Lecture 10: Causality, and Way too little Time Series

D. Alex Hughes

December 9, 2014

1. Put a Hat on it
   - Causal Models
2. Time Series
   - Motivation
   - Simplifying Parameters
   - GLS
   - Do I have serially correlated errors?
   - How Many Observations?
3. The Final Exam

Broadening the Example

- What if schooling has a different effect for everyone?
- People can take on different amounts of the treatment?

$$Y_{si} \equiv f_i(s)$$

Where

- f_i is the individuals' return to schooling
- (s) is the amount of schooling received

- This is a considerably more general statement than the equivalence across all and only one level of treatment

The **CIA** now becomes:

$$Y_{si} \perp s_i | X_i, \forall s$$

This will help us assess situations where s is assigned conditional on X

Broadening the Example

- In this setup, the *average causal effect* of one more year of schooling is
$$E[f_i(s) - f_i(s-1)|X_i]$$
- And four years is
$$E[f_i(s) - f_i(s-4)|X_i]$$
- We will only ever observe $Y_i = f_i(s)$, but if the **CIA** holds, then average earnings across schooling levels have a causal interpretation.
$$E[Y_i|X_i, s_i = s] - E[Y_i|X_i, s_i = (s-1)]$$
$$= E[f_i(s) - f_i(s-1)|X_i]$$
- **Stop.** Look how powerful that is!

Broadening the Example

If the variable you are assessing is independent of potential earnings conditional on X_i, then selection bias vanishes, and and that variable has a causal interpretation.

- Notice that we have the ability - with this set up - to estimate causal effects at *all* values of education.
- This is a lot of potential effects.

This dynamic system leads us back to regression because some τ can summarize the effect of education across all the values.

Broadening the Example

Suppose we have a linear, constant effects, causal model of the form:

$$f_i(s) = \alpha + \tau s + \epsilon_i$$

- This says that $f_i(s)$ is linear, and the effect (τ) is the same for everyone, and the same across all levels of s
- Then, the only individual specific term in $f_i(s)$ is ϵ_i the unobserved things that determine earnings

If we make the substitution of an individual's education into the equation above we have:

$$Y_i = \alpha + \tau s_i + \epsilon_i$$

Which looks like a bi-variate regression, but it is *explicitly* linked to the causal model above! This model would have a causal interpretation.

Causal Models

What limitations are there in the last causal model?

- Of course, there is likely to be selection into $s_i \rightarrow cov(s_i, \epsilon_i)! = 0$
- Ability & Family History

What if we break ϵ_i into parts?

- Observable parts X_i
- Unobservable parts υ_i

Then:

$$\epsilon_i = X_i\gamma + \upsilon_i$$

Where, γ are regression coefficients of ϵ_i, on X_i, and so $cov(\upsilon_i, X_i) = 0$

Causal Models

What limitations are there in the last causal model?

- Of course, there is likely to be selection into $s_i \rightarrow cov(s_i, \epsilon_i) ! = 0$
- Ability & Family History

What if we break ϵ_i into parts?

- Observable parts X_i
- Unobservable parts υ_i

Then:

$$\epsilon_i = X_i\gamma + \upsilon_i$$

Where, γ are regression coefficients of ϵ_i, on X_i, and so $cov(\upsilon_i, X_i) = 0$

Causal Models

Finally, we can write:

$$\begin{aligned} E[f_i(s)|X_i, s_i] = E[f_i(s)|X_i] &= \alpha + \tau s + E[\epsilon_i|X] \\ &= \alpha + \tau s + E[X_i\gamma + \upsilon_i] \\ &= \alpha + \tau s + X_i\gamma \end{aligned}$$

In this form, the observed values Y_i take the form:

$$Y_i = \alpha + \tau s + X_i\gamma + \upsilon_i$$

and *upsilon* is uncorrelated with either s or X, and has a causal interpretation...

... if we believe that the only reason s_i and ϵ_i were correlated was the vector of X_is.

Causal Models

Finally, we can write:

$$\begin{aligned} E[f_i(s)|X_i, s_i] = E[f_i(s)|X_i] &= \alpha + \tau s + E[\epsilon_i|X] \\ &= \alpha + \tau s + E[X_i\gamma + \upsilon_i] \\ &= \alpha + \tau s + X_i\gamma \end{aligned}$$

In this form, the observed values Y_i take the form:

$$Y_i = \alpha + \tau s + X_i\gamma + \upsilon_i$$

and *upsilon* is uncorrelated with either s or X, and has a causal interpretation...
... if we believe that the only reason s_i and ϵ_i were correlated was the vector of X_is.

Motivation

The assumptions of the core OLS model assume a particular distribution of the residuals:

$$\sigma_\epsilon^2 = \mathbf{I}_n \sigma_\epsilon^2$$

Which we know has certain asymptotic and sample properties, and which we can estimate.

We have built more on to allow for more general data-generating processes:

- In the case of **Heteroskedastic Error Variance** we have allowed the disturbance term to be distributed as

$$\sigma_\epsilon^2 = \mathbf{I}_n \sigma_i$$

  for all the $i \in k$ where k index the RHS regressors.
- We could figure out how to estimate this (WLS, Robust SE, ...).

Motivation

But, much of this has been in the context of a single collection (*cross-section*) of data

- In many political science applications, we observe our data through time, potentially observing the same unit several times
- Does this give us any ability to employ a *scientific* solution to the FPCI? Can we get closer?

Motivation

Definition

Panel Data is data that has multiple observations of the same individual(s) collected through time.

If you want names for things:

1. A **cross-section** (what we have worked with to this point) has a single observation of many cases
2. A **time-series** dataset has repeated observations of a single case
3. A **panel** has repeated observations of many cases

Motivation

Some of the largest/most reputable polls in political science contain explicit panels

- ANES

And the most famous internal case can well be thought of as a series of panels

- All IR datasets – country/year observations

Motivation

So, why would just throwing this into an OLS be inappropriate?

- OLS assumes the ϵ are independent
- Very likely not the case in a time series: Time periods that are closer to one another and likely to be more simliar than time periods that are further from one another.
- Analogy with geographic proximity
- Analogy with social proximity
- Heteroskedasticity across unit-repeated observations and time

So what?

Motivation

So, why would just throwing this into an OLS be inappropriate?

- OLS assumes the ϵ are independent
- Very likely not the case in a time series: Time periods that are closer to one another and likely to be more simliar than time periods that are further from one another.
- Analogy with geographic proximity
- Analogy with social proximity
- Heteroskedasticity across unit-repeated observations and time

So what?

So what?

If we have strage covariance structures, that vary across units and time

- $E[\hat{\beta}_{OLS}]$, and
- $V[\hat{\beta}_{OLS}]$

might give wacky results.

But, if we *exactly* specify what we think is going on, there are too many parts to ever estimate.

Let's do some work to show this

Relaxing Sigma

In the past we've made somewhat stronger assumptions on the shape of Σ. Let's relax those.

$$past : \mathbf{y} = \mathbf{X}\beta + \mathbf{I}\epsilon_{\mathbf{OLS}}$$
$$current : \mathbf{y} = \mathbf{X}\beta + \mathbf{\Sigma}_{\epsilon\epsilon}$$

$$V[\epsilon_{OLS}] = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k^2 \end{bmatrix} \quad V[\Sigma_{\epsilon\epsilon}] = \begin{bmatrix} \sigma_1^2 & \rho_{1,2} & \cdots & \rho_{1,k} \\ \rho_{2,1} & \sigma_2^2 & \cdots & \rho_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k,1} & \rho_{k,2} & \cdots & \sigma_k^2 \end{bmatrix}$$

$\epsilon \sim N_n(0, \Sigma_{\epsilon\epsilon})$

- We're looking for some transformation that will move from

$$\mathbf{\Sigma}_{\epsilon\epsilon} \rightarrow \mathbf{I}\sigma^2$$

- If we can find that, we can just run with OLS again. Boom.

Relaxing Sigma

So, continuing... We're looking for a Γ that will transform $\Sigma_{\epsilon\epsilon}^{-1} \rightarrow \sigma^2$. Easy. This is like finding some "square-root" of Σ^{-1}.

Linear algebra, quadratic form, this is looking for some $\Gamma'\Gamma = \Sigma_{\epsilon\epsilon}^{-1}$.

$$\begin{aligned} V[\Gamma\Sigma] &\rightarrow \sigma^2 I_{(n)} \\ &= \Gamma V[\Sigma]\Gamma' \\ &= \sigma^2 \Gamma\Sigma\Gamma \end{aligned}$$

And so, Γ is what we were looking for. This Γ will let us run with OLS if we correct everything according to Γ.

Proofing – like a baker

$$\Gamma Y = \Gamma X\beta + \Gamma\epsilon$$
$$\Gamma\epsilon = \Gamma Y - \Gamma X\beta$$

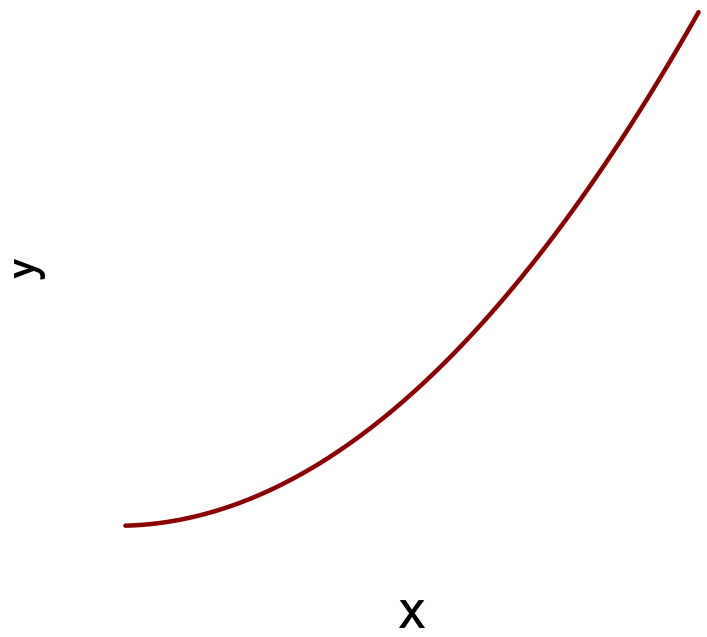
We're in a spot with a derivation like OLS.

$$\begin{aligned}
arg\ min_{\{\beta\}}(\Gamma\epsilon)^2 &= (\Gamma Y - \Gamma X\beta)'(\Gamma Y - \Gamma X\beta) \\
&= Y'\Gamma'\Gamma Y - Y'\Gamma'\Gamma X\beta - \beta' X'\Gamma'\Gamma Y - \beta' X'\Gamma'\Gamma X\beta \\
&= Y'\Gamma'\Gamma Y - 2 * \beta' X'\Gamma'\Gamma Y - \beta' X'\Gamma'\Gamma X\beta \\
\frac{\partial}{\partial\beta'} &= 0 - 2X'\Gamma'\Gamma - 2X'\Gamma'\Gamma X\beta = 0 \\
&= X'\Gamma'\Gamma Y - X'\Gamma'\Gamma X\beta = 0 \\
&= X'\Gamma'\Gamma X\beta = X'\Gamma'\Gamma Y \\
\beta_{GLS} &= (X'\Gamma'\Gamma X)^{-1} X'\Gamma'\Gamma Y \\
\beta_{GLS} &= (X'\Sigma_{\epsilon\epsilon}^{-1} X)^{-1} X'\Sigma_{\epsilon\epsilon}^{-1} Y
\end{aligned}$$

Are we out of the woods? Like a baker?

As ever, we don't typically *know* $\Sigma_{\epsilon\epsilon}$. So we estimate it.

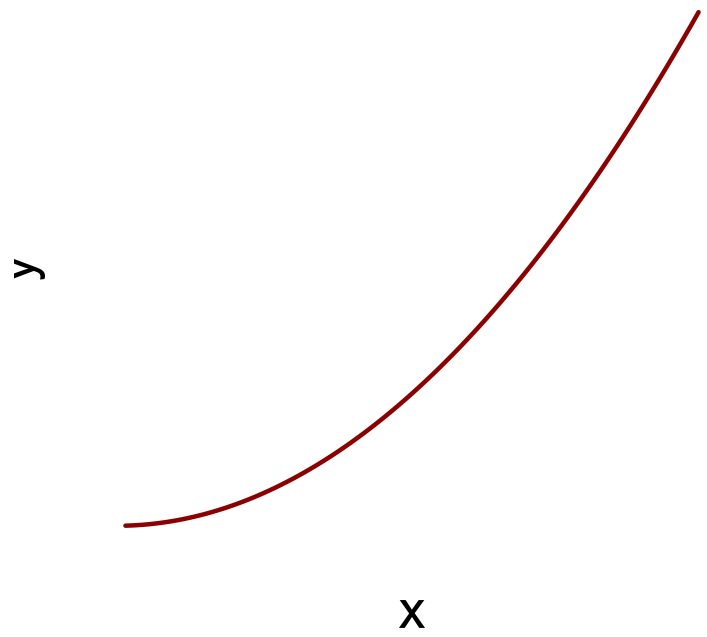
- But, look at the dimensions on $Sigma_{\epsilon\epsilon}$ it's $(n \times n)$. Uh. oh.
- If all elements were distinct, there would be n^2 things to estimate.
- Luckily, it is symmetric so we only have to estimate a diagonal and one triangle.
- Unluckily, the there are $\frac{n(n+1)}{2}$ parameters to plot



Are we out of the woods? Like a baker?

As ever, we don't typically *know* $\Sigma_{\epsilon\epsilon}$. So we estimate it.

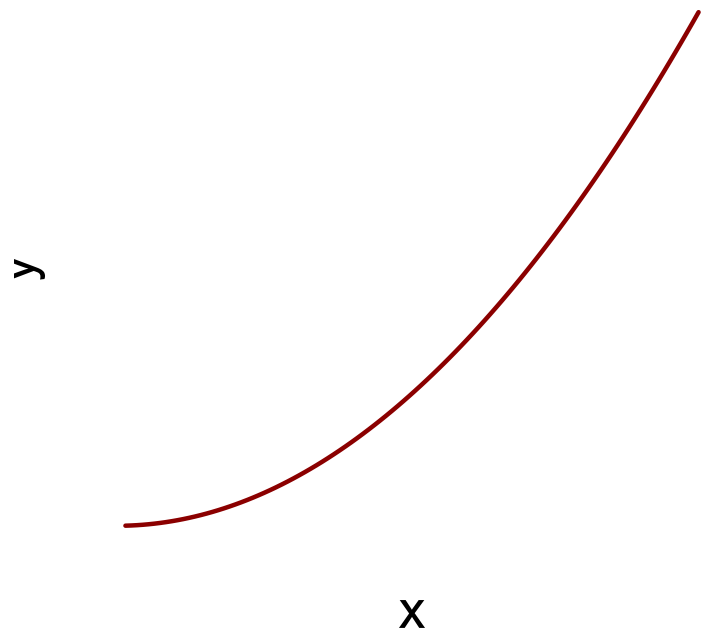
- But, look at the dimensions on $Sigma_{\epsilon\epsilon}$ it's $(n \times n)$. Uh. oh.
- If all elements were distinct, there would be n^2 things to estimate.
- Luckily, it is symmetric so we only have to estimate a diagonal and one triangle.
- Unluckily, the there are $\frac{n(n+1)}{2}$ parameters to plot



Are we out of the woods? Like a baker?

As ever, we don't typically *know* $\Sigma_{\epsilon\epsilon}$. So we estimate it.

- But, look at the dimensions on $Sigma_{\epsilon\epsilon}$ it's $(n \times n)$. Uh. oh.
- If all elements were distinct, there would be n^2 things to estimate.
- Luckily, it is symmetric so we only have to estimate a diagonal and one triangle.
- Unluckily, the there are $\frac{n(n+1)}{2}$ parameters to plot



Nope. We're not out of the woods

Unless we can represent the information in this matrix with some lower-order representation, we will just *never* have enough data to estimate this.

- In past contexts we've been worried about *overfitting* data
- This is a different concern – this isn't a fully determined system – we've got more things to estimate than we have data
- Like solving for $y = 2x$. Non unique solutions abound!

So, what to do?

Are there simplifying statements that we can make to collapse all these estimands into a lower-order space? Are the errors **Stationary Errors:**

1. Do all the error have the same expectation (0)?
2. Do all the errors have a common variance (σ_ϵ^2)
3. Does the covariance between errors depend only on their seperation in time?

$$Cov(\epsilon_t, \epsilon_{t+s}) = E[\epsilon_t \epsilon_{t+s}] = \sigma_\epsilon^2 \rho_s = Cov(\epsilon_t, \epsilon_{t-s})$$

Where ρ_s is the *autocorrelation* or *serial correlation* between ϵ_t and ϵ_s.

$$\Sigma_{\epsilon\epsilon} = \sigma_\epsilon^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix}$$

Comparison

So, by assuming *stationarity* we have more-or-less cut the number of parameters to estimate by a factor of $\frac{2}{\sqrt{n}}$.

- Non-stationary: $\frac{n(n+1)}{2}$
- Stationary: n

This is closer, but if we have to estimate functional parameters (β), as well, we're still sunk.

Are there additional assumptions we can make?

Auto Regressive Processes

Definition

An **autoregressive process** is a data generating procedure where the error in period t (ϵ_t) depends only on the error in the previous term (ϵ_{t-1}) and some random component ν_t.

Often abbreviated $AR(1)$:

$$\epsilon_t = \rho\epsilon_{t-1} + \nu_t$$

Where $\nu_t \sim N(0, \sigma_\nu^2)$ (unlike ϵ_t).

What do we get?

If the errors are stationary – which in an $AR()$ process requires $|\rho| < 1$, then we can compactly write (and consequently estimate) the components of the autocorrelated errors.

What would happen if $|\rho| >= 1$?

$$\begin{aligned}\sigma_\epsilon^2 \equiv V(\epsilon_t) &= E(\epsilon_t^2 - 0) \\ &= V(\epsilon_{t-1}) \\ &= E(\epsilon_{t-1}^2)\end{aligned}$$

By this assumption, we've collapsed everything down in a very small number of parameters.

What do we get?

If the errors are stationary – which in an $AR()$ process requires $|\rho| < 1$, then we can compactly write (and consequently estimate) the components of the autocorrelated errors.

What would happen if $|\rho| >= 1$?

$$\begin{aligned} \sigma_\epsilon^2 \equiv V(\epsilon_t) &= E(\epsilon_t^2 - 0) \\ &= V(\epsilon_{t-1}) \\ &= E(\epsilon_{t-1}^2) \end{aligned}$$

By this assumption, we've collapsed everything down in a very small number of parameters.

What do we get?

We can nicely represent the variance of the "goal" regression in terms of the autocorrelation and the variance of the random shock ν.

$$\begin{aligned} E[\epsilon_t^2] &= \rho^2 E[\epsilon_{t-1}^2] + E[\nu_t^2] + 2\rho E[\epsilon_{t-1}\nu_t] \\ \sigma_\epsilon^2 &= \rho^2 \sigma_\epsilon^2 + \sigma_\nu^2 \\ &= \frac{\sigma_\nu^2}{1-\rho^2} \end{aligned}$$

We get that we can represent the entire variance structure with only two estimands: σ_ν^2 and ρ.

What do we get?

We can nicely represent the variance of the "goal" regression in terms of the autocorrelation and the variance of the random shock ν.

$$
\begin{aligned}
E[\epsilon_t^2] &= \rho^2 E[\epsilon_{t-1}^2] + E[\nu_t^2] + 2\rho E[\epsilon_{t-1}\nu_t] \\
\sigma_\epsilon^2 &= \rho^2 \sigma_\epsilon^2 + \sigma_\nu^2 \\
&= \frac{\sigma_\nu^2}{1-\rho^2}
\end{aligned}
$$

We get that we can represent the entire variance structure with only two estimands: σ_ν^2 and ρ.

Doctor, is this normal?

How would you know if you have serially correlated errors?

- Theory; Common-sense
- Not all panel data have serially correlated errors; but most do

Is there a test for it? Yes.

- Like always, look at the residuals from the OLS regression
- Fit a model using OLS (not including time)

Doctor, is this normal?

How would you know if you have serially correlated errors?

- Theory; Common-sense
- Not all panel data have serially correlated errors; but most do

Is there a test for it? Yes.

- Like always, look at the residuals from the OLS regression
- Fit a model using OLS (not including time)

How many observations?

You **will** at some point present a panel dataset. And someone **will** say, "Well, you've got the right model specified. I see you took 204b at UCSD. But, how many *effective* observations do you have? Does the time component really help you?"
Here's what you calculate, and then respond!

- n matter in so far as it decreases the standard errors of a distribution.
- $SE_y = \frac{\sigma_y^2}{n}$
- But if you've got AR() data, $SE_y = \frac{\sigma^2}{n} \times \frac{1+\rho}{1-\rho}$
- σ^2 is the same, but the "*effective* number of observations" $n^{\#} = n \times \frac{1-\rho}{1+\rho}$.

Put another way: If you're considering running a panel, and you are cost-conscious – more data will not always be better. To the extent that you are measuring a high ρ AR() processes, you aren't actually getting more *data* in your data.

Final Exam

- Comprehensive
- All material fair game
- No explicit programming

Final Exam

1. Introduction to Data
   - Descriptive & Stats
   - Simple Cross Tabs (and χ^2)
   - Hand Drawn Plots
   - Probability
2. Hypothesis Tests and Inference
   - Test statistics
   - Hypotheses
   - Critical regions
   - P-values
   - Power
3. Estimators
   - Properties of Estimators
   - Expectation, Covariance, & Correlation
   - Basics of Regression
   - Regression Fit

Final Exam

4. Linear Algebra and Regression
   - Assumptions of Regression
   - Proofs of Desirable Properties
   - Hypothesis Testing
5. Joint Tests
   - Joint tests for significance
   - Diagnostics and outliers
   - ANOVA
6. Problems?
   - Problems in the errors
7. Causality
   - So much!
8. Serial Correlation