UCI Paul Merage School of Business

Leadership for a Digitally Driven World™

# MFIN 290: Financial Econometrics

Lecture 6-2

# Today

Being clever with Panel Data: Correcting Endogeneity problems with Excess Variation

Intro to Likelihood and Maximum Likelihood introduction

- Key concepts:
- What is MLE doing?
- What are the ways I can do hypothesis tests in an MLE problem?
- What are common measures of model performance?

# Hausman Taylor (1981)

- Imagine the following setup

- $y_{it} = x_{1it}\beta_1 + x_{2it}\beta_2 + z_{1i}\Gamma_1 + z_{2i}\Gamma_2 + \varepsilon_{it} + u_i$

- We have two types of time varying ($X$) and time-invariant ($Z$) data. Some $X$ data is "bad" = correlated with the group specific errors ($u_i$), and some that is not.

(this is a very well known paper. Usually IV papers are storytelling, but it's rare that you get instruments from a model's structure. Lagged dependent variables was another case)

# Hausman Taylor (1981)

- Imagine the following setup
-
- $y_{it} = x_{1it}\beta_1 + x_{2it}\beta_2 + z_{1i}\Gamma_1 + z_{2i}\Gamma_2 + \varepsilon_{it} + u_i$
- $u_i$ is the group specific error term
- $x_{1it}$ is time varying and uncorrelated with $u_i$
- $x_{2it}$ is time varying and correlated with $u_i$ - bad
- $z_{1i}$ is time invariant and uncorrelated with $u_i$
- $z_{2i}$ is time variant and correlated with $u_i$ - bad

# Hausman Taylor (1981)

- $y_{it} = x_{1it}\beta_1 + x_{2it}\beta_2 + z_{1i}\Gamma_1 + z_{2i}\Gamma_2 + \varepsilon_{it} + u_i$

- Because of the correlation with the error, we can't get consistent estimates of the coefficients on $x_{2it}$ or $z_{2i}$…

- But within group estimation can get unbiased $\beta_1$ and $\beta_2$ since $\overline{z_{ji}} = z_{ji}$ and those terms drop (along with the problematic correlated $u_i$ term)
- $y_{it} - \overline{y}_i = (x_{1it} - \overline{x}_{1i})\beta_1 + (x_{2it} - \overline{x}_{2i})\beta_2 + (\varepsilon_{it} - \overline{\varepsilon}_i)$

# Hausman Taylor (1981)

- $y_{it} = x_{1it}\beta_1 + x_{2it}\beta_2 + z_{1i}\Gamma_1 + z_{2i}\Gamma_2 + \varepsilon_{it} + u_i$

- $y_{it} - \overline{y}_i = (x_{1it} - \overline{x}_{1i})\beta_1 + (x_{2it} - \overline{x}_{2i})\beta_2 + (\varepsilon_{it} - \overline{\varepsilon}_i)$

We will also be able to get an estimate of the variance of $\varepsilon_{it}$ from this regression…

But we still have variation left over!

# Hausman Taylor (1981)

- $y_{it} = x_{1it}\beta_1 + x_{2it}\beta_2 + z_{1i}\Gamma_1 + z_{2i}\Gamma_2 + \varepsilon_{it} + u_i$

- Never used the between group variation!… We can instrument for the problematic $z_{2i}$ with the $\overline{x_{1i.}}$. We know this works because the $x_{1it}$ is uncorrelated with $u_i$, so we just need it to be correlated with the $z_{2i}$.

- $=>$ We can use the extra identifying variation in a panel to get around estimation issues and provide natural instruments!

- They find the returns to schooling twice what was previously estimated.

# Hausman Taylor (1981)

- $y_{it} = x_{1it}\beta_1 + x_{2it}\beta_2 + z_{1i}\Gamma_1 + z_{2i}\Gamma_2 + \varepsilon_{it} + u_i$

- Step 1: Estimate within estimator, acquire estimate of within group residual variance $\sigma_e^2$ and $\beta's$
- Step 2: Instrument for the problematic time invariant information with the group means of the non-problematic x variables (if it is correlated)
- Step 3: Residual variance from step 2 is given by $\sigma_u^2 + \sigma_e^2/T$.. And we have an estimate of $\sigma_e^2$
- Step 4: Use our (consistent) coefficient estimates to estimate the residual variance structure to apply FGLS with the estimates of $\sigma_u^2$ and $\sigma_e^2$

# Dynamic Panel Data

Earlier in the course, we talked about some of the issues with lagged dependent variables.

Let's return to this here.

$$y_{it} = x_{it}\beta + \delta y_{it-1} + c_i + \epsilon_{it}$$

# Dynamic Panel Data

$$y_{it} = x_{it}\beta + \delta y_{it-1} + c_i + \epsilon_{it}$$

Assumptions:

$E(\epsilon_{it} | X_i, c_i) = 0$ (Strict Exogeneity)

$E(\epsilon_{jt}\epsilon_{is} | X_i, c_i) = \sigma_\epsilon^2 \; if \; i = j$ and $t = s$, 0 otherwise (Homoskedasticity/nonautocorrelation)

$E(c_i | X_i) = h(X_i)$ (Common Effects)

We don't need time series effects anymore because of the panel structure (essentially embeds them)

# Dynamic Panel Data

$$y_{it} = x_{it}\beta + \delta y_{it-1} + c_i + \epsilon_{it}$$

Common error term $c_i + \epsilon_{it} = u_{it}$

Problem?

$y_{it-1}$ is correlated with this (because it contains $c_i + \epsilon_{it-1} = u_{it-1}$)

# Dynamic Panel Data

$$y_{it} = x_{it}\beta + \delta y_{it-1} + c_i + \epsilon_{it}$$

Problem $y_{it-1}$ is correlated with this because it contains $c_i + \epsilon_{it-1} = u_{it-1}$

In fact, we know:

$cov(y_{it-1}, u_i) = \sigma_c^2 + \delta cov(y_{it-2}, u_i)\ldots$

If $0 < \delta < 1$, this will tend to

$$\frac{\sigma_c^2}{1-\delta}$$

# Dynamic Panel Data

Fixed effects does not solve this problem as panel asymptotics are usually over N and not T!

Let's take first differences between the third and second observation (so we have the lagged term defined):

$$y_{i3} - y_{i2} = (x_{i3} - x_{i2})\beta + \delta(y_{i2} - y_{i1}) + \epsilon_{i3} - \epsilon_{i2}$$

Here, $y_{i2} - y_{i1}$ is correlated with $\epsilon_{i3} - \epsilon_{i2}$, so we need an instrument…

# Dynamic Panel Data

$$y_{i3} - y_{i2} = (x_{i3} - x_{i2})\beta + \delta(y_{i2} - y_{i1}) + \epsilon_{i3} - \epsilon_{i2}$$

Here, $y_{i2} - y_{i1}$ is correlated with $\epsilon_{i3} - \epsilon_{i2}$, so we need an instrument…

What about $y_{i1}$? Obviously correlated with $y_{i2} - y_{i1}$, uncorrelated with $\epsilon_{i3} - \epsilon_{i2}$ due to our non-autocorrelation assumption!

But there is more! Look at the next difference:

# Arellano-Bond Estimators (1991)

$$y_{i4} - y_{i3} = (x_{i4} - x_{i3})\beta + \delta(y_{i3} - y_{i2}) + \epsilon_{i4} - \epsilon_{i3}$$

Now we have BOTH $y_{i2}$ and $(y_{i2} - y_{i1})$ available as instruments.

Keep doing this, and we have the level and difference as an instrument for all but the first observation and the level for the first. We have A LOT of ways to estimate the model (but we can't just run OLS or GLS or fixed effects!).

# Arellano-Bond Estimators (1991)

In fact, we have T-2 + T-3 = 2T-5ways to do it for the Tth difference (T-2 in levels, T-3 in differences).

Consider only the level based IV estimators: there are T-2 ways that we can recover our coefficients. We can estimate each of these in a stacked IV form. The ultimate estimator is a matrix-weighted average of the individual IV estimates.

Intuition is similar to Hausman Taylor, though the notation is more complex.

# Arellano-Bond Estimators (1991)

Takeaways:

**Always check the validity of the exogeneity assumption when you have a lagged dependent variable.**

There are usually ways out, especially in a panel, but regular OLS, GLS, or Fixed Effects will not help.

Panel datasets can provide instruments given how over-identified the coefficients are, if you take the time series effects seriously!

# Griliches Hausman (1985)

- One other scenario where the exclusion/identifying restriction may be violated is with measurement error. We talked about this in a classical context already, so we know that all of our coefficient estimates will be unbiased/inconsistent in the presence of measurement error.
- In a multivariate specification, if any of the other variables covary with the one measured with error, they are also not able to be recovered. Note that this need not be what we think of. We may observe a variable precisely (say, GDP), but what it represents in our model ("Economic Conditions") it only reflects with noise. With this view, measurement error happens all the time!

- But a panel dataset can give us a way out!

# Griliches Hausman (1985)

- Measurement Error Recap:
- $y_{it} = x_{it}^{*}\beta + \varepsilon_{it}$
- Observe $x_{it} = x_{it}^{*} + u_{it}$
- Regress $y_{it} = x_{it}\beta + \varepsilon_{it} - u_{it}\beta = x_{it}\beta + w_{it}$

- We get unbiased betas if $cov(x_{it}, w_{it}) = 0$ …
- $cov(x_{it}, w_{it}) = cov(x_{it}^{*} + u_{it}, \varepsilon_{it} - u_{it}\beta) = -\beta\sigma_u^2$

# Griliches Hausman (1985)

- What if we difference our data?

- $y_{it} - y_{it-1} = (x_{it}^{*} - x_{it-1}^{*})\beta + \varepsilon_{it} - \varepsilon_{it-1}$

- Observe $x_{it} = x_{it}^{*} + u_{it}$

- This will be less accurate, as we have doubled our MSE… but what's the new bias?

# Griliches Hausman (1985)

- Regress $y_{it} - y_{it-1} = (x_{it} - x_{it-1})\beta + (\varepsilon_{it} - \varepsilon_{it-1}) - (u_{it} - u_{it-1})\beta$

- $Cov((x_{it} - x_{it-1}), (\varepsilon_{it} - \varepsilon_{it-1}) - (u_{it} - u_{it-1})\beta)$

# Griliches Hausman (1985)

- Regress $y_{it} - y_{it-1} = (x_{it} - x_{it-1})\beta + (\varepsilon_{it} - \varepsilon_{it-1}) - (u_{it} - u_{it-1})\beta$
- $Cov((x_{it} - x_{it-1}), (\varepsilon_{it} - \varepsilon_{it-1}) - (u_{it} - u_{it-1})\beta)$
- $Cov(x_{it}^{*} + u_{it} - x_{it-1}^{*} - u_{it-1}, (\varepsilon_{it} - \varepsilon_{it-1}) - (u_{it} - u_{it-1})\beta) = -2\,\beta\,\sigma_u^{2}$
- If we estimate both of these, we have two equations and two unknowns $(\beta, \sigma_u^2)$… if we take the functional form of the measurement error seriously, we can correct for it!

# Griliches Hausman (1985)

- The multivariate case is similar. Note that we can do this with nth differences, within estimators, etc. Many options here, including correlated measurement errors, etc.

- Why do I like this paper?

- The same process can apply for *almost any* case of endogeneity. If we take the specification seriously, the coefficients are over-identified in a panel and can be corrected. Inference is generally done via bootstrapping.

# Likelihood Functions

- So far, we have mainly focused on estimation in a particular order:

- We have a model, and we ask what parameter is most consistent with the model given the data that we have.

# Likelihood Functions

- So far, we have mainly focused on estimation in a particular order:
- We have a model, and we ask what parameter is most consistent with the model given the data that we have.
- This is related to the objective function (e.g., least squares), but at the end of the day we are finding a parameter that best fits the data.

# Likelihood Functions

- Imagine that we turned this around:
- Instead we asked *how likely our data would be* given particular parameters?

- Define the probability distribution of $y$ conditional on some set of parameters $\theta$ as

  - $f(y|\theta)$

# Likelihood Functions

- Define the probability distribution of $y$ conditional on some set of parameters $\theta$ as
  - $f(y|\theta)$
- If $y$ is i.i.d., then the joint probability of the data $y_1, y_2, \dots, y_n$ is the product of the marginals:
- $f(y_1, y_2, \dots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$
- Call this the "likelihood function": $L(y|\theta)$

# Likelihood Functions

- $f(y_1, y_2, \ldots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$
- $L(y|\theta)$
- The likelihood function is the probability that you would see your dataset, conditional on the key parameters being given by (the vector) $\theta$

# Likelihood Functions

- $f(y_1, y_2, \ldots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$
- We can estimate the parameters $\theta$ by asking what values of $\theta$ make our data the most likely? That is, which values maximize $L(y|\theta)$?
- Since logs are a monotonic transformation, we can maximize $L(y|\theta)$ by maximizing $\ln(L(y|\theta))$ which lets us break apart the product into a sum (among other things)

# Likelihood Functions

- Like any other problem, when we maximize the likelihood, we are searching for parameters that satisfy:

- $\frac{\partial \ln\left(L(y|\theta)\right)}{\partial \theta} = 0$

- The values of $\theta$ that satisfy this equation are known as the Maximum Likelihood Estimates, and $\hat{\theta}$ is the Maximum Likelihood Estimator

# Properties of the ML Estimator

- Intuitively, this makes some sense

- It turns out that the MLE- which again is the parameters that are most likely to give us the data we see - have meaningful asymptotic properties:

- Consistency
- Asymptotic Normality
- Asymptotic Efficiency
- Invariance

# Example: Normal Distribution

1) Normal PDF for a single observation as a function of parameters

What are the parameters here?

2) Multiply these together for likelihood/add them up for log likelihood

3) Differentiate with respect to each parameter, set FOC to zero

4) Solve for optimal parameters as a function of the raw data ($x, y, n$, etc.)

# Likelihood: Normal Distribution

- $y = x\beta + e$
- $e \sim N(0, \sigma^2)$

- $y \sim N(x\beta, \sigma^2)$

- $f(y_i|x, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-(y_i - x\beta)^2}{2\sigma^2})$

- Note: $\theta$ here has two parameters: $\mu = x\beta$ and $\sigma^2$

# Properties of the ML Estimator

- **Consistency**

- $plim\ \hat{\theta} = \theta$

- With enough data, your estimates will converge to the true value
- Asymptotic Normality
- Asymptotic Efficiency
- Invariance

# Properties of the ML Estimator

- **Asymptotic Normality**

- $\hat{\theta} \sim N(\theta, I(\theta)^{-1})$

- **Asymptotic Efficiency**
- $I(\theta)^{-1} = E(\frac{\partial^2 lnL}{\partial\theta\partial\theta'})$ = "Cramer Rao Lower Bound" (the lowest possible variance for a consistent estimate)

This is enough to conduct inference here as usual…

# Properties of the ML Estimator

- **Invariance**

- $\gamma = c(\theta)$ for any continuous and continuously differentiable c( )

- $\hat{\gamma} = c(\hat{\theta})$

This is an asymptotic property that will not hold in finite samples for generic functions.

In finite samples, this only holds for linear functions!

# Evaluating Fit

- There are typically three ways that we can evaluate likelihood for MLE estimates.

- Likelihood Ratio Tests
- Wald Tests
- LaGrange Multiplier Tests

- They are all different, but all the same…

# Evaluating Fit: Likelihood Ratio Test



- When we maximize $\ln(L(y|\theta))$ we end up with a value (called the log likelihood).
- Imagine adding a restriction to $\theta$: $\theta_R$ one of the parameters within is equal to zero
- We could re-maximize the likelihood in the presence of this constraint, and calculate a new value for $L(y|\theta_R) = L_R$

# Evaluating Fit: Likelihood Ratio Test



- We know that $L_R$ has to be less than $L_U$ (it was maximized before without a constraint), so that the ratio

- $\frac{L_R}{L_U} = \lambda$ must lie between zero and one

- Intuitively, if $\lambda$ here is small, it means the restriction changes the likelihood a great deal, and the restriction "matters"

# Evaluating Fit: Likelihood Ratio Test



- Under the null hypothesis that the restriction is true ($\theta_R = \theta_U = \theta$), the likelihoods should be "close" and

- $-2\ln(\lambda) \sim \chi^2_K$ where K is the number of restrictions imposed.

- Note: This can only be used for simple hypotheses, and requires that we use the same distribution in the likelihood under the null and alternative hypotheses (can't have a likelihood be normal under the null and t distributed under the alternative).

# Evaluating Fit: Wald Test

- MLEs are asymptotically normal, with variance equal to the Cramer Rao Lower Bound.
- We can construct a Chi-Squared Statistic by summing the standardized product of normals.
- So, if we applied a parameter restriction that was true, we could take the difference from our estimates, standardize them, square them, and add them up, and we should have a Chi-Squared statistic...

# Evaluating Fit: Wald Test

- In matrix form, that's just

- $W = \left(c\left(\hat{\theta}\right) - q\right)' Var\left(c\left(\hat{\theta}\right) - q\right)^{-1}\left(c\left(\hat{\theta}\right) - q\right)$

- Under the null hypothesis that $c(\theta) = q$, this has a chi-squared distribution with degrees of freedom equal to the number of restrictions.

# Evaluating Fit: Lagrange Multiplier Test



- Now consider maximizing the likelihood subject to some constraint.
- Constrained maximization = Lagrange multipliers.
- If we did this, we could calculate the parameters, the log likelihood, and the multipliers themselves.

# Evaluating Fit: Lagrange Multiplier Test



- If the restrictions are valid, they will not be binding at the optimum, and their impact on the log likelihood will be zero => the Lagrange multiplier should be zero.

- It turns out that the value of the multiplier here also is asymptotically normal, and we know the variance as well => we can use this information to construct a chi-squared test!

# Evaluating Fit: Lagrange Multiplier Test



- $\ln\left(L(\theta_R)\right) = \ln\left(L(\theta)\right) + \lambda(c(\theta) - q)$
- $\frac{\partial \ln\left(L(\theta_R)\right)}{\partial \theta} = \frac{\partial \ln\left(L(\theta)\right)}{\partial \theta} + \left[\frac{\partial c(\theta)}{\partial \theta'}\right]' \lambda = 0$
- At the restricted maximum,
- $\frac{\partial \ln\left(L(\theta)\right)}{\partial \theta} = -\left[\frac{\partial c(\theta)}{\partial \theta'}\right]' \lambda$
- If $\lambda = 0$, $\frac{\partial \ln\left(L(\theta)\right)}{\partial \theta} = 0$

# Evaluating Fit: Lagrange Multiplier Test



- $\frac{\partial \ln(L(\theta))}{\partial \theta}$ has variance equal to the information matrix $I(\theta)$(the variance of the first derivative of the log likelihood is equal to the negative of the second derivative)

- $\frac{\partial \ln(L(\theta))}{\partial \theta}' (I(\theta)^{-1}) \frac{\partial \ln(L(\theta))}{\partial \theta} \sim \chi_K^2$

- Where K is the number of restrictions

# Summary

- LR test: uses objective function changes (as a ratio)
- Wald test: uses coefficients
- Lagrange multiplier test: uses the Lagrange multiplier from a constrained optimization

You want to have good intuition about what these tests look at and why they makes sense.

- Asymptotically, these three tests are equivalent! But in finite samples they will give you different levels of confidence…. This is not a shortcoming of the tests, but is often misunderstood in practice.

# Evaluating Fit: Pseudo $R^2$

- Imagine if we compare the maximized likelihood to what we would see in a model with no explanatory variables (only a constant term). Call this $L_0$
- The likelihood for the unrestricted model with our covariates and coefficients would be higher… so
- $1 - \frac{\ln(L)}{\ln(L_0)}$ would be less than one
- L is between 0 and 1 (it is a probability) => $\ln(L)$ is less than zero => if the restriction is very unlikely, $\ln(L_0)$ is very negative & $\ln(L)$ much less so => this ratio is closer to one.

# Evaluating Fit: Pseudo $R^2$

- $Pseudo\ R^2 = 1 - \frac{\ln(L)}{\ln(L_0)}$

- It does not have the same proportion of variance interpretation either, but it is bounded, and a higher number is better (though it doesn't penalize extra model degrees of freedom).

- Has more to do with an improvement over no model than how much of the dependent variable we can characterize with our regressors…

# Issues with Pseudo $R^2$

- In the linear model,
- $\ln(L) = -\frac{n}{2}[1 + \ln(2\pi) + \ln\left(\frac{e'e}{n}\right)]$

- The Pseudo $R^2$ here can be written as a function of the model $R^2$:

- $-\frac{\ln(1-R^2)}{1+\ln(2\pi)+\ln(s_y^2)}$, which is not the real $R^2$ … more than that, the Pseudo $R^2$ is unit dependent
- with $s_y^2$, it can be less than zero.
- With count or categorical dependent variables, this doesn't make much sense as a measure, but people use it as if it did not have these shortcomings.