed foundations. One can learn a great deal from walking through fine details. Each day, near the end, I will signpost what of the readings to focus for the following day.

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: One then Multiple Time Series

- 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. First t.

Data Examples

- The time series of Oregon's Bond rating. $\rightarrow$ Bond ratings among U. S. states.
- Nine justices vote on a case $\rightarrow$ US Supreme Court justices/votes
- a dyad fights $\rightarrow$ Wars/Military conflicts
- OECD countries/global samples in political economy.
- FDI inflows/outflows, etc.
- Voter turnout across nations or US states/counties/municipalities.

Some Basic Things

- Time is complicated.
- Time-zones
- Daylight savings
- Irregular periodicity

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, χ^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 $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_1, \ldots, v_n)$ 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

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n$$

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

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 $\sum_i \sum_j a_i a_j Cov(\mathbf{X}_i \mathbf{X}_j)$

Some (Loosely Stated) Distributional Results

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

Normal random variables

A random variable X has a normal distribution with mean μ and variance σ^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.

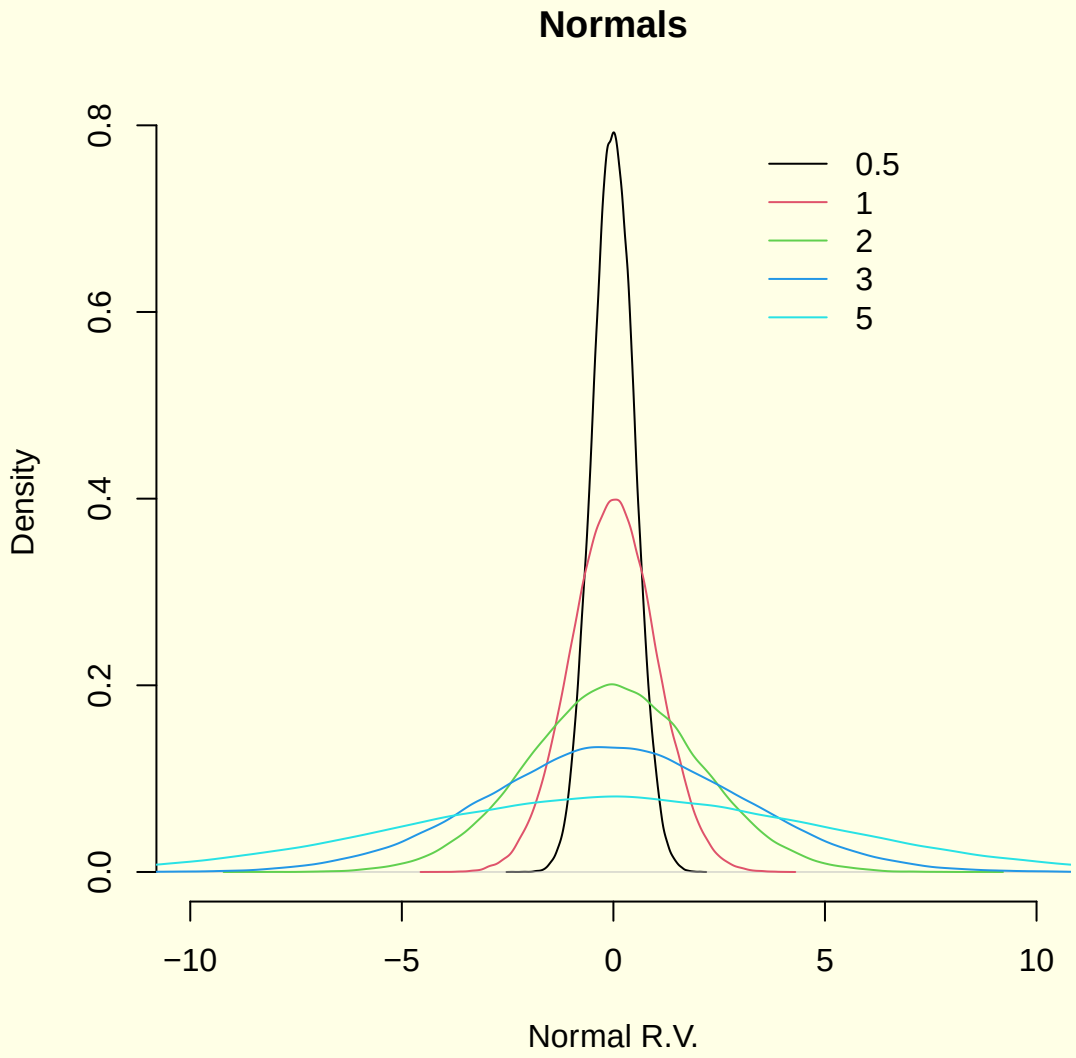


Figure 1: The Normal distribution (with varying σ^2)

Basu's Theorem

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

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

the sample mean is a complete sufficient statistic – it is informationally optimal to estimate μ.

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

the sample variance, is an ancillary statistic – its distribution does not depend on μ. 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.

χ^2 random variables

If a random variable X has a χ^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 $X_1, \ldots, X_k$ are independent and if X_i has a χ^2 distribution with n_i degrees of freedom, then $\sum_{i=1}^{k} X_i$ has a χ^2 distribution with $\sum_{i=1}^{k} n_i$ degrees of freedom.

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

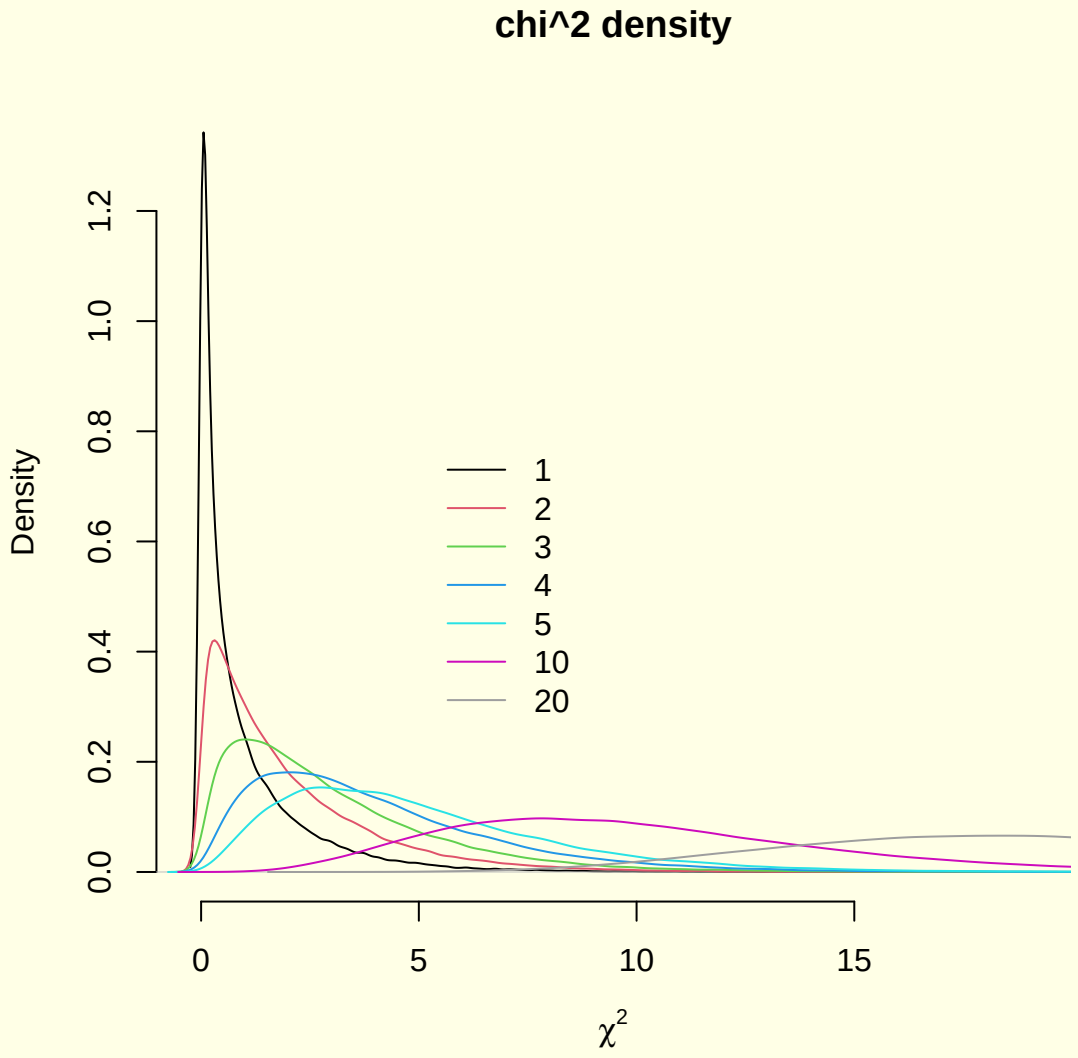


Figure 2: The χ^2 distribution

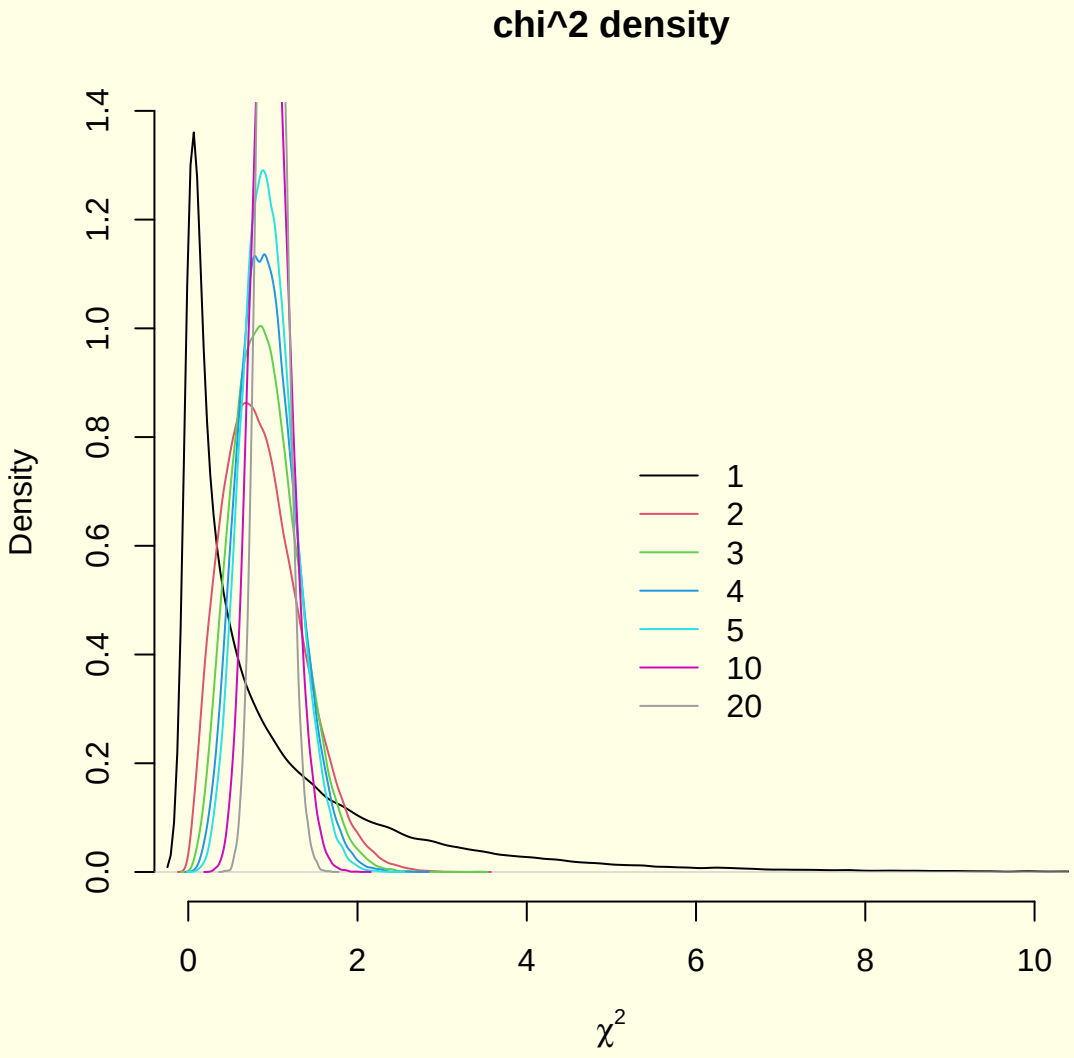


Figure 3: The χ^2 distribution (/n root)

Student's t distributed random variables

Consider two independent random variables Y and Z, such that Y has a χ^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)

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

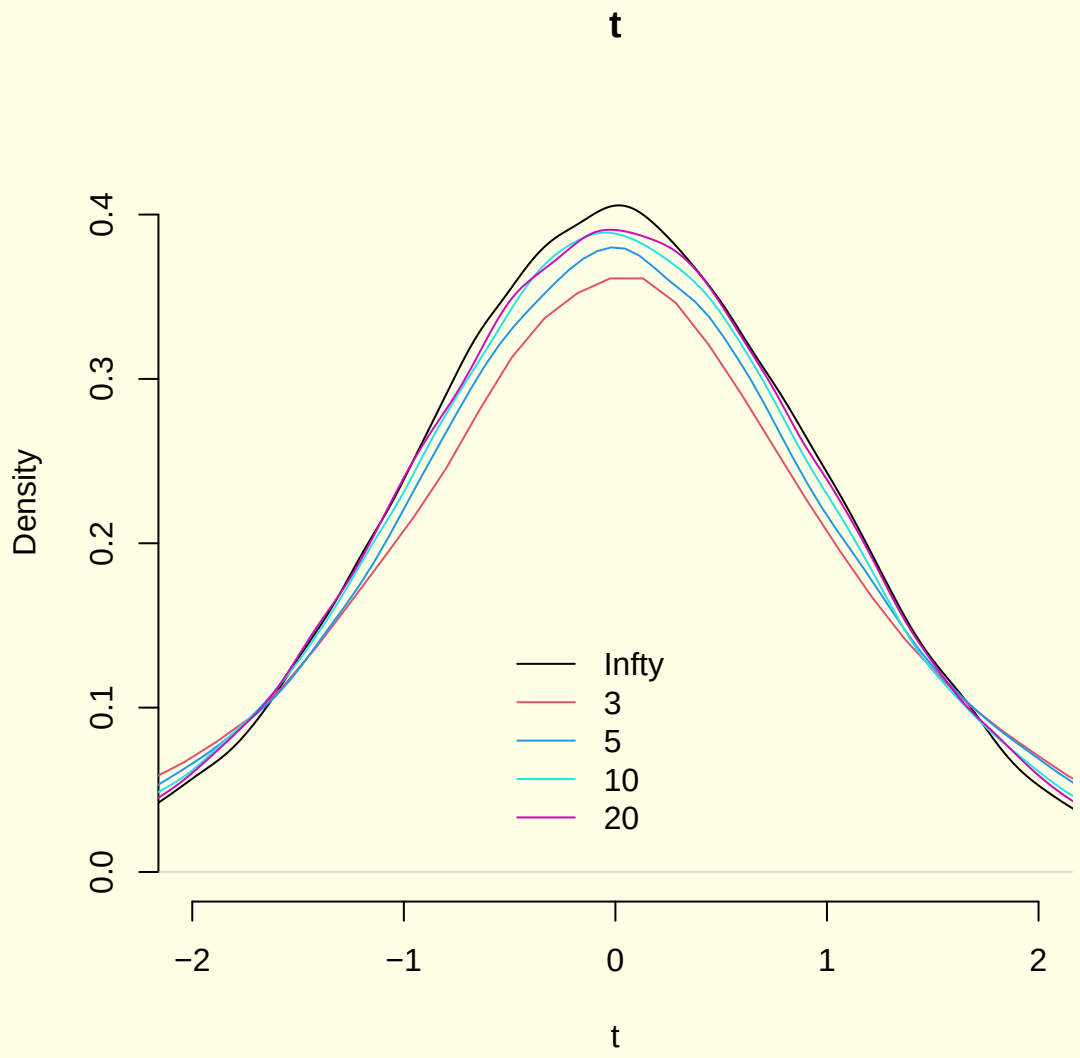


Figure 4: The t distribution

F distributed random variables (Variance-Ratio)

Consider two independent random variables Y and W, such that Y has a χ^2 distribution with m degrees of freedom and W has a χ^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.

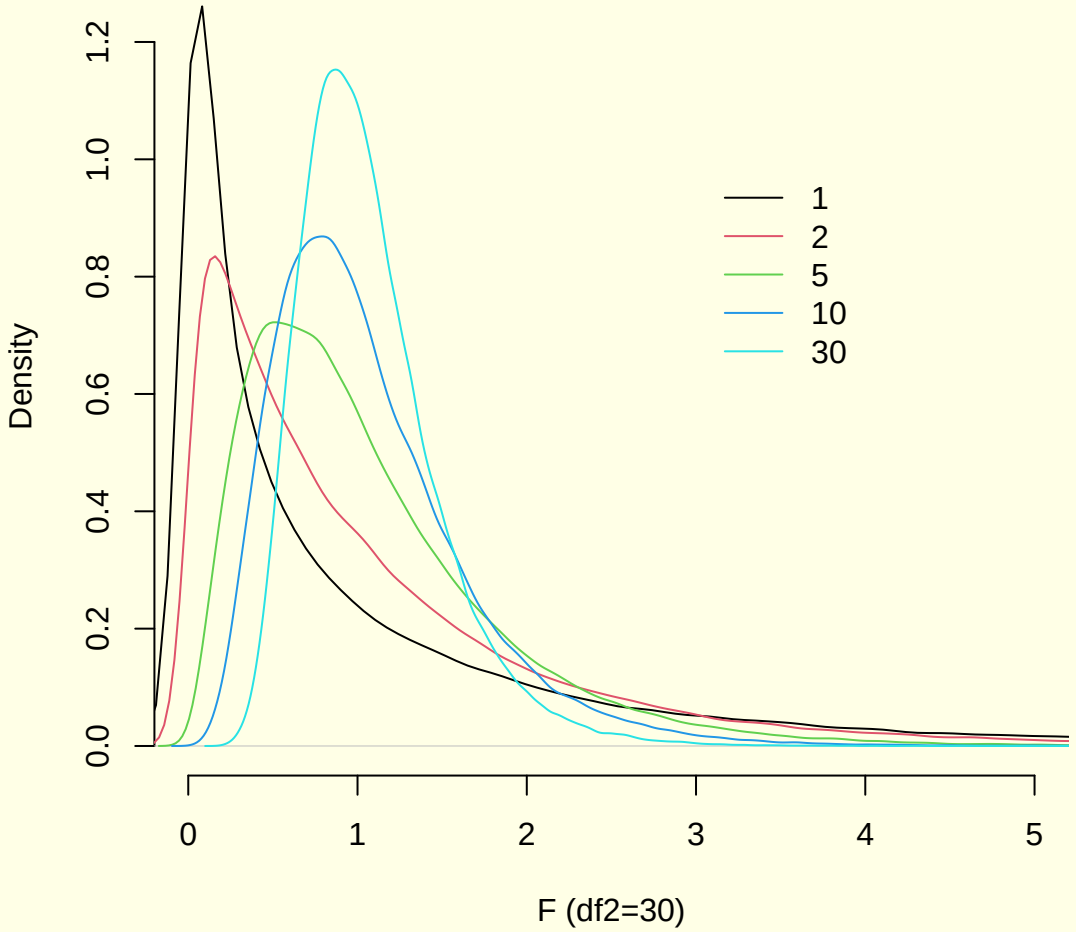


Figure 5: The F distribution: Numerator

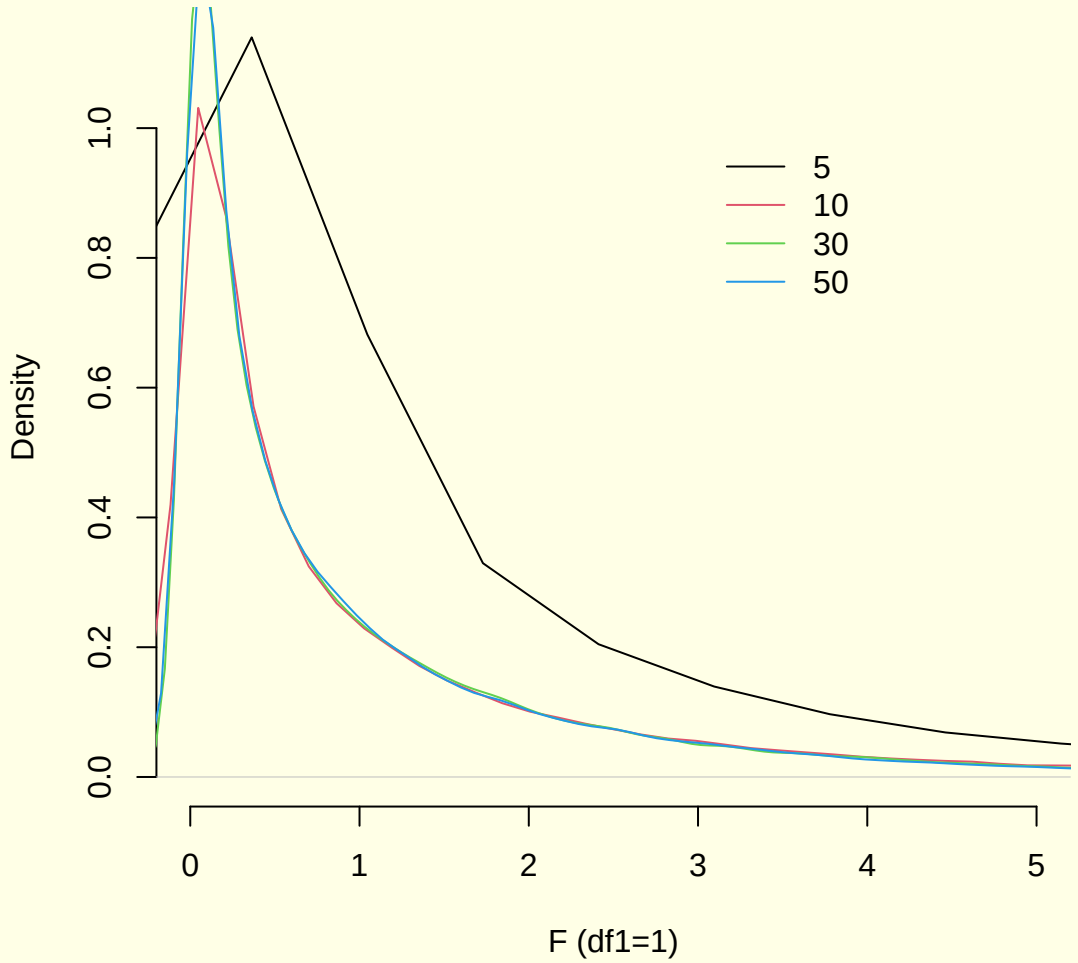


Figure 6: The F distribution: Denominator

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 ther inverses. We have some distributional results that link the normal, χ^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:
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

2. Strict Exogeneity
$$E[\boldsymbol{\epsilon}|\boldsymbol{X}] = 0$$

3. No [perfect] multicollinearity:
Rank of the NxK data matrix X is K with probability 1 ($N > K$).

4. **X** is a nonstochastic matrix.

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

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 $\boldsymbol{e}$.

$$\begin{aligned} \boldsymbol{e}'\boldsymbol{e} &= (\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta) && (1) \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\beta - \beta'\boldsymbol{X}'\boldsymbol{y} + \beta'\boldsymbol{X}'\boldsymbol{X}\beta && (2) \end{aligned}$$

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

$$\begin{aligned} \frac{\partial \boldsymbol{e}'\boldsymbol{e}}{\partial \beta} &= -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\beta \therefore && (3) \\ \boldsymbol{X}'\boldsymbol{y} &= \boldsymbol{X}'\boldsymbol{X}\beta && (4) \\ \hat{\beta}_{OLS} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} && (5) \end{aligned}$$

Properties of OLS Estimators

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

The first two

Need nothing about the distribution other than the two moment defintions. 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 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$, $\mathbb{E}[\hat{\boldsymbol{\beta}} - \beta] = 0$ 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

$$\begin{aligned} \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon) - \beta] &= 0 \quad (6) \\ \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon - \beta] &= 0 \quad (7) \end{aligned}$$

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

$\mathbb{E}[(\hat{\boldsymbol{\beta}} - \beta)(\hat{\boldsymbol{\beta}} - \beta)']$ 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,

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

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 $\mathbf{I}$. 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**: $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$

2. Residual Maker **M**: $\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$

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

On M and P

1. M

$$
\begin{aligned}
M &= \mathbf{I} - X(X'X)^{-1}X' && (12)\\
\mathbf{My} &= (\mathbf{I} - X(X'X)^{-1}X')\mathbf{y} && (13)\\
\mathbf{My} &= \mathbf{Iy} - X\underbrace{(X'X)^{-1}X'\mathbf{y}}_{\hat{\beta}} && (14)\\
\mathbf{My} &= \mathbf{y} - X\hat{\beta} && (15)\\
\mathbf{My} &= \hat{\epsilon} && (16)
\end{aligned}
$$

2. $\mathbf{P}$

$$\mathbf{P} = \mathbf{I} - \mathbf{M} \tag{17}$$

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

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

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

$$\mathbf{P}\mathbf{y} = X\hat{\beta} \tag{21}$$

$$\mathbf{P}\mathbf{y} = \hat{y} \tag{22}$$

General Linear Regression Model: Multiple Regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

1. $\boldsymbol{X}$ is a $n \times k$ matrix of regressors (where the first column is 1)
2. $\boldsymbol{\beta}$ is a $k \times 1$ matrix of unknown (partial) slope coefficients
3. ϵ 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

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

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 $\beta_1 = \beta_2 = \ldots = \beta_k = 0$.

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.

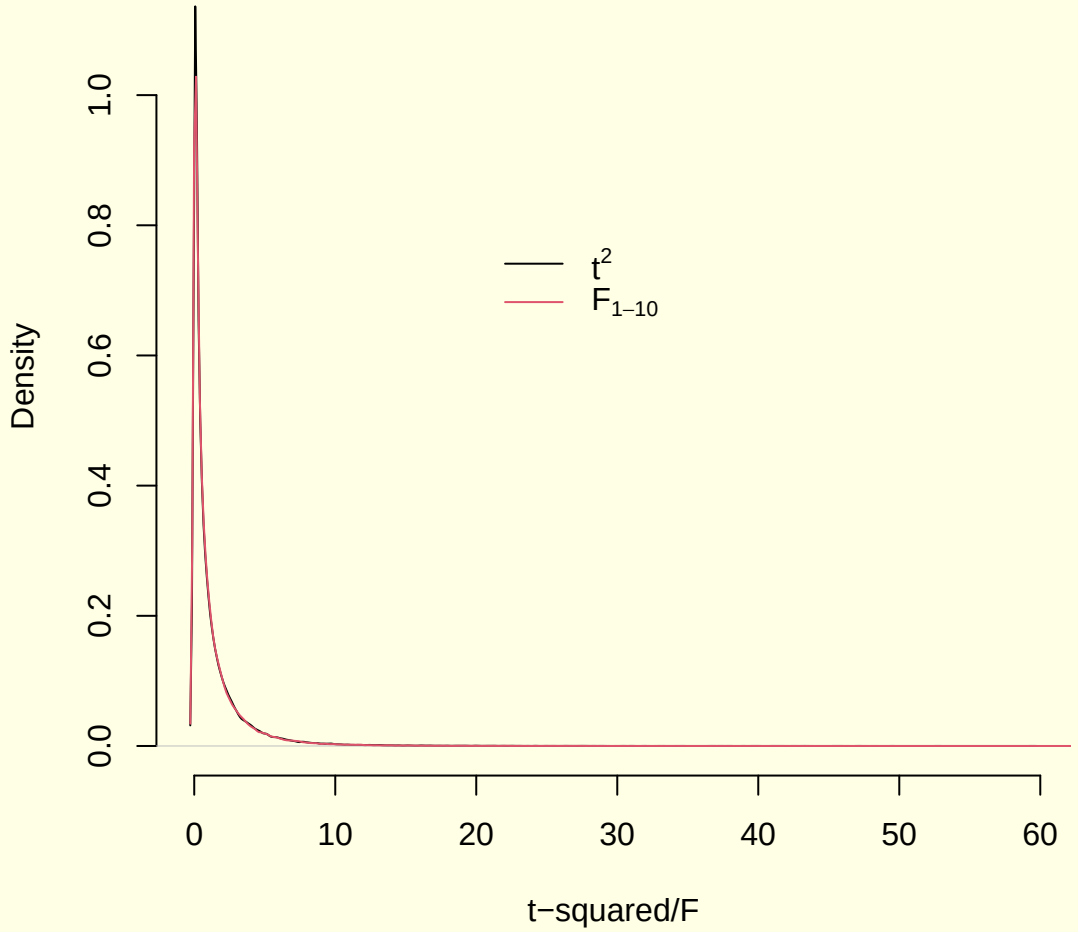


Figure 7: t and F distributions for Simple Hypotheses ($t_{10}, F_{1,10}$)

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:

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

  - ⋆ A mean response:

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

- $\hat{\beta}$ Confidence Intervals
  - ⋆ Individual:

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

    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:

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

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,

$$\begin{aligned} b_1 &= (X_1'X_1)^{-1}X_1y && (23)\\ &= (X_1'X_1)^{-1}X_1'[X_1\beta + X_2\beta_2 + \epsilon] && (24)\\ &= (X_1'X_1)^{-1}X_1'X_1\beta + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\epsilon && (25)\\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\epsilon && (26)\\ & && (27) \end{aligned}$$

Two elements worthy of consideration.

1. If $\boldsymbol{\beta_2} = 0$ 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 $\boldsymbol{X_1' X_2} = 0$ then the second term also vanishes (even though $\boldsymbol{\beta_2} \neq 0$). If, on the other hand, neither of these conditions hold, but we estimate the regression in any case, the estimate of $\boldsymbol{b_1}$ will be biased by a factor of (defining $\boldsymbol{P} = (\boldsymbol{X_1' X_1})^{-1} \boldsymbol{X_1' X_2}$)

$$\boldsymbol{P}_{X_1 X_2} \boldsymbol{\beta_2}$$

What is $\boldsymbol{P}_{X_1 X_2}$?

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 there 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, $\boldsymbol{y = \beta_0 + X_1\beta_1 + \epsilon}$ is nested in $\boldsymbol{y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon}$ using the restriction that $\boldsymbol{\beta_2 = 0}$. 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
- `xtdes`: Describing `xt` data structure
- `xtsum`: Summarizing `xt` data
- `xttab`: Summarizing categorical `xt` data.
- `xttrans`: Transition matrix for `xt` data.

Basic `xt` commands

```
library(foreign)
HR.Data <- read.dta("ISQ99-Essex.dta")
library(skimr)
skim(HR.Data)
library(tidyverse)
library(plm)
source(url("https://raw.githubusercontent.com/robertwwalker/DADMStuf
# Be careful with the ID variable, the safest is to make it factor;
xtsum(IDORIGIN~., data=HR.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]:

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

- B(etween):

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

- T(otal):

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

[1] Thus the total within variance would be a summary over all $i \in N$

Outline for Day 2

Big picture: Models for Single Time Series

1. Stationarity and differencing
2. Spurious regressions: Yule (1926)
3. Autoregressive and moving average terms.
4. Unit-root testing
5. Event Studies
6. The model/substance interaction

Day 2: Univariate Time Series: Stationarity and ARIMA

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-α 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.

Differencing

In the case of integer orders of integration, the widely available and simple solution is to take sufficient differences to render the variable stationary. If the levels are not stationary, try changes; if changes are non-stationary, try the change in change. And so on. But differencing has rather pernicious substantive consequences.

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.

Stationarity and the I: Integration

The basic idea: time-consistency. James D. Hamilton's Time Series Analysis defines two notions of stationarity.

1. Strict stationarity: A process is said to be strictly stationary if, for any values of $j_1, j_2, \ldots, j_n$ the joint distribution of $Y_t, Y_{t+j_1}, Y_{t+j_2}, \ldots, Y_{t+j_n}$ depends only on the intervals separating the dates $(j_1, j_2, \ldots, j_n)$ and not on the date itself t.

2. Weak stationarity: If neither the mean μ, nor the autocovariances γ_j, depend on the date t, then the process for Y_t is said to be covariance-stationary or weakly stationary.

$$\mathbb{E}[(Y_t)] = \mu \; \forall \; t$$
$$\mathbb{E}[(Y_t - \mu)(Y_{t-j} - \mu)] = \gamma_j \; \forall t \; [any] \; j$$

Why It Matters?

Everything is calculated from deviates: $y_i - \overline{y}$

We assume stationarity in so doing. **And we do so at our peril.**

A Simple Example

Let's examine a simple form of first-order dependence. Let's suppose that the current observation depend on the immediately prior observation by some elasticity ρ. Let $\epsilon_t \sim N(0, \sigma_\epsilon)$ This yields:

$$y_t = \alpha + \rho_1 y_{t-1} + \epsilon_t$$

One sufficient condition for stationarity would be that $abs(\rho) < 1$ – we will see this again shortly. Why? Suppose it is one.

$$\begin{aligned} y_t &= \alpha + \rho_1 y_{t-1} + \epsilon_t && (28) \\ y_t - y_{t-1} &= \alpha + \epsilon_t && (29) \\ & && (30) \end{aligned}$$

The over time difference in y is a constant plus a well-behaved error. This is known as a random walk with drift. This problem can be easily estimated with the difference in y on the left hand side. But, keeping in mind that ϵ has zero expectation, this will grow or shrink endlessly by virtue of α. The mean is a function of time.

Yule and Spurious Regressions

Let's have a look at a little simulation on the box.

As you see, this applies to non-stationary series. But not everything with dependency through time is non-stationary. ARMA process provide the middle ground while the I in ARIMA relates to required differences for stationarity.

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

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

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

Three Standard Time-Serial Structures [ARMA]

Auto Regressive Moving Average (ARMA) structures characterize most time series of interest (virtually all with the inclusion of their seasonal counterparts). In general, we write

- Autoregression [AR(p)]:

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

- Moving Average [MA(q)]:

$$e_t = \nu_t + \theta_1 \nu_{t-1} + \theta_2 \nu_{t-2} + \cdots + \theta_p \nu_{t-q}$$

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

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

Illustrations

One way of getting a handle on this is to attempt to measure it. From the inimitable Allison Horst....

An Illustration

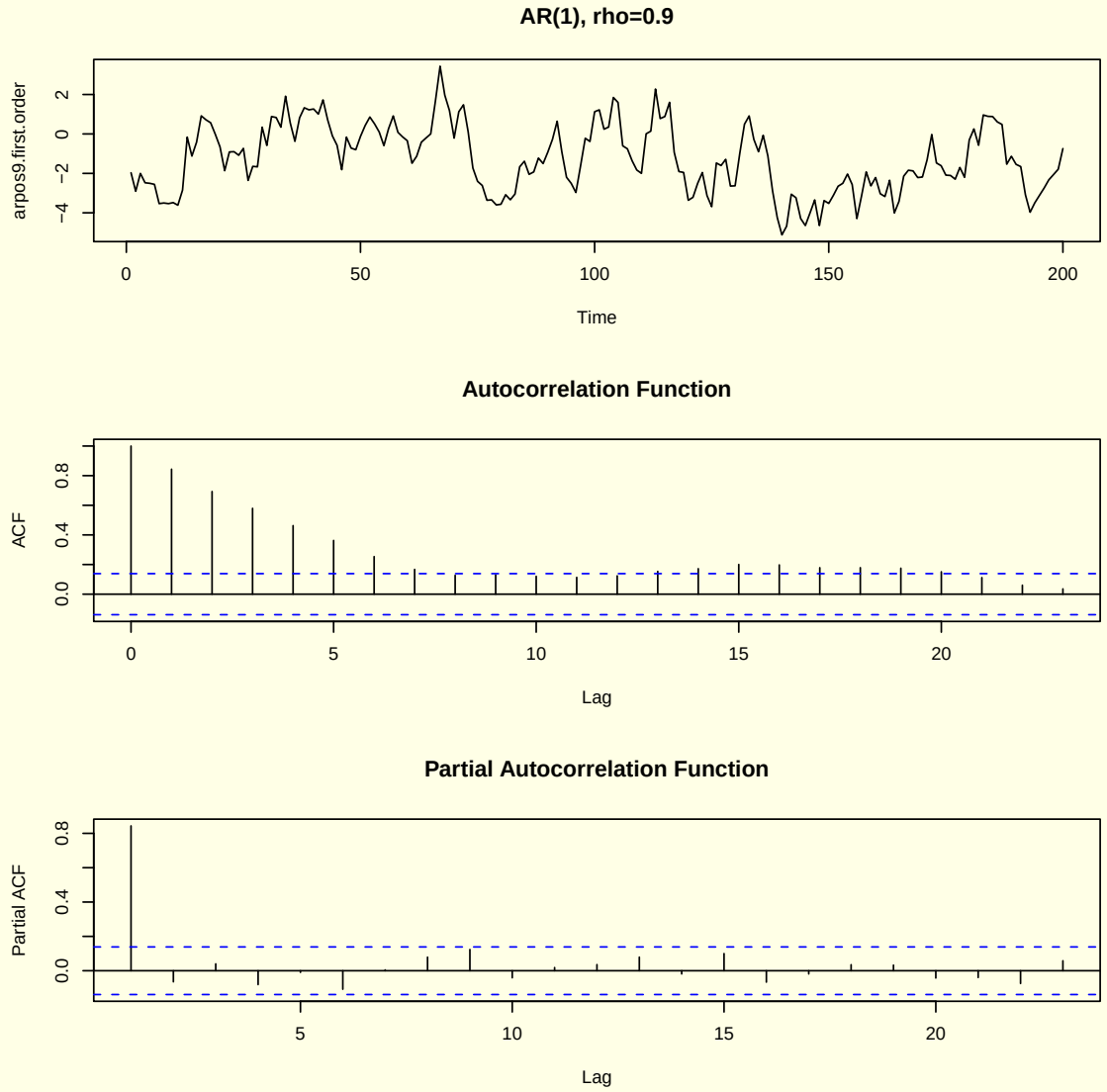
ACF and PACF

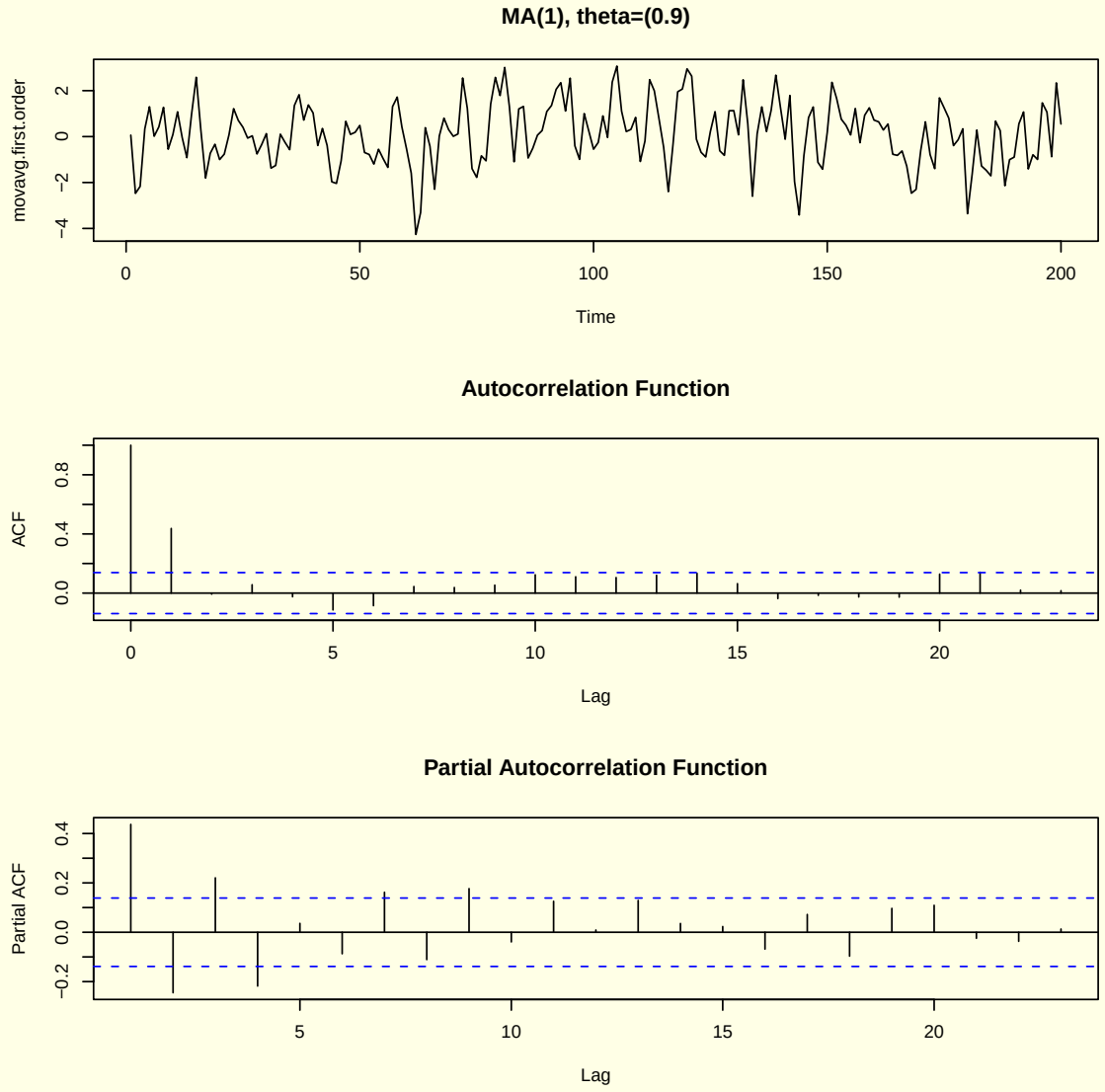
Two relevant autocorrelations:

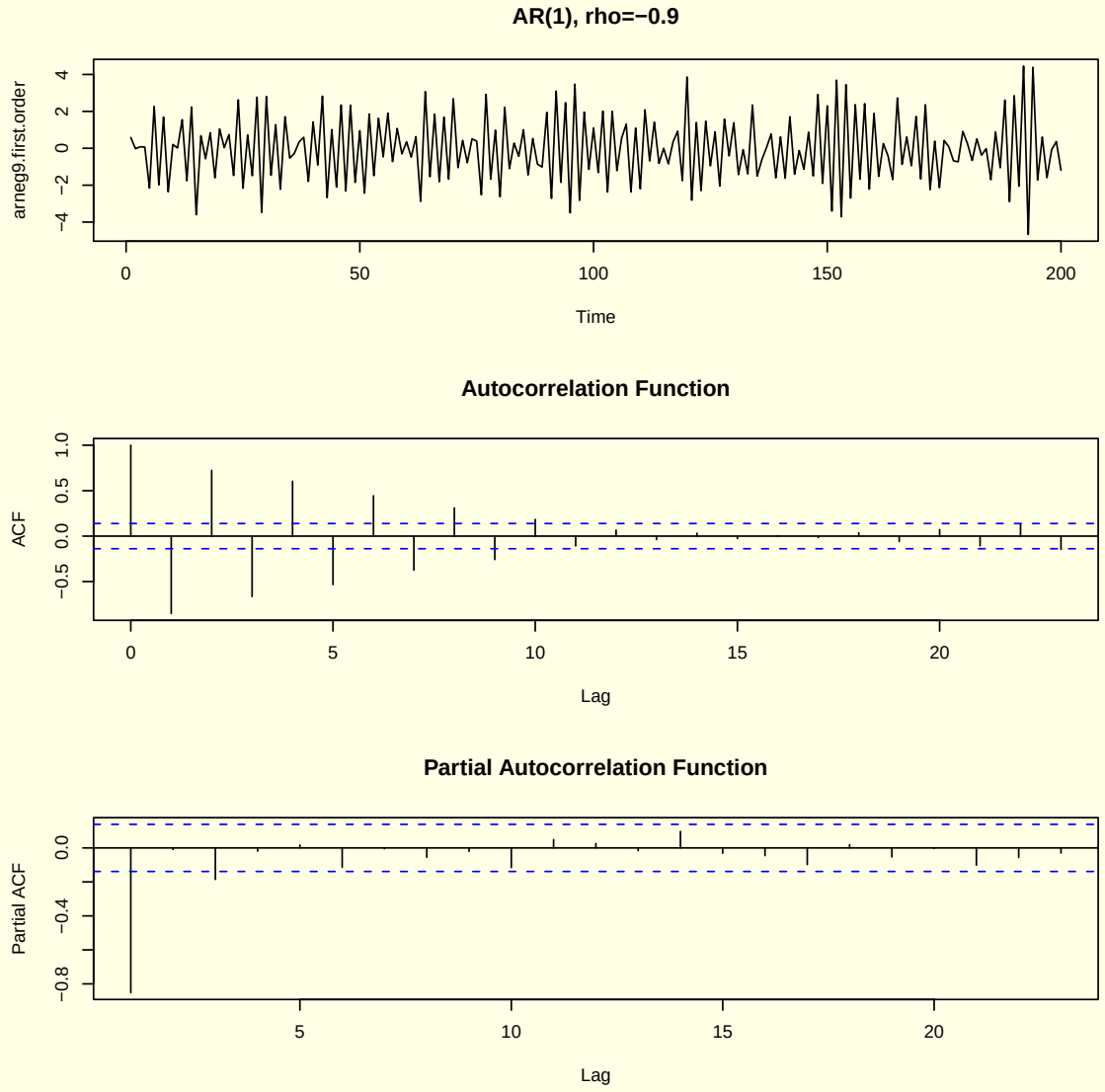
1. Autocorrelation: $\rho_s = \frac{\sum_{t=s+1}^{T}(y_t - \overline{y})(y_{t-s} - \overline{y})}{\sum_{t=1}^{T}(y_t - \overline{y})^2}$

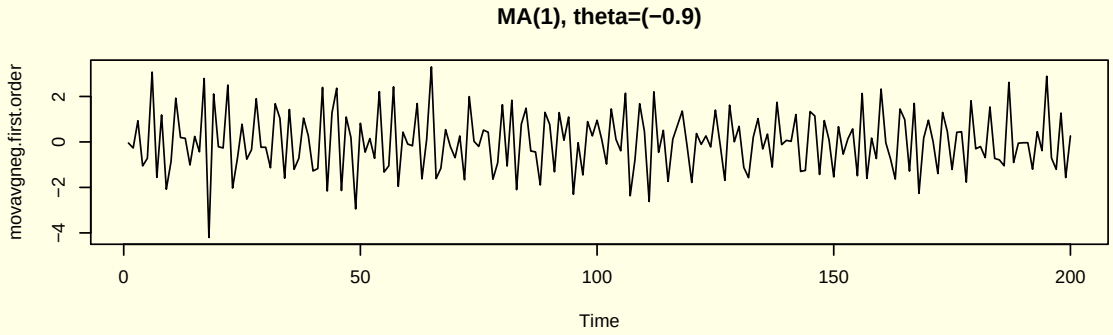
2. Partial Autocorrelation $\phi_s = \frac{\rho_s - \sum_{j=1}^{s-1}\phi_{s-1,j}\rho_{s-j}}{1 - \sum_{j=1}^{s-1}\phi_{s-1,j}\rho_j}$

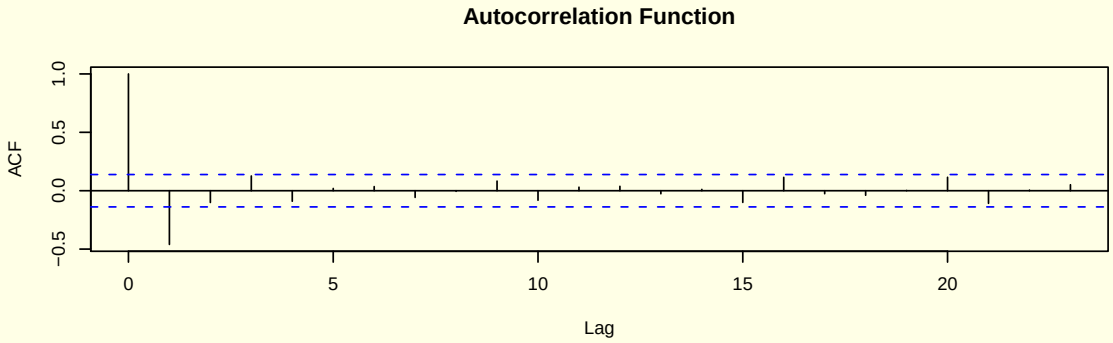
In ARIMA modeling, these are two critical components as each process has a characteristic signature. An autoregressive process typically exhibits geometric decay in the autocorrelation function and spikes in the partial; moving average processes exhibit the reverse. Nonstationary series decay very slowly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

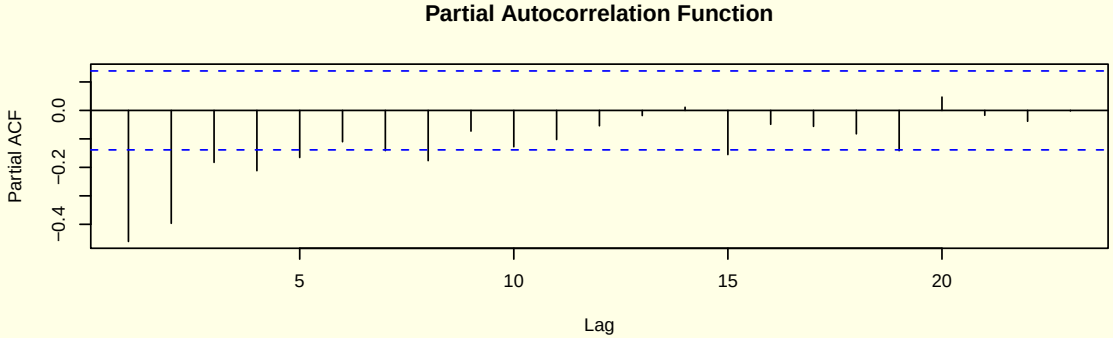

AR(1), rho=0.9
arpos9.first.order
−4
−2
0
2
0
50
100
150
200
Time
Autocorrelation Function
ACF
0.0
0.4
0.8
0
5
10
15
20
Lag
Partial Autocorrelation Function
Partial ACF
0.0
0.4
0.8
5
10
15
20
Lag

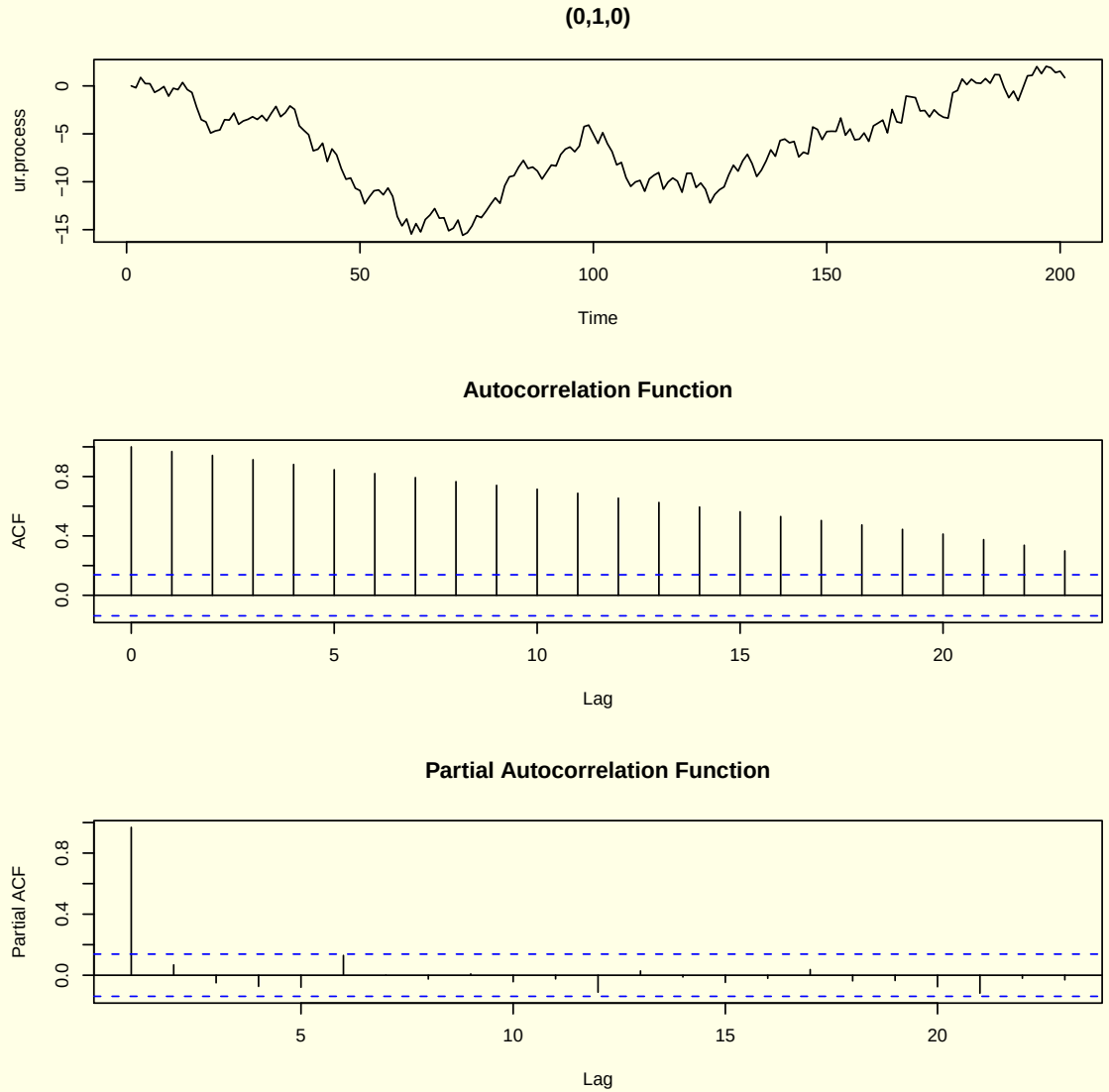

MA(1), theta=(0.9)
movavg.first.order
Time
Autocorrelation Function
ACF
Lag
Partial Autocorrelation Function
Partial ACF
Lag


AR(1), rho=−0.9
arneg9.first.order
−4
−2
0
2
4
0
50
100
150
200
Time
Autocorrelation Function
ACF
−0.5
0.0
0.5
1.0
0
5
10
15
20
Lag
Partial Autocorrelation Function
Partial ACF
−0.8
−0.4
0.0
5
10
15
20
Lag


MA(1), theta=(−0.9)
movavgneg.first.order
−4
−2
0
2
0
50
100
150
200
Time


Autocorrelation Function
ACF
−0.5
0.0
0.5
1.0
0
5
10
15
20
Lag


Partial Autocorrelation Function
Partial ACF
−0.4
−0.2
0.0
5
10
15
20
Lag


(0,1,0)
ur.process
0
−5
−10
−15
0
50
100
150
200
Time
Autocorrelation Function
ACF
0.0
0.4
0.8
0
5
10
15
20
Lag
Partial Autocorrelation Function
Partial ACF
0.0
0.4
0.8
5
10
15
20
Lag

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.

Dickey-Fuller tests

The Dickey-Fuller testing philosophy relies on the following base equation that mirrors our earlier basic presentation of random walks. To obtain a test equation, subtract y_{t-1} from both sides.

$$y_t = \alpha + \beta t + \rho_1 y_{t-1} + \epsilon_t \quad (31)$$

$$\Delta(y_t) = \alpha + \beta t + (\rho_1 - 1) y_{t-1} + \epsilon_t \quad (32)$$

The α – drift – and β – trend – terms are optional depending on the series in question. If and only if $\rho - 1 < 0$ or $\rho < 1$ can we reject the claim of nonstationarity.

The critical values are tabulated.

KPSS

Is unique in having a null hypothesis of [trend/drift/trend and drift] stationary. It is also fairly easy to construct.

1. Regress y on the chosen option above [trend/drift/trend and drift] and isolate the residual u.

2. Calculate S_t, the partial sum of the residuals:

$$S_t = \sum_{i=1}^{t} u_i$$

3. The KPSS statistic is

$$KPSS = \frac{1}{T^2} \sum t = 1^T \frac{S_t}{s_{u^2}}$$

where s_{u^2} is an estimate of the long-run variance (typically done by a Newey-West procedure).

4. If KPSS is large, reject the claim of *** stationary.

The critical values are tabulated.

Perron

A stationary time-series may look like non-stationary when there are structural breaks in the intercept or trend.

The unit root tests lead to false non-rejection of the null when we do not consider the structural breaks. A low power problem.

A single known breakpoint in 1987, Perron extended it to a case of unknown breakpoint in the 1990s.

Perron considers the null and alternative hypotheses

H0: $y_t = a_0 + y_{t-1} + \mu_1 D_P + \epsilon_t$ $(y_t \sim Stationary$ with a jump)
H1: $y_t = a_0 + a_2 t + \mu_2 D_L + \epsilon_t$ $(y_t \sim TS$ with a jump)
Pulse break: $D_P = 1\ if\ t = TB + 1$ and zero otherwise,
Level break: $D_L = 0\ for\ t = 1, ..., TB$ and one otherwise.

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, $nR^2 \sim \chi^2_p$).

- The above alongside:
  (1) is the temporal process common or distinct? and
  (2) if distinct, how and why?

Panel Unit Root Testing

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 $T \to \infty$ 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 ρ to be ρ_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.

Language of Time Series

Trend : pattern exists when there is a long-term increase or decrease in the data.

Seasonal : pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week).

Cyclic : pattern exists when data exhibit rises and falls that are not of fixed period (duration usually of at least 2 years).

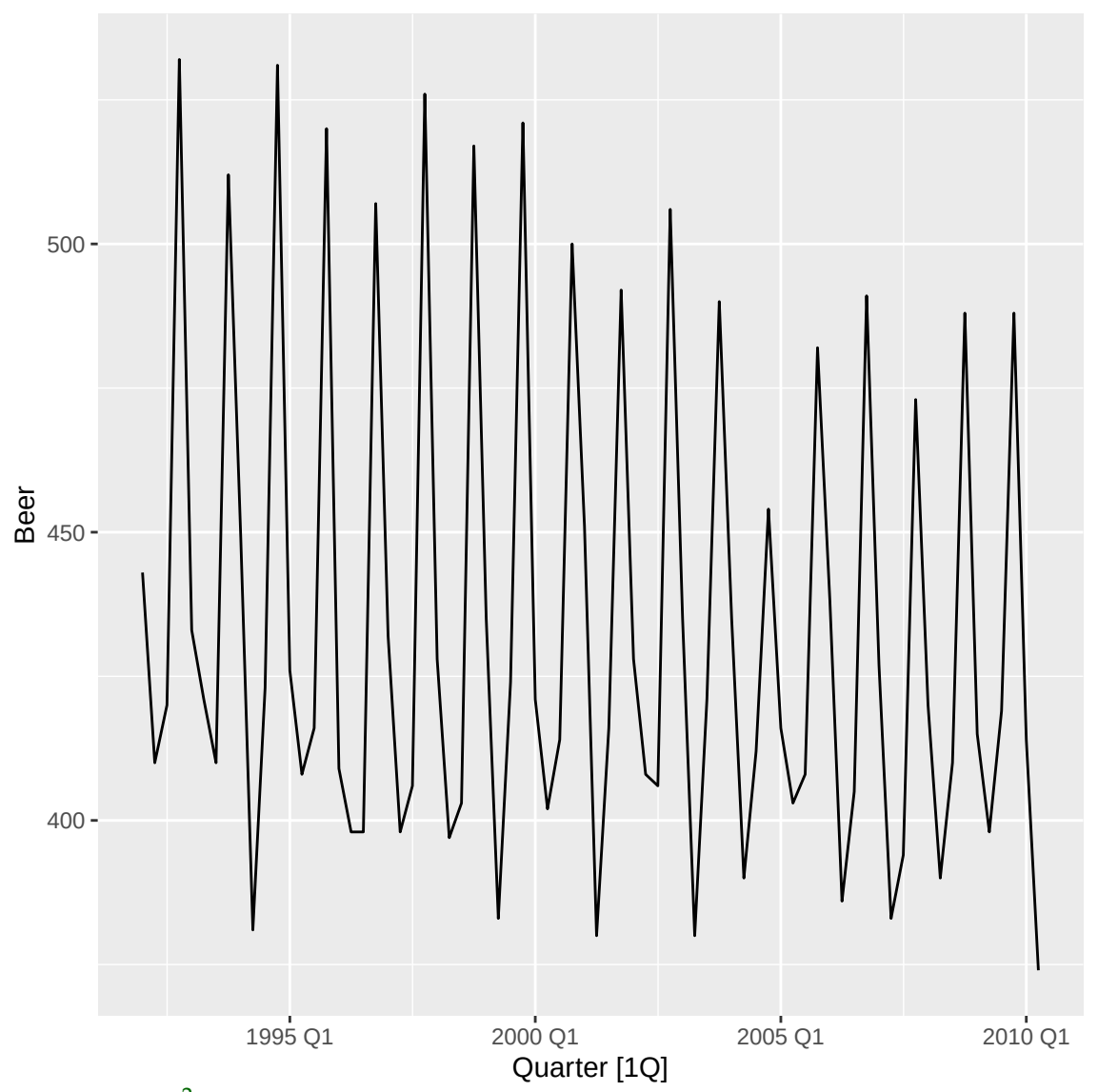
Visuals: Seasonal plots

Data plotted against the individual "seasons" in which the data were observed. (In this case a "season" is a month.)

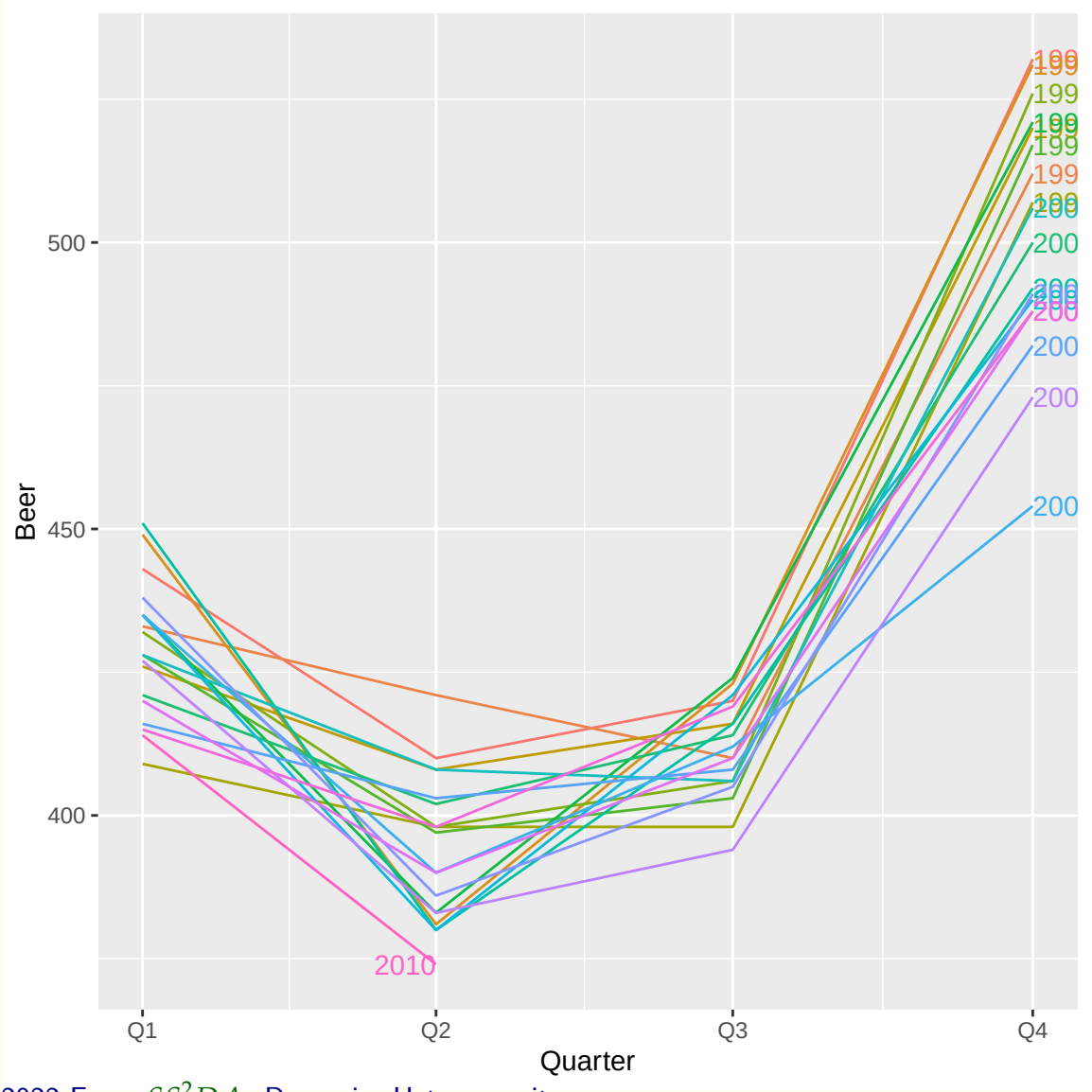
Something like a time plot except that the data from each season are overlapped.

Enables the underlying seasonal pattern to be seen more clearly, and also allows any substantial departures from the seasonal pattern to be easily identified.

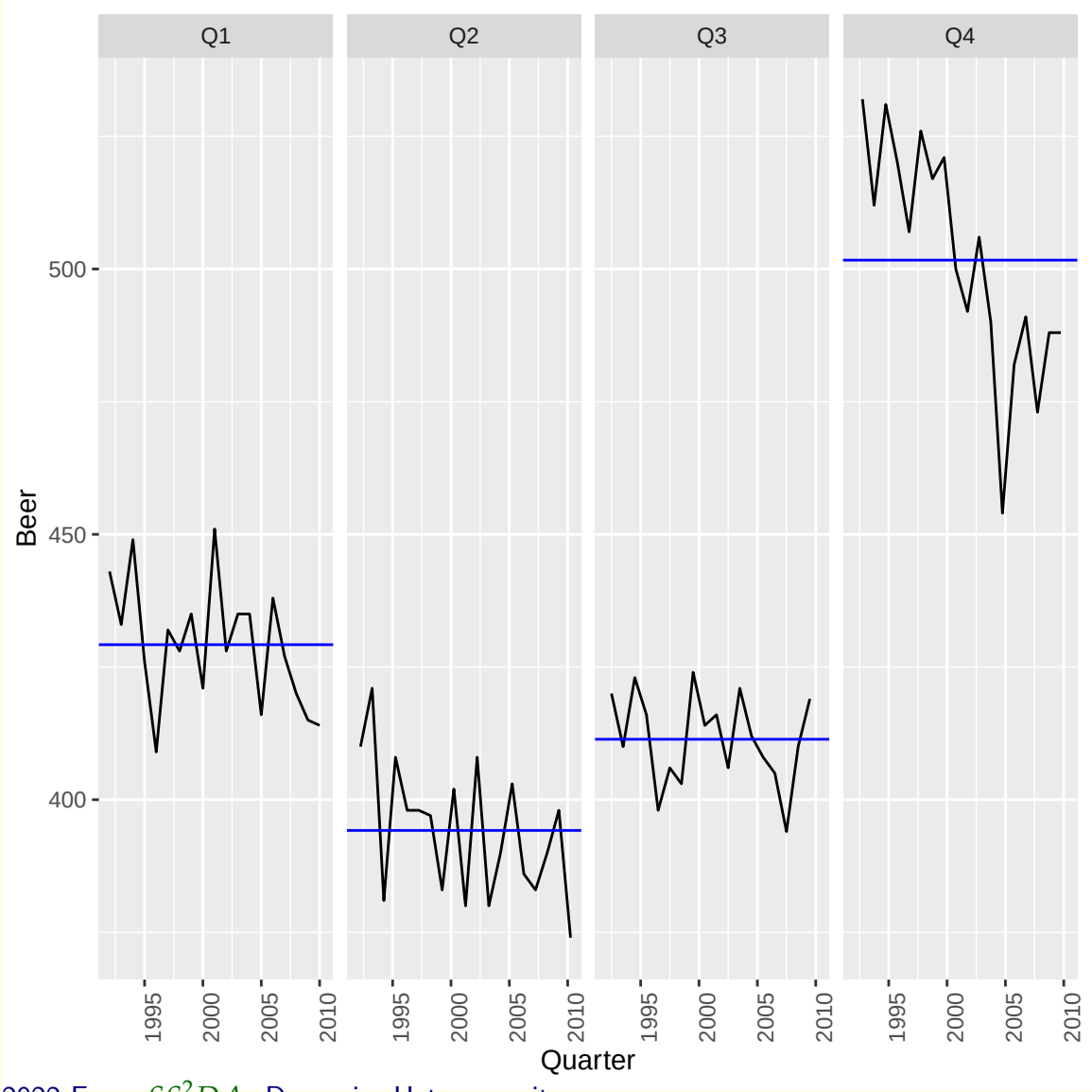
```
> library(fpp3); library(tidyverse)
> beer <- aus_production %>%
+   select(Quarter, Beer) %>%
+   filter(year(Quarter) >= 1992)
> beer %>% autoplot(Beer)
```


Beer
500
450
400
1995 Q1
2000 Q1
2005 Q1
2010 Q1
Quarter [1Q]

```
> beer %>% gg_season(Beer, labels="right")
```


Beer
500
450
400
Q1
Q2
Q3
Q4
Quarter
2010
200

```
> beer %>% gg_subseries(Beer)
```


Q1
Q2
Q3
Q4
500
450
400
Beer
1995
2000
2005
2010
Quarter

Day 3: Dynamic Models

The ARIMA approach is fundamentally inductive. The workflow involves the use of empirical values of ACF's and PACF's to engage in model selection. Dynamic models engage theory/structure to impose more stringent assumptions for producing estimates.

Time Series Linear Models/Dynamic Models

First, a result.

Aitken's Theorem?

In a now-classic paper, Aitken generalized the Gauss-Markov theorem to the class of Generalized Least Squares estimators. It is important to note that these are GLS and not FGLS estimators. What is the difference? The two GLS estimators considered by Stimson are not strictly speaking GLS.

Definition:

$$\hat{\beta}_{GLS} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y} \tag{33}$$

Properties: (1) GLS is unbiased. (2) Consistent. (3) Asymptotically normal. (4) MV(L)UE

A Quick Example

The variance/covariance matrix of the errors for a first-order autoregressive process is easy to derive. The matrix is banded; observations separated by one point in time are correlated ρ. Period two is ρ^2; the corners are ρ^{T-1}. The diagonal is one. What I have actually described is the correlation; the relevant autocovariances are actually defined by $\frac{\sigma^2\rho^s}{1-\rho^2}$ where s denotes the time period separation. It is also straightforward to prove (tediously through induction) that this is invertible; it is square and the determinant is non-zero having assumed that $|\rho < 1|$. The $2x2$ determinant is $\frac{1}{1-\rho^2}$. The $3x3$ is $1 * (1-\rho^2) - \rho(\rho - \rho^3) + \rho^2(\rho^2 - \rho^2)$. The first term is positive and the second term is non-zero so long as $\rho \neq 0$. But even if $\rho = 0$, we would have an identity matrix which is invertible.

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

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

If the variance is stationary, we can rewrite,

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

A comment on characteristic roots....

Cochrane-Orcutt

We have the two key elements to implement this except that we do not know ρ; we will have to estimate it and estimates have uncertainty. But it is important to note this imposes exactly an AR(1). If the process is incorrectly specified, then the optimal properties do not follow. Indeed, the optimal properties also depend on an additional important feature.

What does the feasible do?

We need to estimate things to replace unknown covariance structures and coverage will depend on properties of the estimators of these covariances. Consistent estimators will work but there is euphemistically "considerable variation" in the class of consistent estimators. Contrasting the Beck and Katz/White approach with the GLS approach is a valid difference in philosophies.[2]

[2] We will return to this when we look at Hausman because this is the essential issue.

Prais-Winsten/Cochrane-Orcutt

$$y_{it} = X_{it}\beta + \epsilon_{it}$$

where

$$\epsilon_{it} = \rho\epsilon_{i,t-1} + \nu_{it}$$

and $\nu_{it} \sim N(0, \sigma_\nu^2)$ with stationarity forcing $|\rho| < 1$. We will use iterated FGLS. First, estimate the regression recalling our unbiasedness condition. Then regress $\hat{\epsilon}_{it}$ on $\hat{\epsilon}_{i,t-1}$. Rinse and repeat until ρ doesn't change. The transformation applied to the first observation is distinct, you can look this up.... In general, the transformed regression is:

$$y_t - \rho y_{t-1} = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + \nu_t$$

with ν white noise.

Incremental Models

$$y_t = a_1 y_{t-1} + \epsilon_t$$

is the simplest dynamic model but it cannot be estimated consistently, in general terms, in the presence of serial correlation. Why? The key condition for unbiasedness is violated because $\mathbb{E}(y_{t-1}\epsilon_t) \neq 0$. OLS will not generally work.

Incremental Models with Covariates

$$y_t = a_1 y_{t-1} + \beta X_t + \epsilon_t$$

The problem is fitting and the key issue is white noise residuals post-estimation. But we have to assume a structure and implement it.

Distributed Lag Models

$$y_t = \alpha + \beta_0 X_t + \beta_1 x_{t-1} + \ldots + \epsilon_t$$

The impact of x occurs over multiple periods. It relies on theory, or perhaps analysis using information criteria/F [owing to quasi-nesting and missing data]. OLS is a fine solution to this problem but the search space of models is often large.

In response to this problem, we have structured distributed lag models; there are many such schemes.

- Koyck/Geometric decay:
  short run and long-run effects are parametrically identified

$$y_t = \alpha + \beta(1-\lambda)\sum_{j=0}^{\infty} \lambda^j X_{t-j} + \epsilon$$

- Almon (more arbitrary decay):

$$y_{it} = \sum_{t_A=0}^{T_F} \rho_{t_A} x_{t-t_A} + \epsilon_t$$

with coefficients that are ordinates of some general polynomial of degree $T_F >> q$. The $\rho_{t_A} = \sum_{k=0}^{T_F} \gamma_k t^k$.

Autoregressive Distributed Lag Models

$$y_t = \alpha + \gamma_1 y_{t-1} + \beta_0 X_t + \beta_1 x_{t-1} + \ldots + \epsilon_t$$

OLS is often used if iid; ϵ_t is unrelated to y_{t-1}. If not iid: GLS is needed. The authors argue that the lagged dependent variable often yields white noise for free. As they also note, there is a deBoef and Keele paper showing the relationship between these models and a form of error correction models. More on that tomorrow.

Structural vs. Non-structural

Data analysis can quite yield models comparisons among competing dynamic structures. The key issue is that the analyst need divine the process; what is the relevant error process and what is the structure and timing of effects alongside the potential question of incremental adjustment. We need good theory for that.

Given such theory, we can take an equations as analysis approach, measure the variables, and derive reduced forms, and then recover parameter estimates deploying simultaneous equations methods. Very large such systems were a core part of early empirical macroeconomics. The failures of such systems led to the proposal of alternatives.

Chris Sims suggested a more flexible approach: the VAR.

On Reduced and Structural Forms and Systems of Equations

What are structural forms and what are reduced forms? Are we familiar with what the rank and order conditions are?

Systems of Simultaneous Equations

Which beg questions about identification as above. Let's briefly discuss the system that they present.

Vector AutoRegression

Choose a relevant set of lag lengths and write each variable in the system as a function of lags of itself and other variables to the chosen lengths. The key insight is that this VAR is the reduced form to some more complicated as yet unspecified structural form. But if the goal is to specify how variables related to one another and to use data to discover Granger causality and responses to impulse injected in the system.

On Granger Causation

Granger causation[3] in a panel data setting. This is a rather complex topic because of heterogeneity.

- An identical causal relationship could exist for each $i \in N$
- No causal relationship could exist for any $i \in N$.
- Something in between the above two extremes.

Matching the theoretical claim and the empirical analysis here, like everywhere else, is absolutely crucial. Also, in some ways, this is just our earliest ANOVA example but now we will use the lags of the dependent variable to establish the alternative hypotheses.

[3] In the remainder of this discussion, when I use the word cause, I mean it in a Granger sense which may differ dramatically from more common understandings of causal.

Implementation

- Look at stationarity. We want to make certain that this holds.
- Test a hypothesis, many forms can be implied.

Conditional on stationarity, the procedure goes like this. For example, to test the hypothesis that, for all i, x does not Granger cause y, we choose a lag length, call it k. We then regress lags of y and k lags of x on y in a model with unit specific slopes and unit specific intercepts. We do not include contemporaneous x because Granger causality is all about temporal priority. This is the unrestricted model. The hypothesis implies the restriction that all the coefficients on all of the lagged x are zero for each cross-section and for all k lags of x.

In other words, the unrestricted model is:

$$y_{it} = \rho_{i,1} y_{i,t-1} + \rho_{i,2} y_{i,t-2} + \ldots + \rho_{i,m} y_{i,t-m} + \beta_{i,1} x_{i,t-1} + \ldots + \beta_{i,k} x_{i,t-k} + \epsilon_{it}$$

while the restricted model is

$$y_{it} = \rho_{i,1} y_{i,t-1} + \rho_{i,2} y_{i,t-2} + \ldots + \rho_{i,m} y_{i,t-m} + \epsilon_{it}$$

because the null hypothesis implies that all the β are zero for all lags $t-1$ to $t-k$ lags of x.

Backfilling the Lagged Dependent Variable problem

Let's analyse this. This will preview dynamic interpretation in panel settings also. The long-run multiplier.

Let's Talk about simultaneous equations and systems of equations.

Stationarity Invertibility

On Equilibrium

What does equilibrium mean? Equilibration?

Cointegration

- Is the alternative conjecture for simultaneously analysing the short run and the long-run.
- Is premised on a dynamic system of sorts.
- The idea was largely responsible for winning Clive Granger the Nobel.

Cointegration: The Practice

We begin with a need for nonstationary data. If the data were stationary, they are "self-equilibrating".

Cointegration: The Criteria

- Some linear combination must yield a stationary series.
- Variables must be integrated of the same order.
- With more than two variables, there can be multiple cointegrating vectors linking sets of variables.

Cointegration: The Process

The ECM is:

1. Pretest the variables for the order of integration. If integrated, then

2. Perform a seemingly troubled regression: the spurious regression of

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t$$

3. Is ϵ_t stationary? Check using

$$\Delta\hat{\epsilon}_t = \alpha_1 \hat{\epsilon}_{t-1}$$

4. If all complies, then estimate the ECM:

$$\Delta x_t = \alpha + \gamma_{xy}\hat{\epsilon_{t-1}} + G + \epsilon C x$$
$$\Delta y_t = \alpha + \gamma_{yx}\hat{\epsilon_{t-1}} + G + \epsilon C y$$

where G is the set of appropriate lagged differences in an OLS VAR [an aside on instrumental variables here]

5. Assess model adequacy.

Multiple Cointegrating Vectors: Johansen

Non-integrated ECM?

de Boef and Keele [referenced in BFHP on p. 90]