Preliminaries

Fu Zixuan[1]

[1] Last compiled on May 18, 2024

Contents

1 Year 2016-2022 . . . 2

1.1 Regression tables . . . 3

1.1.1 WG, BG, Random effect, Correalted random effect(Mundlak) . . . 3

1.1.2 Inference on $\hat{\beta}$ and $\hat{\theta}_i$ . . . 4

1.2 Mixture model . . . 6

1.2.1 Location mixture . . . 6

1.2.2 Scale mixture . . . 8

1.2.3 Location-scale mixture (independent) . . . 9

1.2.4 Location-scale mixture (general) . . . 9

Chapter 1

Year 2016-2022

1.1 Regression tables

1.1.1 WG, BG, Random effect, Correalted random effect(Mundlak)

Notation

- Dependent variable:

$$\underbrace{y}_{NT\times 1} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \quad \underbrace{y_i}_{T\times 1} = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}$$

- Independent variable:

$$\underbrace{X}_{NT\times K} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} \quad \underbrace{x_i}_{T\times K} = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iT} \end{pmatrix}$$

- Matrix to calcualte the mean:

$$B_T = d_T(d_T' d_T)^{-1} d_T' \quad \text{where} \quad d_T = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

  Thus

$$B_T y_1 = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_1 \end{pmatrix}$$

  we set $B = I_N \otimes B_T$

- Matrix to demean the variable:

$$W_T = I_T - B_T$$

  Thus,

$$W_T y_i = \begin{pmatrix} y_{i1} - \bar{y}_1 \\ \vdots \\ y_{iT} - \bar{y}_1 \end{pmatrix}$$

  we set $W = I_N \otimes W_T$

Estimators

- WG:

$$\hat{\beta}_{WG} = (X'WX)^{-1}X'Wy$$

- BG:

$$\hat{\beta}_{BG} = (X'BX)^{-1}X'By$$

- RE: A linear combination of WG and BG
- CRE (Mundlak): equivalent to WG

> In empirical analysis of data consisting of repeated observations on economic units (time series on a cross section) it is often assumed that the coefficients of the quantitiative variables (slopes) are the same, whereas the coefficients of the qualitative variables (intercepts or effects) vary over units or periods. This is the constant-slope variable- intercept framework. In such an analysis an explicit account should be taken of the statistical dependence that exists between the quantitative variables and the effects. It is shown that when this is done, the random effect approach and the fixed effect approach yield the same estimate for the slopes, the "within" estimate. Any matrix combination of the "within" and "between" estimates is generally biased. When the "within" estimate is subject to a relatively large error a minimum mean square error can be applied, as is generally done in regression analysis. Such an estimator is developed here from a somewhat different point of departure.

1.1.2 Inference on $\hat{\beta}$ and $\hat{\theta}_i$

Inference on β Because

$$\hat{\beta}_{WG} - \beta = (X'WX)^{-1}X'W\epsilon$$

we have

$$\sqrt{N}(\hat{\beta}_{WG} - \beta) \rightarrow^d N(0, A^{-1}CA^{-1})$$

where $A = \mathbb{E}\left[X'WX\right]$ and $C = \mathbb{E}\left[X'W\epsilon\epsilon'W'\right] = \mathbb{E}\left[X'W\Omega W'X\right]$.
We may greatly simplify Ω if we assume that all individual and all observations' ϵ are iid. Then $\Omega = \sigma^2 I_{NT}$ (something that we generally do not assume.)

Inference on θ_i If we apply the WG estimator $\hat{\beta}_{WG} = (X'WX)^{-1}X'Wy$. The estimated fixed effect is

$$\begin{aligned}
\hat{\theta}_i &= \frac{1}{T}\sum_{t=1}^{T} y_{it} - x_{it}\hat{\beta}_{WG} \\
&= \frac{1}{T}\sum y_{it} - x_{it}\beta + \frac{1}{T}\sum x_{it}\beta - x_{it}\hat{\beta}_{WG} \\
&= \theta_i + \frac{1}{T}\sum_t \epsilon_{it} + \left(\frac{1}{T}\sum x_{it}\right)(\beta - \hat{\beta}_{WG})
\end{aligned}$$

It is clear that the second part follows a normal distribution $N(0, \sigma_i^2)$. The third part is where asymptotic kicks since

$$\begin{aligned} \sqrt{N}\hat{\beta}_{WG} - \beta &\sim N(0, \Sigma) \\ \hat{\beta}_{WG} - \beta &\sim N(0, \frac{1}{N}\Sigma) \\ \frac{1}{T}\sum x_{it}(\hat{\beta}_{WG} - \beta) &\sim N(0, \left(\frac{1}{T^2}\sum x_{it}^2\right)\frac{1}{N}\Sigma) \end{aligned}$$

Then $\hat{\theta}_i - \theta_i$ is a sum of two (non independent) normal variables. Each following $N(0, \frac{\sigma_i^2}{T})$ and $N(0, \left(\frac{1}{T^2}\sum x_{it}^2\right)\frac{1}{N}\Sigma)$ respectively.

Remark. We don't know the Σ and there's a need to replace it by $\hat{\Sigma}$.

Recall the lectures... Assume a simple stripped down model where $y_{it} = \alpha_i + \epsilon_{it}$, two cases

1. $y_{it} = \alpha_i + \epsilon_{it} \sim N(\alpha_i, \sigma^2)$
2. $y_{it} = \alpha_i + \epsilon_{it} \sim N(\alpha_i, \sigma_i^2)$

The estimator of α_i is

$$\hat{\alpha}_i = \bar{y}_i$$

Estimate the σ^2 and σ_i^2 The classical incidental parameter problem appears when we want to estimate σ^2 or σ_i^2 because we only have an estimate of $\hat{\alpha}_i = \bar{y}_i$.

1. σ^2:

$$\hat{\sigma}^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \hat{\alpha}_i)^2$$

   - The mean of the estimator is $\mathbb{E}\left[\hat{\sigma}^2\right] = \sigma^2(1 - \frac{1}{T})$. Asymptotically **BIASED**. (Though We can deploy bias correction techniques.)
   - The variance of the estimator is $\mathrm{Var}\left(\hat{\sigma}^2\right) = \frac{2\sigma^4}{NT}(1 - \frac{1}{T})$

2. σ_i^2:

$$\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T}(y_{it} - \hat{\alpha}_i)^2$$

   This corresponds to the *sufficient statistics* we will be talking about later.

Estimate the variance of α If we treat α_i for each i as a fixed number, and be agnostic about the possible latent distribution G_α from which α_i is drawn from. We don't talk about the variance of α. Yet if we instead believe that all α_i are drawn iidly from a distribution G, then it is interesting to estimate something about G (e.g. the moment of α).
For example, we want to estimate the **first moment** of α by

$$\frac{1}{N}\sum \hat{\alpha}_i = \frac{1}{N}\sum \alpha_i + \frac{1}{N}\sum \bar{\epsilon}_i$$
$$\mathbb{E}\left[\frac{1}{N}\sum \hat{\alpha}_i\right] = \mathbb{E}\left[\alpha_i\right] + 0$$

This is unbiased.
Yet if we want to estimate the **second moment** of α by

$$\begin{aligned}\frac{1}{N}\sum (\hat{\alpha}_i)^2 &= \frac{1}{N}\sum (\alpha_i + \bar{\epsilon}_i)^2 \\ &= \frac{1}{N}\sum \alpha_i^2 + \frac{1}{N}\sum \bar{\epsilon}_i^2 + \frac{2}{N}\sum \alpha_i \bar{\epsilon}_i \\ \mathbb{E}\left[\frac{1}{N}\sum (\hat{\alpha}_i)^2\right] &= \mathbb{E}\left[\alpha_i^2\right] + \frac{1}{T}\sigma_i^2 + 0\end{aligned}$$

This is biased and the bias is $\frac{1}{T}\sigma_i^2$.
This is the issue we encountered when we go from the first moment of α to the second moment. It would be interesting to estimate the $G_\alpha(\alpha_i)$ directly using non parametric convex optimization methods.
Or when σ_i^2 is heterogeneous, it would be (challenging/interesting) to estimate $H_{\alpha,\sigma}(\alpha_i, \sigma_i)$ directly.

1.2 Mixture model

1.2.1 Location mixture

The model is the following

$$y_{it} = \theta_i + \epsilon_{it} \quad \text{where} \quad \epsilon_{it} \sim N(0,1)$$

Thus

$$\hat{\theta}_i = \frac{1}{T_i}\sum y_{it} \sim N(\theta_i, 1/T_i)$$

The likelihood function $L(F|y)$ (optimizing over distribution function F given the observed y) is

$$L(F|y) = \prod_{i=1}^{N} \int \prod_{1}^{T_i} \phi(y_{it} - \theta_i) dF(\theta_i)$$

Instead of focusing on each observation y_{it} we can also focus on the mean $\hat{\theta}_i = \bar{y}_i$. If we utilize $\hat{\theta}_i - \theta_i \sim N(0, 1/T_i)$, we can write the likelihood function as

$$L(F|y) = \prod_{i=1}^{N} \int \phi((\hat{\theta}_i - \theta_i)\sqrt{T_i})\sqrt{(}T_i)dF$$
$$l(F|y) = \sum_{i=1}^{N} \log \int \phi((\hat{\theta}_i - \theta_i)\sqrt{T_i})\sqrt{(}T_i)dF$$

Optimizing over all possible function F neccesitates some kind of discrete approximation. The most common one is the grid approximation. We can also use the EM algorithm to optimize the likelihood function. Let f_j approximate the value of dF on the grid

$$\max_f \left\{ \sum_{i=1}^{N} \log g_i \big| g = Af, \sum_j f_j \Delta_j = 1, f \geq 0 \right\}$$

Remark. For a reader not as versed in mathematics as she should be. $A_{i*}f = \sum \sqrt{T_i}\phi((\hat{\theta}_i - \theta_j)\sqrt{T_i})f_j\Delta_j$. We use $\sum f_j\Delta_j$ to approximate the integral $\int dF$ as one can imagine.

This a convex objective function subject to linear equality and inequality constraints. The EM algorithm is a natural choice to optimize this function. The E-step is to calculate the expectation of the log-likelihood function given the observed data and the current estimate of the parameter. The M-step is to maximize the expectation of the log-likelihood function with respect to the parameter. The algorithm iterates between these two steps until convergence. Often, the dual formulation of a convex objective is more efficient than the primal.

$$\max_f \left\{ \sum_{i=1}^{N} \log(v_i) | A^T v = n1_p, v \geq 0 \right\}$$

Question 1. Derive on your own for practice

Solution. We define the multiplier μ_1', μ_2, λ' for the two equality and one inequality contstraints. The dual objective function is

$$\min_{f,g} \left\{ -\sum_{i=1}^{N} \log g_i + \mu_1'(g - Af) + \mu_2(\sum_j f_j\Delta_j - 1) - \lambda' f \right\}$$

Minimize over each g_i gives the condition

$$\frac{1}{g_i} = \mu_{1i}$$

Minimize over each f_i gives the condition

$$-(\mu_1' A)_j + \mu_2 - \lambda_j = 0$$

Therefore the objective is

$$\begin{array}{ll} & \{\sum \log \mu_{1i} + n - \mu_2\} \\ \text{subject to} & -\mu_1' A + \mu_2 1_p - \lambda = 0 \\ & \lambda \geq 0 \end{array}$$

Thus the dual problem is

$$\begin{array}{ll} \text{maximize}_{\mu_1',\mu_2} & \{\sum \log \mu_{1i} + n - \mu_2\} \\ \text{subject to} & \mu_1' A \leq \mu_2 1_p \\ & \mu_1 \geq 0 \end{array}$$

□

1.2.2 Scale mixture

The model is the following:

$$y_{it} = \sigma_i \epsilon_{it} \quad \text{where} \quad \epsilon_{it} \sim N(0,1)$$

Similarly

$$s_i = \hat{\sigma}_i^2 = \frac{1}{m_i} \sum_{t=1}^{T_i} y_{it}^2$$

But what's the distribution of $\hat{\sigma}_i$ (which is not so obvious relative to $\hat{\theta}_i$)? Well, $\frac{\sum y_{it}^2}{\sigma_i^2}$ follows a Gamma distribution with shape parameter $r_i = (T_i)/2$ and scale parameter $s_i = \sigma_i^2 / r_i = 2\sigma_i^2 / T_i$.

Question 2. Why is it gamma distribution?

Solution. The sum of k independent standard normal variable X follows a $\chi^2(k)$ distribution. The mean of n independent $\chi^2(k)$ distribution variables K follows a gamma distribution $\gamma(nk/2, 2/n)$. The equivalence lies in here: if $X \sim \gamma(v/2, 2)$ (in the shape-scale parametrization), then X is identical to $\chi^2(v)$, the chi-squared distribution with v degrees of freedom. Conversely, if $Q \sim \chi^2(v)$ and c is a positive constant, then $cQ\ \gamma(v/2, 2c)$. □

Remark. The Gamma distribution $\gamma(k, \theta)$ (shape, scale) has the following distribution function

$$f(x|k,\theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} \quad F(x|k,\theta) = \frac{1}{\Gamma(k)} \gamma(k, x/\theta)$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$.

Thus the likelihood function is

$$L(F|y) = \prod_{i=1}^{N} \int \gamma(s_i | r_i, \sigma_i) dF(\sigma_i)$$

$$l(F|y) = \sum_{i=1}^{N} \log \int \gamma(s_i | r_i, \sigma_i) dF(\sigma_i)$$

which we can proceed just as in the location mixture case.

1.2.3 Location-scale mixture (independent)

The model is

$$y_{it} = \theta_i + \sigma_i \epsilon_{it} \quad \text{where} \quad \epsilon_{it} \sim N(0,1)$$

The sufficient statistics $\hat{\theta}_i$ and $\hat{\sigma}_i$ are

$$\hat{\theta}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} y_{it} \sim N(\theta_i, \sigma_i^2/T_i)$$
$$\hat{\sigma}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} (y_{it} - \hat{\theta}_i)^2 \sim \gamma(s_i|\alpha = r_i, \beta = \sigma_i^2/r_i)$$

Just as in the previous two cases, we can write the (log) likelihood as

$$l(G_\theta, F_\sigma|y) = \sum_{i=1}^{N} \log \int\int \left[\phi((\hat{\theta}_i - \theta_i)\sqrt{T_i})\sqrt{T_i}\right][\gamma(s_i|r_i,\sigma_i)]\, dG_\theta(\theta_i) dF_\sigma(\sigma_i)$$

For **estimation**, we can first solve for $\hat{F}_\sigma$ and solve for $\hat{G}_\theta$ given $\hat{F}_\sigma$. There are two compuatation methods.

- Reexpress the Gaussian component as Student's t therefore eliminating the dependence on σ_i.
- Iterate between the Gamma and Gaussian component of the likelihood. (Specific to this independent prior assumption.)

1.2.4 Location-scale mixture (general)

The most general Gaussian location-scale mixture with covariate effects.

Question 3. It seems that in the paper. They assume that the **TRUE** covariate effect is known by constructing a profile likelihood to estimate the true β (covariate effects).w

Remark. The $\hat{\beta}$ that maximizes profile likelihood function of the parameter of interests β $L_{\hat{\alpha}(\beta)}(\beta)$ follows

1. $\hat{\alpha}(\beta) = \arg\max_\alpha L(\alpha, \beta)$
2. $\hat{\beta} = \arg\max_\beta L_{\hat{\alpha}(\beta)}(\beta)$

Note that maximizing profile likelihood function to get estimates gives rise to the infamous **incidental parameter problem** due to the fact that

$$\hat{\alpha}(\beta) = \frac{1}{T_i}\sum \log(f(\alpha,\beta)) \not\to \mathbb{E}\left[\log f(\alpha,\beta)\right]$$

$$y_{it} = x_{it}\beta + \theta_i + \sigma_i\epsilon_{it} \quad \text{where} \quad \epsilon_{it} \sim N(0,1)$$

Given a true β, it is straightforward that

$$y_{it}|\mu_i, \sigma_i, \beta \sim N(x_{it}\beta + \mu_i, \sigma_i^2)$$

The sufficient statistics for

- μ_i: $\bar{y}_i - \bar{x}_i\beta \sim N(\theta_i, \frac{\sigma_i^2}{T_i})$

  contains the between information

- σ_i^2: $\frac{1}{T_i-1}\sum_{t=1}^{T_i}(y_{it} - x_{it}\beta - \mu_i)^2$ It is worth mentioning that

  $$S_i|\mu_i, \sigma_i^2, \beta \sim \gamma(r_i, \sigma_i^2/r_i) \quad \text{where} \quad r_i = (m_i - 1)/2$$

  contains the within information (deviations from the individual means)

Remark. Unlike in the pure scale mixture, we don't know the true location parameter. instead of $y_{it} - x_{it}\beta - \theta_i$ as in the scale mixture case, we need to use the sufficient statistics for θ_i, which is $\frac{1}{T}\sum(y_{it} - x_{it}\beta)$. Each $y_{it} - x_{it}\beta - (\bar{y}_i - \bar{x}_i\beta)$ is a normal variable with mean 0 and variance $(1 - 1/T_i)\sigma_i^2$.

$$y_{it} - x_{it}\beta - \bar{y}_i + \bar{x}_i\beta \sim N(0, (1 - 1/T_i)\sigma_i^2)$$

If we *may* assume that $z_{it} = y_{it} - x_{it}\beta - \bar{y}_i + \bar{x}_i$ is independent of $z_{it'}$ (which is not, the covariance is $\frac{1}{T(T-1)}\sigma_i^2$), then the sufficient statistics

$$S_i' = \sum(\frac{1}{\sqrt{T_i - 1}} z_{it})^2 \underbrace{\sim}_{\text{if we may assume}} \sum \frac{\sigma_i^2}{T_i} Z_{it}^2 \sim \frac{\sigma_i^2}{T_i}\chi^2(T_i) \sim \gamma(\frac{T_i - 1}{2}, \frac{2\sigma^2}{T_i})$$

where $Z_{it} \sim N(0,1)$ and are iid.

Question 4. Why is the derived distribution of S_i' different from $\gamma(\frac{T_i-1}{2}, \frac{2\sigma_i^2}{T_i-1})$?

Remark. The orthogonality between the within and between information no longer holds here. (Why does it hold in the classical Gaussian panel data?)

The likelihood function is

$$\begin{aligned} l(\beta, h(\theta,\sigma)|y) &= \prod_{i=1}^{N} g_i(\beta, \theta_i, \sigma_i | y_{i1}, \ldots, y_{iT}) \\ &= \prod_{i=1}^{N} \int\int \prod \frac{1}{\sigma}\phi(\frac{y_{it} - x_{it}\beta - \theta_i}{\sigma}) h(\theta,\sigma) d\theta d\sigma \\ &= K \prod_{i=1}^{N} \int S_i^{1-r_i} \int\int \frac{1}{\sigma}\phi(\frac{\bar{y}_i - \bar{x}_i\beta - \theta_i}{\sigma}) \frac{e^{-R_i}R_i^{r_i}}{S_i\Gamma(r_i)} h(\theta,\sigma) d\theta d\sigma \end{aligned}$$

where

$$R_i = \frac{r_i S_i}{\sigma_i^2} \quad K = \prod_{i=1}^{N} \left(\frac{\Gamma(r_i)}{r_i^{r_i}}(\frac{1}{\sqrt{2\pi}})^{T_i - 1}\right)$$

Question 5. The true β is unknown. Therefore, we can not condition on it. How about the so called profile likelihood? How to compare it with the *FORBIDDEN* approach of getting fixed effect estimates from the WG estimation? Since the θ_i is regarded as *NUISANCE* parameters in the WG estimation, how low the status is...! Poor θ_i!

Example 1.2.1. Consider a simple model where

$$y_{it} = \alpha_i f(x_{it}; \beta)\epsilon_{it}$$

and $\epsilon | \alpha_i, x_{it} \sim \mathcal{P}(1)$ Then we have

$$y_{it} | \alpha_i, (m_{it} = f(x_{it}; \beta)) \sim \mathcal{P}(\alpha_i m_{it})$$