Clustering, random effects, mixed models, sandwiches

February 22, 2021

Grouped data

Notation: When a variable z_{ig} is indexed by i and g, $z_{\cdot g}$ refers to the vector consisting of all the z_{ig}'s with the given value of g.

$$y_{ig} = X_{ig}\beta + \varepsilon_{ig}\,, \quad i = 1, \ldots, n, \quad g = 1, \ldots, M\,.$$

g indexes groups (states, gender, age brackets, etc.), i indexes observations (people, years, firms, etc.).

GLS: two step

Assume residual correlation within groups, but not between:

$$\mathrm{Var}(\varepsilon_{\cdot g}) = \Omega, \text{ all } g\,, \qquad \mathrm{Cov}(\varepsilon_{\cdot g}, \varepsilon_{\cdot h}) = 0\,.$$

Estimate by OLS on all data, use the estimated β to form $\hat{\varepsilon} = y - X\beta$, estimate Ω as

$$\hat{\Omega} = \sum_{g=1}^{M} \hat{\varepsilon}_{\cdot g}\hat{\varepsilon}'_{\cdot g}\,.$$

Then the full $nM \times nM\ E[\varepsilon\varepsilon']$ matrix is block diagonal, with copies of Ω down the diagonal. The estimated Ω converges in probability to the true value as $M \to \infty$ by the law of large numbers, so we can use it for feasible GLS estimation.

Because Ω is symmetric, it contains $(n^2 + n)/2$ coefficients. If $(n^2 + n)/2 \geq M$, $\hat{\Omega}$ will be singular (hence not usable for GLS).

GLS: Likelihood approach

Assume normality, write down the likelihood for the sample, base inference on it:

$$|\Omega|^{-M/2}\,(2\pi)^{-Mn/2}e^{-\frac{1}{2}\sum_{g=1}^{M}(y_{\cdot g}-X_{\cdot g}\beta)'\Omega^{-1}(y_{\cdot g}-X_{\cdot g}\beta)}\,.$$

Here $(n^2+n)/2 \geq M$ implies the likelihood is unbounded above.

Parametric form for Ω

- If group size n is fairly large GLS my require using a parametric form for Ω.

- For example, if observations within each group form a time series, we could assume a first or second order autoregressive model for the residuals, making Ω a function of just two or three parameters.

- Or, one can model the dependence of Ω on X with a **mixed model**, which we will discuss below.

Clustered covariance matrix for β.

Assume
$$E[X'\varepsilon\varepsilon'X] = E\Big[\sum_g X'_{\cdot g}\varepsilon_{\cdot g}\varepsilon'_{\cdot g}X_{\cdot g}\Big] \ .$$

That is, no correlation of $X'_{\cdot g}\varepsilon_{\cdot g}$ with $X'_{\cdot h}\varepsilon_{\cdot h}$ for $g \neq h$, but arbitrary covariance, constant across g, when $g = h$.

Then apply the usual "robust" standard error form:

$$T\,\mathrm{Var}(\hat{\beta}_{OLS}) \doteq TE\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right] \ ,$$

which is consistently estimated under stationarity assumptions by

$$T(X'X)^{-1}\left(\sum_{g=1}^{M} X'_{\cdot g}\hat{\varepsilon}_{\cdot g}\hat{\varepsilon}'_{\cdot g}X_{\cdot g}\right)(X'X)^{-1} \ .$$

Which: OLS, GLS 2-step, GLS Likelihood, OLS with sandwich?

- A very common set of trade-offs here. Robustness increases inaccuracy.

- Straight OLS with $\sigma^2(X'X)^{-1}$ covariance matrix is the most accurate and efficient if its assumptions are correct.

- If the only deviation from the SNLM assumption is $\Omega \neq \sigma^2 I$, GLS is the most accurate and efficient if its assumptions are correct.

- OLS with sandwich is easiest on the researcher's brain. No need to think about a structure for Ω or to defend the assumption of $E[\varepsilon_{\cdot g}\varepsilon'_{\cdot g}]$ constant across groups.

A fundamental distinction

- GLS and OLS without the sandwich produce covariance matrices for β that condition on the observed X. The sandwich covariance matrix treats X as random and averages across its behavior in repeated samples as we extend the number of groups.

- This means, for example, that if X is drawn from a distribution with infinite variance (as would be possible for firm size or individual income, for example), the sandwich estimator provides meaningless results, whereas if their assumptions are correct, OLS or GLS with their usual associated covariance matrices are unaffected by the infinite variance of X.

Another fundamental distinction

- OLS with sandwich does not require a balanced panel — that is, the number of observations can differ across groups.

- GLS estimating Ω by averaging $\varepsilon_{.g}\varepsilon'_{.g}$ across groups can't work with unbalanced data, though this is not a problem for most parametric models for Ω.

- For example if each group consists of time series data. One can model the residuals as an MA or AR process, making Ω a function of the AR or MA parameters.

Efficiency vs robustness

- OLS plus sandwich will provide, in large enough samples, the same covariance matrix as the SNLM $\sigma^2(X'X)^{-1}$ formula, yet also provides "correct" answers under broader assumptions.

- So isn't using OLS plus sandwich "conservative", or "safer"?

Efficiency vs robustness

- OLS plus sandwich will provide, in large enough samples, the same covariance matrix as the SNLM $\sigma^2(X'X)^{-1}$ formula, yet also provides "correct" answers under broader assumptions.

- So isn't using OLS plus sandwich "conservative", or "safer"?

- Finding an effect to be "insignificant" when it is in fact present can be as bad a mistake as its opposite.

- Finding a big difference between $\sigma^2(X'X)^{-1}$ and the sandwich covariance matrix does suggest that the SNLM model assumptions may be incorrect, but the right response could be to use GLS or a mixed model, rather than be satisfied with the sandwich.

Group-specific shifts

Same grouped data model, with one change:

$$y_{ig} = X_{ig}\beta + \nu_g + \varepsilon_{ig}\,, \quad i = 1, \ldots, n, \quad g = 1, \ldots, M\,.$$

What's new is the ν_g, a "disturbance" that changes all observations within a group by the same amount. I've used a greek letter for it, which makes it seem natural to treat it as part of the error term.

Applying GLS

If we assert that

$$\varepsilon \mid X \sim N(0, \sigma^2 \underset{Mn \times Mn}{I}), \qquad \nu \mid X \sim N(0, \tau^2 \underset{M \times M}{I}),$$

and ε, ν independent, then this is a special case of grouped data with non-scalar covariance matrix. Instead of an unconstrained residual covariance matrix Ω for data within each group, we have the parametric form

$$\Omega = \sigma^2 I + \tau^2 \underset{n \times n}{\mathbf{1}}.$$

We could stop here, simply referring back to the discussion of likelihood-based GLS, but it is worth noting that there is an analogue of weighted least squares available because of the special structure of Ω.

Between and within regressions

As usual, if we can find a matrix W such that $W'\Omega W = I$, then OLS on the transformed data $X^*_{\cdot g} = W'X_{\cdot g}$, $y^*_{\cdot g} = W'y_{\cdot g}$ is equivalent to GLS. With this group-effect Ω matrix, we can choose W to have the symmetric form

$$W = I_M \otimes \left(\sigma^{-1}(I - \tfrac{1}{n}\mathbf{1}) + \frac{\delta}{n}\mathbf{1} \right).$$

It is possible to compute δ from σ^2 and τ^2, but this involves messy algebra. What matters for our purposes is that there is a symmetric inverse square root W of Ω of this form, and that

$$\delta \xrightarrow[\tau^2 \to 0]{} \frac{1}{\sigma}$$

$$\delta \xrightarrow[\tau^2 \to \infty]{} 0 \, .$$

Between and within regressions

Note that with this choice of W, $X^* = W'X_{\cdot g} = \sigma^{-1}\tilde{X}_{\cdot g} + \delta\bar{X}_{\cdot g}$, where $\bar{X}$ is a matrix in which each row is the mean of $X_{\cdot g}$ within group g and $\tilde{X}_{\cdot g}$ is the deviation of $X_{\cdot g}$ from its group mean. Note also that $\tilde{X}'_{\cdot g}\bar{X}_{\cdot g} = 0$, because the columns of $\bar{X}_{\cdot g}$ are constant and the sum of each column of $\tilde{X}_{\cdot g}$ is zero.

Therefore

$$X^{*\prime}X^* = \sigma^{-2}\tilde{X}'\tilde{X} + \delta^2\bar{X}'\bar{X}\,, \qquad X^{*\prime}y^* = \tilde{X}'\tilde{y} + \bar{X}'\bar{y}\,,$$

where the $*$'d vectors and matrices are of full length Mn, consisting of the grouped data stacked up vertically.

Between and within regressions

This lets us write the GLS estimator as a matrix weighted average of the OLS estimator $\hat{\beta}_w$ using $\tilde{X}, \tilde{y}$, called the "**within**" regression, and the OLS estimator $\hat{\beta}_b$ using the group means, called the "**between**" regression:

$$\hat{\beta}_{GLS} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^* = (\sigma^{-2}\tilde{X}'\tilde{X} + \delta^2\bar{X}'\bar{X})^{-1}(\sigma^{-2}\tilde{X}'\tilde{X}\hat{\beta}_w + \delta^2\bar{X}'\bar{X}\hat{\beta}_b).$$

This decomposition is of some use to programmers and to you if you try to look at small data sets with a calculator. But the important insight from it is that the GLS estimator, which is the classic **random effects** estimator, becomes the ordinary OLS estimator in the limit as τ^2 becomes very small (as we might have expected) and becomes purely the "within" regression in the limit as τ^2 gets very big. But this pure within regression is also what is known as the **fixed effects** estimator.

Fixed effects

The fixed effects estimator is what emerges if we assert dogmatically that the variance of the group means $\nu_{\cdot g}$ is infinite — i.e. we put a flat prior on ν_g. It is also, as we verified above, the result of using data on deviations from group means in an OLS estimation. And finally, it is also the result if we estimate by OLS the equation

$$ y_{ig} = \sum_{g=1}^{M} D_{ig} c_g + X_{ig} \beta + \varepsilon_{ig} \, , $$

where D_{ig} is a "group g dummy variable", meaning that it is one for observations in group g and zero otherwise. We are now treating the c_g's (a new name for ν_g) as parameters to be estimated along with β.

Asymptotic properties of $\hat{c}_g$

- If the data set grows by increasing the number of groups, M, while the number of observations within each group remains fixed, the estimates of c_g will not be consistent.

- Under standard assumptions they will be unbiased, so we can write $\hat{c}_g = c_g + \xi_g$, where $E[\xi_g] = 0$. Thus the distribution of estimated "fixed effects" will have higher variance than the distribution of the underlying true c_g's.

- Perhaps surprisingly, the fact that the c_g's under these assumptons are not consistently estimated does not prevent β from being consistently estimated.

- Note that this depends strongly on the model's linearity. In nonlinear regression models the inconsistency of the $\hat{c}_g$'s generally makes other parameter estimates also inconsistent.

Fixed vs. random effects

- Fixed effects requires giving up any attempt to estimate coefficients of variables that are constant within groups. Random effects models can do so, because they exploit the assumption that ν_g and X are uncorrelated.

- Fixed effects gives consistent estimates of β as $M \to \infty$, even if ν_g and $X_{\cdot g}$ are correlated, while random effects does not.

Unfortunate terminology

- The fixed-effects estimator consistently estimates β even if $X_{\cdot g}$ and ν_g are correlated, but wastes information if in fact they are uncorrelated.

- The difference between the two estimators is not that one treats ν_g as random and the other does not, however.

- The difference is in whether $E[\nu_g \mid X_{\cdot g}] = 0$ is assumed. We can treat ν_g as random, yet allow for correlation of ν_g with $X_{\cdot g}$.

Random effects correlated with X

$$y_{ig} = X_{ig}\beta + c_g + \varepsilon_{ig}\,, \qquad E[\varepsilon_{.g} \mid X_{.g}] = 0 \tag{1}$$

$$X_{.g} = \gamma c_g + \tilde{X}_{.g}\,, \qquad E[\tilde{X}_{.g} \mid c_g] = 0 \qquad \text{or} \tag{2}$$

$$c_g = \vec{X}_{.g}\psi + \xi_g\,, \qquad E[\xi_g \mid X_{.g}] = 0\,. \tag{3}$$

Wooldridge calls this the "Chamberlain-Mundlak device". Using (3), we can substitute into (1) to obtain

$$y_{ig} = X_{ig}\beta + \vec{X}_{.g}\psi + \xi_g + \varepsilon_{ig}$$

Random effects correlated with X

$$y_{ig} = X_{ig}\beta + \vec{X}_{.g}\psi + \xi_g + \varepsilon_{ig}$$

In this equation, $\vec{X}_{.g}$ is a single vector of length nK containing all the values of X_{ig} that occur in the group. It is the same vector for every i in the group. This is nk additional parameters. That is still usually smaller than the number of c_g parameters that enter the straight fixed-effects estimator. In some applications it might be reasonable to claim that the correlation of ν_g with the $\vec{X}_{.g}$ vector should be only via the group means of the X's. In that case the $\vec{X}_{.g}$ vector could be replaced by the $\bar{X}_{.g}$ vector, making the number of extra parameters much smaller.

Random effects correlated with X

$$y_{ig} = X_{ig}\beta + \bar{X}_{.g}\psi + \xi_g + \varepsilon_{ig}$$

This is an equation that can be estimated by standard grouped-data GLS. It does allow consistent estimation of $\mathrm{Var}(\nu_g)$, but it does not allow estimation of coefficients of possible $X_{.g}$ variables that are constant within groups, because for such variables the corresponding columns of $X_{.g}$ and $\bar{X}_{.g}$ are identical, and thus collinear.

Mixed models

This term does not have a precise and widely accepted definition, but generally refers to models in which not only the constant, but also coefficients of $X_{\cdot g}$ variables, are allowed to be random and vary with g. The general form, assuming the constant vector is treated as part of the $X_{\cdot g}$ matrix, is

$$y_{ig} = X_{ig}\beta_g + Z_{ig}\gamma + \varepsilon_{ig}$$

$$E[\varepsilon \mid X, Z] = 0$$

$$E[\beta_g \mid X, Z] = \bar{\beta}]$$

24

Mixed models

$$y_{ig} = X_{ig}\beta_g + Z_{ig}\gamma + \varepsilon_{ig}$$

$$E[\varepsilon \mid X, Z] = 0$$

$$E[\beta_g] \mid X, Z] = \bar{\beta}]$$

Some assumption on the joint distribution of β_g and ε_{ig} is needed. A common choice would be to make β_g and ε_{ig} jointly normal and independent of each other, with the full $Mn \times 1$ ε_{ig} vector $N(0, \sigma^2 I)$ and

$$\mathrm{Var}(\beta_g) = I \otimes \Sigma_\beta \,,$$

where Σ_β is an unknown and unrestricted covariance matrix and the "$A \otimes B$" notation refers to a Kronecker product.

Why mixed models?

They give a fully articulated probability model for the data, and thus a likelihood function, while addressing the possibility that $E[\varepsilon_{\cdot g} \varepsilon'_{\cdot g} \mid X_{\cdot g}]$ might depend on $X_{\cdot g}$. This possibility is what clustered standard errors allow for that GLS does not. Clustered standard errors allow any kind of dependence between $X_{\cdot g}$ and $\varepsilon_{\cdot g} \varepsilon'_{\cdot g}$, while mixed models restrict the dependence to linearity.

Mixed models used to be intractably difficult to estimate, but with Gibbs sampling MCMC, they can be handled in a very straightforward way.

It might be a good way for you to test your understanding of Gibbs sampling to see if you can describe a convenient Gibbs sampling scheme for a mixed model.

"Fixed effects" for mixed models?

Mixed models almost always are used treating β_g as random. But there is an analogue to the simple fixed effects approach for these models: just as we split the constant into a bunch of group-dummy variables for fixed effects, we can split up the X matrix into a block diagonal form with $X_{.g}$ blocks down the diagonal and zeros off diagonal, giving each column of this matrix its own free parameter and applying OLS.

The problem with this is that if there are very many columns in X, the fact that OLS with fixed effects allocates too much explanatory power to the fixed effects is multiplied in a mixed model — here it is not only the constant terms, but the coefficients on all the group-specific variables, that have an over-dispersed distribution.