Thoughts on Heterogeneity in Econometric Models

Presidential Address
Midwest Economics Association
March 19, 2011

Jeffrey M. Wooldridge
Michigan State University

1. Introduction

- Much of current econometric methodology is aimed at allowing lots of individual heterogeneity (cross section and panel data).
- Even if we are not explicitly modeling heterogeneity, the modern way we interpret standard estimators is colored by the belief that partial effects (treatment effects) can vary widely across individuals – in unobserved ways. This belief is the primary motivation for LATE, marginal treatment effect, and so on.
- Sometimes we rely too much on unobserved heterogeneity for explaining certain results.

- Linear and nonlinear panel data models often incorporate heterogeneity. Naive approaches can result in our attaching too much importance to unobserved heterogeneity.
- Models for cluster samples, including “heirarchical linear models” (HLMs), place a premium on individual and group heterogeneity but often to the detriment of the econometric analysis. (Confusion about allowing lots of heterogeneity versus heterogeneity that is correlated with observables.)

Questions

1. Can we distinguish heterogeneity from other model features (identification)?

2. When does explicitly recognizing heterogeneity matter?

3. Do stories about heterogeneity mask other econometric problems?

4. When do good intentions regarding heterogeneity lead to poor estimators or inference?

5. How come we abandon heterogeneity in some situations?

2. Quantities of Interest in Models with Heterogeneity

- Lots of empirical work focuses on estimating "partial" or "marginal" effects.
- A general conditional mean function with observed covariates, $\mathbf{x}_i$ and unobserved heterogeneity, $\mathbf{c}_i$:

$$E(y_i|\mathbf{x}_i, \mathbf{c}_i) = m(\mathbf{x}_i, \mathbf{c}_i)$$

So the mean function is $m(\mathbf{x}, \mathbf{c})$.

- If x_j is continuous, its partial (or marginal) effect is

$$PE_j(\mathbf{x}, \mathbf{c}) = \frac{\partial m(\mathbf{x}, \mathbf{c})}{\partial x_j}$$

• What should we do about the argument $\mathbf{c}$? If we know $D(\mathbf{c}_i)$, or some features of it, we can plug in interesting values, such as the mean, $\boldsymbol{\mu}_{\mathbf{c}} = E(\mathbf{c}_i)$:

$$PEA_j(\mathbf{x}) = \frac{\partial m(\mathbf{x}, \boldsymbol{\mu}_{\mathbf{c}})}{\partial x_j}$$

• The "partial effect at the average" (PEA) is is not usually the same quantity that underlies the modern treatment effect literature. Plus, if $\mathbf{c}_i$ is continuous (or generally takes on lots of values) then $PEA_j(\mathbf{x})$ applies to only a small part of the population.

• Of course, if we can identify $D(\mathbf{c}_i)$ along with $m(\mathbf{x}, \mathbf{c})$, then we can replace $\mathbf{c}$ with lots of interesting values, such as certain quantiles or a given number of standard deviations from the mean.

• A different way to handle heterogeneity is to average the partial effects across $D(\mathbf{c}_i)$, giving the "average partial effect" (APE):

$$APE_j(\mathbf{x}) = E_{\mathbf{c}_i}\left[\frac{\partial m(\mathbf{x}, \mathbf{c}_i)}{\partial x_j} \right]$$

• If x_j is binary and we look at the difference rather than the derivative, we get the average treatment effect (ATE).

• Idea of an APE is closely related to Blundell and Powell's (2004, REStud) "average structural function" (ASF):

$$ASF(\mathbf{x}) = E_{\mathbf{c}_i}[m(\mathbf{x}, \mathbf{c}_i)]$$

• Definitions of PEA and APE are the same whether or not $\mathbf{c}_i$ is correlated with $\mathbf{x}_i$. Of course, whether we can estimate APEs, and how, depends on what we assume about the relationship between $\mathbf{c}_i$ and $\mathbf{x}_i$.

3. Can we Distinguish Heterogeneity from Other Model Features?

Cross Section, Linear Model

- Consider a random coefficient model for random draws from a cross section:

$$y_i = a_i + \mathbf{x}_i \mathbf{b}_i$$

$$\alpha = E(a_i), \boldsymbol{\beta} = E(\mathbf{b}_i)$$

For example, "ability," b_{i1}, interacts with schooling, x_{i1}.

- In this case, the PEAs and APEs are the same: the β_j.

- Letting $a_i = \alpha + e_i$ and $\mathbf{b}_i = \boldsymbol{\beta} + \mathbf{d}_i$, we can write

$$\begin{aligned} y_i &= \alpha + \mathbf{x}_i\boldsymbol{\beta} + e_i + \mathbf{x}_i\mathbf{d}_i \\ &\equiv \alpha + \mathbf{x}_i\boldsymbol{\beta} + u_i \end{aligned}$$

- If we assume $(a_i, \mathbf{b}_i)$ is independent of $\mathbf{x}_i$ then

$$\begin{aligned} E(y_i|\mathbf{x}_i) &= \alpha + \mathbf{x}_i\boldsymbol{\beta} \\ Var(y_i|\mathbf{x}_i) &= \sigma_e^2 + 2\mathbf{x}_i\boldsymbol{\sigma}_{\mathbf{v}e} + \mathbf{x}_i'\boldsymbol{\Sigma}_{\mathbf{v}}\mathbf{x}_i \end{aligned}$$

• Can we really feel comfortable asserting that heteroskedasticity in $Var(y_i|\mathbf{x}_i)$ is due only to heterogeneity in the slopes? It could be that $\mathbf{b}_i = \boldsymbol{\beta}$ and $Var(a_i|\mathbf{x}_i)$ is heteroskedastic.

• This lack of identification is of little concern here because for estimating partial effects, the reason $Var(y_i|\mathbf{x}_i)$ is heteroskedastic is irrelevant: the APEs (= PEAs) are identified from $E(y_i|\mathbf{x}_i) = \alpha + \mathbf{x}_i\boldsymbol{\beta}$ (and we can use OLS or weighted least squares to estimate them).

Cross Section, **Nonlinear Model**

- Let y_i be a binary response, and consider two underlying models for y_i.

Model 1:

$$y_i = 1[a_i + \mathbf{x}_i\mathbf{b}_i > 0]$$
$$(a_i, \mathbf{b}_i) \text{ independent of } \mathbf{x}_i$$
$$(a_i, \mathbf{b}_i) \sim \text{Multivariate Normal}$$

- This model has lots of unobserved heterogeneity: all slopes can vary by i.

Model 2:

$$y_i = 1[a_i + \mathbf{x}_i\boldsymbol{\beta} > 0]$$
$$D(a_i|\mathbf{x}_i) = Normal(\alpha, \gamma_0 + \mathbf{x}_i\boldsymbol{\gamma}_1 + \mathbf{x}_i'\boldsymbol{\Gamma}_2\mathbf{x}_i)$$
$$(a_i, \mathbf{b}_i) \sim \text{Multivariate Normal}$$

- This model has very little unobserved heterogeneity: a_i is the only observable.

• Models 1 and 2 are observationally equivalent because they both lead to "heteroskedastic probits" for $P(y_i = 1|\mathbf{x}_i)$:

$$P(y_i = 1|\mathbf{x}_i) = \Phi\left[\frac{\alpha + \mathbf{x}_i\boldsymbol{\beta}}{\sqrt{\gamma_0 + \mathbf{x}_i\boldsymbol{\gamma}_1 + \mathbf{x}_i'\boldsymbol{\Gamma}_2\mathbf{x}_i}}\right]$$

(Need a normalization, usually $\gamma_0 = 1$.)

• Unlike in the linear model, the models have very different implications for computing APEs. Model 1 APEs are derivatives of

$$\Phi\left[\frac{\alpha + \mathbf{x}\boldsymbol{\beta}}{\sqrt{1 + \mathbf{x}\boldsymbol{\gamma}_1 + \mathbf{x}'\boldsymbol{\Gamma}_2\mathbf{x}}}\right]$$

whereas in Model 2 they are derivatives of

$$E_{\mathbf{x}_i}\left\{\Phi\left[\frac{\alpha + \mathbf{x}\boldsymbol{\beta}}{\sqrt{1 + \mathbf{x}_i\boldsymbol{\gamma}_1 + \mathbf{x}_i'\boldsymbol{\Gamma}_2\mathbf{x}_i}}\right]\right\}$$

• For Model 1, the APE of x_j is complicated and need not have the same sign as β_j. For Model 2, the APE of x_j has the same sign as β_j, and the relative APEs for continuous x_j and x_h is β_j/β_h.

• For Model 2, the ASF can be consistently estimated as

$$\widehat{ASF}(\mathbf{x}) = N^{-1} \sum_{i=1}^{N} \Phi\left[\frac{\hat{\alpha} + \mathbf{x}\hat{\boldsymbol{\beta}}}{\sqrt{1 + \mathbf{x}_i\hat{\boldsymbol{\gamma}}_1 + \mathbf{x}_i'\hat{\boldsymbol{\Gamma}}_2\mathbf{x}_i}} \right].$$

• Conclusion: Unlike with the linear model, here we cannot tell how to compute the APEs unless we take a stand on the underlying model. Lack of identification.

• The claim about APEs for Model 1 is a special case of a general situation. If we start with

$$E(y_i|\mathbf{x}_i, \mathbf{c}_i) = m(\mathbf{x}_i, \mathbf{c}_i)$$

and *assume* $\mathbf{c}_i$ is independent of $\mathbf{x}_i$, the APEs are obtained simply from

$$r(\mathbf{x}) \equiv E(y_i|\mathbf{x}_i = \mathbf{x}),$$

which is nonparametrically identified.

• In other words, if APEs are of interest, we can just ignore the heterogeneity entirely and search for flexible models for $E(y_i|\mathbf{x}_i)$ [or $D(y_i|\mathbf{x}_i)$ more generally].

- A very useful extension: Suppose

$$E(y_i|\mathbf{x}_i, \mathbf{c}_i, \mathbf{w}_i) = E(y_i|\mathbf{x}_i, \mathbf{c}_i) = m(\mathbf{x}_i, \mathbf{c}_i) \text{ (redundancy of } \mathbf{w}_i)$$
$$D(\mathbf{c}_i|\mathbf{x}_i, \mathbf{w}_i) = D(\mathbf{c}_i|\mathbf{w}_i) \text{ ($\mathbf{w}_i$ contains good ``proxies'' for } \mathbf{c}_i)$$

Define

$$r(\mathbf{x}, \mathbf{w}) = E(y_i|\mathbf{x}_i = \mathbf{x}, \mathbf{w}_i = \mathbf{w}).$$

It can be shown that

$$ASF(\mathbf{x}) = E_{\mathbf{w}_i}[r(\mathbf{x}, \mathbf{w}_i)].$$

- $r(\mathbf{x}, \mathbf{w})$ is nonparametrically identified if $\mathbf{w}_i$ has enough separate variable from $\mathbf{x}_i$. And we can always use flexible parametric approximations.
- We can consistently estimate the ASF as

$$\widehat{ASF}(\mathbf{x}) = N^{-1} \sum_{i=1}^{N} \hat{r}(\mathbf{x}, \mathbf{w}_i)$$

- Has lots of applications to true proxy situations (for example, *IQ* proxies for "ability") but also for correlated random effects panel data models and control function methods to handle endogenous explanatory variables.

- Economists like to think we can be more structural, but can we in a convincing way? The previous probit example shows it is heroic to think that we can generally distinguish between models with "a little heterogeneity" and "a lot of heterogeneity."
- But if we maintain assumptions of the form

$$D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i) \text{ or } D(\mathbf{c}_i|\mathbf{x}_i, \mathbf{w}_i) = D(\mathbf{c}_i|\mathbf{w}_i)$$

then we can at least identify APEs even though we cannot necessarily identify $D(\mathbf{c}_i|\mathbf{x}_i)$ or $D(\mathbf{c}_i)$.

- See Wooldridge (2010, MIT Press) for an even simpler probit example that shows APEs are identified even the the "structural" parameters and heterogeneity distribution are not.
- Introducing endogeneity in $\mathbf{x}_i$ does not help identify structural parameters. But APEs can still be identified using instruments and control function methods [Blundell and Powell (2004), Wooldridge (2005, Rothenberg Festschrift)].

Linear Panel Data Model

• With panel data, separating heterogeneity from other sources of randomness is more promising but still has its pitfalls.

• Consider the standard unobserved effects model with a single, additive source of heterogeneity:

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \\ &\equiv \mathbf{x}_{it}\boldsymbol{\beta} + v_{it} \end{aligned}$$

where $v_{it} \equiv c_i + u_{it}$.is the composite error.

- Typically we identify σ_c^2 and σ_u^2 from the expressions

$$\sigma_c^2 = Cov(v_{it}, v_{ir}),\ t \neq r$$
$$\sigma_u^2 = \sigma_v^2 - \sigma_c^2$$

- We often draw conclusions about the importance of heterogeneity from

$$\rho = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_u^2}$$

which is routinely estimated after random effects or fixed effects estimation of β.

• But the usual estimate of ρ is valid only when $\{u_{it} : t = 1, \ldots, T\}$ is serially uncorrelated and homoskedastic. With positive serial correlation, ρ is almost certainly overestimated.

• In the extreme case where $Var(\mathbf{u}_i)$ is a general $T \times T$ covariance matrix, identification of σ_c^2 is lost.

- Ironically, will often see "cluster robust" standard errors reported with RE and FE estimation – to guard against heteroskedasticity/serial correlation in $\{u_{it}\}$ – while at the same time relying on standard estimates of ρ to determine the importance of c_i.
- For estimating $\boldsymbol{\beta}$ and conducting inference, it does not matter whether we can identify features of $D(c_i)$.

```
. xtreg lpassen lfare concen ldist ldistsq y98 y99 y00, re

Random-effects GLS regression                   Number of obs      =      4596
Group variable: id                              Number of groups   =      1149

------------------------------------------------------------------------------
     lpassen |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfare |  -1.108796    .022065   -50.25   0.000    -1.152043   -1.065549
      concen |   .1119717   .0387787     2.89   0.004     .0359668    .1879766

-------------+----------------------------------------------------------------
         rho |  .97188941   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. xtreg lpassen lfare concen ldist ldistsq y98 y99 y00, re cluster(id)

                                (Std. Err. adjusted for 1149 clusters in id
------------------------------------------------------------------------------
             |               Robust
     lpassen |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfare |  -1.108796   .1034778   -10.72   0.000    -1.311609   -.9059832
      concen |   .1119717   .0850196     1.32   0.188    -.0546636     .278607
-------------+----------------------------------------------------------------
         rho |  .97188941   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

- What about panel data models with lots of heterogeneity?

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}\mathbf{b}_i + c_i + u_{it} \\ &\equiv \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \mathbf{x}_{it}\mathbf{d}_i + u_{it} \end{aligned}$$

under the assumptions

$$\begin{aligned} D(c_i, \mathbf{b}_i | \mathbf{x}_i) &= D(\mathbf{b}_i) = \boldsymbol{\Lambda}_{\mathbf{b}} \\ E(\mathbf{u}_i | \mathbf{x}_i, \mathbf{b}_i, c_i) &= \mathbf{0} \\ Var(\mathbf{u}_i | \mathbf{x}_i, \mathbf{b}_i, c_i) &= \sigma_u^2 \mathbf{I}_T \end{aligned}$$

- Estimation can be hard even under these strong assumptions; simulation and Bayesian methods often relied on.
- Empirically, the conclusion is usually that $Var(\mathbf{b}_i) \neq \mathbf{0}$, so there is "lots of heterogeneity" – even though none or very little of it is observable
- But the entire analysis hinges on the strong assumption that $Var(\mathbf{u}_i|\mathbf{x}_i, \mathbf{b}_i, c_i) = \sigma_u^2 \mathbf{I}_T$ (particularly, no serial correlation in $\{u_{it}\}$).
- By contrast, consistent estimation the APEs $\boldsymbol{\beta} = E(\mathbf{b}_i)$ along with fully robust inference are straightforward.

Nonlinear Panel Data Models

- Correlated random effects (CRE) probit model:

$$P(y_{it} = 1|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)$$
$$D(c_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) = D(c_i|\bar{\mathbf{x}}_i) = Normal(\psi + \bar{\mathbf{x}}_i\boldsymbol{\xi}, \sigma_a^2)$$

where $\bar{\mathbf{x}}_i = T^{-1}\sum_{r=1}^{T}\mathbf{x}_{ir}$.

- These assumptions are enough to estimate the APEs, but not the PEAs or effects at other values of c: $\boldsymbol{\beta}$ is identified only up to scale, and ψ, $\boldsymbol{\xi}$, and σ_a^2 are not separately identified (which means $D(c_i)$ is not identified).

- If we add conditional independence,

$$D(y_{i1}, \ldots, y_{iT} | \mathbf{x}_i, c_i) = \prod_{t=1}^{T} D(y_{it} | \mathbf{x}_i, c_i),$$

then we can identify all parameters and so also the unconditional moments μ_c, σ_c^2 [and even all of $D(c_i)$].

- How badly do we want to identify the heterogeneity distribution and at what cost? The joint MLE ("random effects probit") is not robust for estimating the APEs; the pooled MLE is.

• What if we want to allow more heterogeneity?

$$P(y_{it} = 1|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{c}_i) = \Phi(a_i + \mathbf{x}_{it}\mathbf{b}_i)$$

and not impose conditional independence (or other specific dependence restrictions)?

• Short of trying to estimate $(a_i, \mathbf{b}_i)$ for each i (with only T observations each), it suffices to restrict $D(\mathbf{c}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ in some way, such as $D(\mathbf{c}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) = D(\mathbf{c}_i|\bar{\mathbf{x}}_i)$ or, more generally,

$$D(\mathbf{c}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) = D(\mathbf{c}_i|\mathbf{w}_i)$$

for suitable "sufficient statistics" $\mathbf{w}_i$.

• Have to impose some restrictions on $\mathbf{w}_i$, but the APEs are nonparametrically identified from

$$r(\mathbf{x}_{it}, \mathbf{w}_i) \equiv P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{w}_i)$$

without restricting $D(\mathbf{c}_i|\mathbf{w}_i)$:

$$ASF(\mathbf{x}_t) = E_{\mathbf{w}_i}[r(\mathbf{x}_{it}, \mathbf{w}_i)].$$

[Altonji and Matzkin, (2005, Econometrica), Wooldridge (REStat, 2005).]

• The focus on APEs can be liberating and allows *a lot* of underlying heterogeneity. Generally, for any response variable y_{it}, the ASF is identified under

$$E(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},\mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it},\mathbf{c}_i) \text{ (strict exogeneity of } \{\mathbf{x}_{it}\})$$

and

$$D(\mathbf{c}_i|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}) = D(\mathbf{c}_i|\mathbf{w}_i) \text{ } (\mathbf{w}_i \text{ a "sufficient statistic"})$$

with suitable restrictions on $\mathbf{w}_i$.

- As a radical suggestion, just use flexible flexible functions of $\mathbf{x}_{it}$ and $\mathbf{w}_i$ in standard parametric models, and then average out $\mathbf{w}_i$.
- For example,

$$\hat{r}(\mathbf{x}_{it}, \mathbf{w}_i) = \Phi[\hat{\psi}_t + \mathbf{x}_{it}\hat{\boldsymbol{\beta}} + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}} + (\mathbf{x}_{it} \otimes \bar{\mathbf{x}}_i)\hat{\boldsymbol{\eta}} + \mathbf{g}_i\hat{\boldsymbol{\gamma}} + (\mathbf{x}_{it} \otimes \mathbf{g}_i)\hat{\boldsymbol{\delta}}]$$

where the $\mathbf{g}_i$ are, say, average trends in $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$

- The APEs are estimated as derivatives of or discrete changes in

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N}\Phi[\hat{\psi}_t + \mathbf{x}_t\hat{\boldsymbol{\beta}} + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}} + (\mathbf{x}_t \otimes \bar{\mathbf{x}}_i)\hat{\boldsymbol{\eta}} + \mathbf{g}_i\hat{\boldsymbol{\gamma}} + (\mathbf{x}_t \otimes \mathbf{g}_i)\hat{\boldsymbol{\delta}}]$$

- Note: Similar strategies can be justified for linear panel data models (that is, include polynomials and interactions of time averages and trends).
- Some complain about CRE approaches as being restrictive, but “fixed effects” approaches *do not* assume less, and they deliver less, too (no magnitudes of effects).

When Can Appealing to Heterogeneity Lead Us to Harmful Conclusions?

Attributing Findings to Heterogeneity Rather than Poor Instruments

- The interpretation of LATE – as measuring the effect of a policy on a particular, unidentifiable subset of the population – has proven useful, but it can be abused.
- For example, why are IV estimates of the return to schooling often larger than OLS estimates? "Ability" bias suggests it should be otherwise. The LATE interepretation is heterogeneous returns to schooling: the return for "compliers" is larger than the average return. But might the instrument just be poor (even just "slightly" endogenous)?

Ignoring Other Econometric Problems when Conducting Inference

- Less serious but serious enough: standard inference in models with even lots of heterogeneity can be very misleading. Only fairly recently has “cluster-robust” inference become popular for linear models estimated by fixed effects and random effects. Traditionally it was taken as a given that all other sources of randomness were i.i.d. shocks. (Airfare example gives a counterexample.)
- Robust inference is not as commonly used in nonlinear panel data models (for example, random effects versions of probit, Tobit, and count models). But it should be.

```
. xtpqml passen lfare concen  y98 y99 y00, fe

Conditional fixed-effects Poisson regression    Number of obs      =      4596
Group variable: id                              Number of groups   =      1149

------------------------------------------------------------------------------
      passen |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfare |  -.8658171   .0069057  -125.38   0.000     -.879352   -.8522822
      concen |  -.1289482   .0123807   -10.42   0.000    -.1532138   -.1046825
------------------------------------------------------------------------------


Calculating Robust Standard Errors...
------------------------------------------------------------------------------
      passen |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
passen       |
       lfare |  -.8658171    .036619   -23.64   0.000     -.937589   -.7940452
      concen |  -.1289482   .0544245    -2.37   0.018    -.2356182   -.0222781
```

```
. xtpoisson passen lfare concen ldist ldistsq y98 y99 y00, re

------------------------------------------------------------------------------
      passen |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfare |  -.8621147     .00688  -125.31   0.000    -.8755993   -.8486301
      concen |  -.1353661   .0123453   -10.96   0.000    -.1595624   -.1111697
------------------------------------------------------------------------------


. xtpoisson passen lfare concen ldist ldistsq y98 y99 y00, re vce(boot,r(200

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
      passen |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfare |  -.8621147   .0355057   -24.28   0.000    -.9317046   -.7925249
      concen |  -.1353661   .0565836    -2.39   0.017    -.2462678   -.0244643
------------------------------------------------------------------------------
```

- Ignoring key features, namely heteroskedasticity and serial correlation in the idiosyncratic errors, is rampant in the “hierarchical linear models” (HLMs) or “mixed models” literature. Defaults are always to assume the idiosyncratic errors are independent over time while allowing for group-level, as well as individual-level, hetereogeneity.
- Applications of HLMs are where RE was a decade ago: attribute serial correlation entirely to unobserved heterogeneity.
- Remember, having the variance-covariance matrix misspecified does not cause inconsistency in feasible GLS. But assuming the V-C matrix is correctly specified can lead to very misleading inference.

```
. xtreg math4 lavgrexp lunch lenrol y95 y96 y97 y98, re

Random-effects GLS regression                   Number of obs      =      7150
Group variable: schid                           Number of groups   =      1683

------------+----------------------------------------------------------------
    lavgrexp |   7.838068   1.454286     5.39   0.000     4.987721    10.68842
------------------------------------------------------------------------------

. xtreg math4 lavgrexp lunch lenrol y95 y96 y97 y98, re cluster(schid)

                               (Std. Err. adjusted for 1683 clusters in schid
------------+----------------------------------------------------------------
    lavgrexp |   7.838068   1.578525     4.97   0.000     4.744217    10.93192

------------------------------------------------------------------------------

. xtreg math4 lavgrexp lunch lenrol y95 y96 y97 y98, re cluster(distid)

                               (Std. Err. adjusted for 467 clusters in distid
------------+----------------------------------------------------------------
    lavgrexp |   7.838068   2.157833     3.63   0.000     3.608793    12.06734

------------------------------------------------------------------------------
```

```
. xtmixed math4 lavgrexp lunch lenrol y95 y96 y97 y98 || distid: || schid:

Mixed-effects REML regression                   Number of obs      =      7150

------------------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+-------------------------------------------
         distid |      467          3       15.3        623
          schid |     1683          3        4.2          5
------------------------------------------------------------------

------------+-----------------------------------------------------------------
    lavgrexp |   5.674265   1.577373     3.60   0.000     2.582671    8.765859
------------------------------------------------------------------------------
```

Pitfalls of Structural Dynamic Models with Heterogeneity

- When the nature of the problem requires us to incorporate heterogeneity, we have no choice. Leading example is determining the amount of "state dependence" in

$$D(y_{it}|y_{i,t-1}, y_{i,t-2}, \ldots, y_{i0}, \mathbf{c}_i),$$

or add covariates. Some nontrivial assumptions are required, but fairly convincing analyses are available.

- But what if we are mainly interested in the effects of covariates – particularly policy interventions – on average outcomes?

• Example: Value-added models for estimating the effectiveness of teachers, schools, or programs. The "educational production function" approach is a structural approach with heterogeneity. With achievement A_{it} and educational inputs $\mathbf{E}_{it}$, a typical starting point is a distributed lag on school inputs, with student heterogeneity, c_i:

$$A_{it} = \alpha_t + \mathbf{E}_{it}\boldsymbol{\beta}_0 + \mathbf{E}_{i,t-1}\boldsymbol{\beta}_1 + \ldots + \mathbf{E}_{i0}\boldsymbol{\beta}_t + c_i + u_{it}$$

• If we impose a geometric distributed lag restriction on the $\boldsymbol{\beta}_j$ then $\boldsymbol{\beta}_s = \lambda^s \boldsymbol{\beta}_0$ for some $0 \leq \lambda \leq 1$. Leads to

$$\begin{aligned} A_{it} &= \alpha_t - \lambda\alpha_{t-1} + \lambda A_{i,t-1} + \mathbf{E}_{it}\boldsymbol{\beta}_0 + (1-\lambda)c_i + u_{it} - \lambda u_{i,t-1} \\ &\equiv \tau_t + \lambda A_{i,t-1} + \mathbf{E}_{it}\boldsymbol{\beta}_0 + a_i + e_{it} \end{aligned}$$

• If we make some extra assumptions – $\{e_{it} = u_{it} - \lambda u_{i,t-1}\}$ is serially uncorrelated, $\{\mathbf{E}_{it}\}$ is strictly exogenous with respect to $\{u_{it}\}$ – then the methods of Arellano and Bond (1991, REStud) can be used to estimate λ and $\boldsymbol{\beta}_0$.

• The AB method is based on IV estimation of

$$\Delta A_{it} = \omega_t + \lambda \Delta A_{i,t-1} + \Delta \mathbf{E}_{it} \boldsymbol{\beta}_0 + \Delta e_{it}$$

where $\Delta \mathbf{E}_{it}$ acts as its own instruments.

• A key worry is the nature of assignment of $\mathbf{E}_{it}$ (for example, teachers or class size). Is assignment dependent on c_i or $A_{i,t-1}$ or past shocks?

• Guarino, Reckase, and Wooldridge (2011, Working Paper): In simulations, the AB approach is very sensitive to nonrandom assignment mechanisms (and serial correlation). Dynamic OLS – ignoring the nature of $\{e_{it}\}$ and the heterogeneity c_i – is much better behaved for estimating $\boldsymbol{\beta}_0$ even though it is technically inconsistent for λ and $\boldsymbol{\beta}_0$.

• DOLS is only slightly worse than random effects when random effects is the "right" thing to do (random assignment, no extra serial correlation).

How Come We Sometimes Ignore Heterogeneity?

- Suppose we are interested in the conditional median rather than the conditional mean:

$$Med(y_i|\mathbf{x}_i, a_i, \mathbf{b}_i) = a_i + \mathbf{x}_i\mathbf{b}_i.$$

How should we define the population paramters of interest?

$\beta_j = E(b_{ij})$? $\beta_j = Med(\beta_{ij})$?

- Even if $\mathbf{b}_i$ is independent of $\mathbf{x}_i$, neither the APEs nor median partial effects are generally identified. Need multivariate symmetry. Things are even harder with panel data.

• What about endogenous treatment effects? Asymmetry in focusing on heterogeneity in the response equation but assuming it away in the treatment equation:

$$y_i = a_i w_i + \mathbf{x}_i \mathbf{b}_i + u_i$$
$$w_i = 1[f_i + \mathbf{z}_i \mathbf{g}_i > 0]$$

where the treatment w_i can be correlated with $(a_i, \mathbf{b}_i)$ and $\mathbf{z}_i$ includes exogenous variables $\mathbf{x}_i$ in the response equation and extra instruments.

- Card (2001, Econometrica) proposes an economic model that determines wage and level of schooling. The "reduced form" for schooling turns out to be a random coefficient model. Can we justify arguing for heterogeneity in a wage equation but not a schooling equation?
- Evidently, the problem is theoretical: generally, the presence of $\mathbf{g}_i$ violates the monoticity assumption in the LATE setting and (necessarily) the index structure of Heckman and Vytlacil (2005, Econometrica).

- See also Florens, Heckman, Meghir, and Vytlacil (2008, Econometrica): If the cost function for schooling is sufficiently heterogenous, monotonicity fails.
- Of course, if we make full parametric assumptions on $D(a_i, \mathbf{b}_i, f_i, \mathbf{g}_i|\mathbf{z}_i)$ then $\alpha = E(a_i)$ (the ATE) can be identified and ATEs for subpopulations.

- Why not just admit there are limits on how much heterogeneity can be allowed in nonparametric and semi-parametric approaches to identifying average treatment effects and use flexible parametric models? Are distributional assumptions so much worse than assuming away heterogeneity in selection equations?
- The kinds of monotonicity assumptions used in the nonparametric work are not strictly weaker than parametric assumptions. Why should we necessarily place more trust in monotonicity?
- As a corollary, if nonparametric and parametric methods give different answers we cannot tell which one is “better.”

Summary

- Introducing heterogeneity independent of covariates can be a useful functional form device, but some important quantities of interest (APEs) do not depend on whether or how heterogeneity is in the model.
- The same is true under a conditional independence assumption: we can often identify average partial effects even though we know nothing about the shape of the heterogeneity distribution.
- Trying to identify heterogeneity distributions from cross section data seems heroic. Identification fails in leading cases.

- Identifying heterogeneity distributions with panel data is more promising but relies on restrictions on the dynamics of the process and (usually) imposes strict exogeneity on a set of covariates.
- Very flexible methods for estimating APEs require much weaker assumptions on dynamics and the heterogeneity distribution. Parameters and the heterogeneity distribution are not always fully identified but APEs are.

- Lately it has become popular to appeal to heterogeneity to explain counterintuitive outcomes, but there may be competing explanations (such as poor instruments).
- There are problems for which modeling heterogeneity is fundamental, such as distinguishing between heterogeneity and state dependence. But in other cases – particularly for analyzing policy interventions – dynamic, structural models can produce notably inferior results.

- Nonparametric approaches to identifying treatment effects rule out even simple forms of heterogeneity in the selection equation. Is this always better than parametrically modeling lots of heterogeneity in the response and selection equations? There still appears to be plenty of scope for flexible parametric analysis.