R301b Panel Data

James Legrand[1]

Lent Term, 2023-2024

Contents

[1] A changelog and archive can be found at github.com/james-legrand/Metrics-Notes.

1 Binary Response Models

1.1 General Textbooks

- Wooldridge, J. (2010). Econometric Analysis of Cross-Section and Panel Data. MIT.
- A. C. Cameron and P. K. Trivedi (2005) Microeconometrics: Methods and Applications, Cambridge University Press
- J. Freidman, T. Hastie, R. Tibshirani (2009). The Elements of Statistical Learning. Springer, publishers.

1.2 Limited Dependent Variables

1.2.1 Observational rules

Assume $\exists -\infty < y^* < \infty$ and:

$$y = \mathbf{1}(\infty < y^* < \infty)y*$$

Trivially this *observational rule* represents the case with no "limit" on the dependent variable.

Binary Response:

$$y = \mathbf{1}(y^* > \alpha)$$

Multinomial Response:

$$y = argmax(\mathbf{y*})$$

1.3 Revealed vs Stated Preference

RP:
+ Using a random utility maximisation framework, RP theory can extract information from choice behaviour based on discrete information embodied in the chosed alternative within a well defined choice set, together with attributes of the DM and a set of characterstics defining the alternatives.
- Reliable to estimate current market behaviour, but usually insufficient variation across key factors along with colinearity
- Deficient data as number of attributes ovserved not measurable

SP:
+ Elicits preferences of consumer prior to choice made.
= Uncertainty here rooted in DM's ability to accurately report preferences, instead of with the analyst in RP world
- Difficult to design a survey that is not subject to bias

1.4 Binary Response Model

We have a random sample of observations on y_i and x_i, where y_i is a binary variable and x_i is a vector of explanatory variables. $i = 1, ..., N$.

$$E(y_i = 1|x_i) = Pr(y_i = 1|x_i) = F(x_i'\beta)$$

where F is some monotonically increasing CDF.

Note:-

Here we do not need necessarily to choose a parametric form of F(.)
We identify K parameters in β but unlike OLS we cannot separately identify σ

1.4.1 Deriving $\hat{\beta}$ (low emphasis)

We have N independent observations on y_i and x_i.

The probability density of y_i conditional on x_i is: $\begin{cases} F(y_i|x_i,\beta) & \text{if } y_i = 1 \\ 1 - F(y_i|x_i,\beta) & \text{if } y_i = 0 \end{cases}$

The joint distribution of the n data point sequence we observe is:

$$f(y_1, ..., y_n|x_1, ..., x_n, \beta) = \prod_{i=1}^{n} F(y_i|x_i,\beta)^{y_i}(1 - F(y_i|x_i,\beta))^{1-y_i}$$

We can write this equivalently as a likelihood function, depending on β:

$$L(\beta) = \prod_{i=1}^{n} F(y_i|x_i,\beta)^{y_i}(1 - F(y_i|x_i,\beta))^{1-y_i}$$

If the model is correctly specified then MLE $\hat{\beta_N}$ is consistent, efficient and asymptotically normal.

We work with the log-likelihood function:

$$l(\beta) = \sum_{i=1}^{n} y_i log(F(x_i'\beta)) + (1 - y_i)log(1 - F(x_i'\beta))$$

We then find the critical point:

$$\begin{aligned} \frac{\partial l(\beta)}{\partial \beta} &= \sum_{i=1}^{n} y_i \frac{f(x_i'\beta)}{F(x_i'\beta)} x_i - (1 - y_i)\frac{f(x_i'\beta)}{1 - F(x_i'\beta)} x_i \\ &= \sum_{i=1}^{n} \frac{y_i - F(x_i'\beta)}{F(x_i'\beta)(1 - F(x_i'\beta))} f(x_i'\beta) x_i \\ &= \sum_{i=1}^{n} \tilde{u_i} x_i \quad (*) \end{aligned}$$

where $\tilde{u_i}$ is the *generalised-residual*.

Note:-

Note that $\frac{\partial l(\beta)}{\partial \beta}$ is non linear in β, in general no analytical solution can be found so we resort to numerical methods.
If $l(\beta)$ is (strictly) concave, as in the Probit and Logit, then MLE $\hat{\beta_N}$ is unique.
We can use (*) to test for model misspecification based upon, for example LM tests.

1.4.2 Issues with Binary Response Models

Heterogenous effects in non-linear models:

Potential outcomes in the binary model are given by:

$$Y(x) = \mathbf{1}(\alpha + \beta x + u \geq 0)$$

The effect of a change from x to x' for an idvidual with error u is:

$$Y(x') - Y(x) = \mathbf{1}(\alpha + \beta x' + u \geq 0) - \mathbf{1}(\alpha + \beta x + u \geq 0)$$

Note:-

This is distinct from a linear model as here partial effects are a function of both x *and u*.
In linear regression u is additive and thus partial effects are independent of u.
Now we need to specficy (in the parametric case) a distribution for u on top of GM assumptions and beyond iid and first moments. Now need F()

Suppose $\beta > 0$ and $x' > x$. Then:

value of u	$Y(x)$	$Y(x')$	$Y(x') - Y(x)$
$-u \leq \alpha + \beta x \leq \alpha + \beta x'$	1	1	0
$-u \geq \alpha + \beta x' \geq \alpha + \beta x$	0	0	0
$\alpha + \beta x \leq -u \leq \alpha + \beta x'$	0	1	1

The effects are either 0 or 1 dependent on the value of u and this differs across individuals.
As this u is unobservable this formulation is not yet of immediate use.

1.4.3 The Partial Effects

Consider the partial effect conditional on two values of X, but averaged wrt u.

$$\begin{aligned} E_u[Y(x') - Y(x)] &= E_u[1(\alpha + \beta x' + u \geq 0)] - E_u[1(\alpha + \beta x + u \geq 0)] \\ &= Pr(-u \leq \alpha + \beta x') - Pr(-u \leq \alpha + \beta x) \\ &= F(\alpha + \beta x') - F(\alpha + \beta x) \end{aligned}$$

Then $E_u[Y(x') - Y(x)]$ denotes the proportion of units in the population had their outcomes been affected from a change from x to x' i.e. those with $\alpha + \beta x \leq -u \leq \alpha + \beta x'$.

1.4.4 The General Partial Effect over continuous X

For continuous X we can instead write:

$$\frac{\partial E_u[Y(x)]}{\partial x} = \frac{\partial F(\alpha + \beta x)}{\partial x} = \beta f(\alpha + \beta x)$$

In binary response models we can regard PEs as a random variable associated with X.
This means we can compute a PE for each observation in a sample.
In this sense it may be of interest to obstain summary measures of its distribution, like the mean or median.

Definition 1.4.1

Partial Effect of x_k:
Initial value of (x, ε) is (x_0, ε_0).
The change is $x_k + \Delta_k$, where Δ_k is a vector of zeros except at position k, a 1.

$$PE_k(x_0) = g(x_0 + \Delta_k, \beta) - g(x_0, \beta)$$

Average Partial Effect:
APE_k is $PE_k(x_0)$ averaged over the distribution of observables x.

$$\text{Continuous } X : APE_k = E_x[PE_k(x_0)] = E_X[\beta_k f(x'\beta)]$$

$$\text{Discrete } X : APE_k = E_x[PE_k(x_0)] = E_X[F([x + \Delta_k]'\beta) - F(x'\beta)]$$

Partial Effect at the Average:

$$PEA_k = PE_k(x_0 = E(x))$$
$$= g(E(x) + \Delta_k, \beta) - g(E(x), \beta)$$

Note:-

PEs in Linear Regression:
In this model with constant linear partial effects:

$$PE_k(x_0) = APE_k = PEA_k = \beta_k$$

But we can introduce non constant PEs through multiplicative effects- interactions between observables.

Definition 1.4.2

The Probit Model:

$$Y_i* = x'\beta + u_i; \quad u_i \sim N(0, \sigma^2)$$
$$Y_i = \mathbf{1}(Y_i* > 0)$$

We can think of Y_i* as an index of net utility from an action involving two (or more) choices/states.

$$\Rightarrow Pr(Y_i = 1|x_i) = Pr(Y_i* > 0|x_i)$$
$$= Pr(u_i > -x_i'\beta|x_i) = 1 - F(-x_i'\beta)$$
$$= Pr(\frac{u_i}{\sigma} > -\frac{x_i'\beta}{\sigma}|x_i) = 1 - \Phi(-\frac{x_i'\beta}{\sigma})$$

Thus β is not separately identified to σ, which is thus usually normalised to 1

Example. Common Distributions and their Conditional PEs:
LPM: $F() = I()$, $PE_k = \frac{\partial(x'\beta)}{\partial x_k} = \beta_k$
Probit: $F() = \Phi()$, $PE_k = \beta_k \phi(x'\beta)$

Logit: $F() = \Lambda()$, $PE_k = \beta_k \Lambda(x'\beta)(1 - \Lambda(x'\beta))$

While the LPM may give probability estimates outside the unit interval, and have constant PEs (and het by construction). If we only care about PEs for the average individual then this may not matter, especially in large samples.

1.4.5 Sparse Covariates

With no covariates or sparse and discrete covariates, linear models and the associated estimation techniques (e.g. 2SLS) are no less appropraite for LDVs than for other kinds of DVs. (Angrist JBES 2001)

1.4.6 Unobserved Heterogeneity

Proposition 1.4.1. In probit models, neglected (unobserved) heterogeneity is a much more serious problem than in linear models as even if it is independent of X, probit coefficients are inconsistent.

We will show that this statement is only **partially** true.

Proof. Probit Parameters become inconsistent:
We illustrate with a one period cross sectional probit model:

$$Pr(y_i = 1|x_i, c_i) = \Phi(x_i'\beta + c_i)$$

where x_i is $k \times 1$, $x_{1i}/equiv1$ and c_i is an unobserved scalar, denoting the unobserved heterogeneity
Our parameter of interest is the partial effect estimation of a given x_j on $Pr(y_i = 1|x_i, c_i)$

Consider the standard latent variable model:

$$y_i* = x_i'\beta + \gamma c_i + \varepsilon_i$$

$$y_i = \mathbf{1}(y_i* > 0)$$

$$\varepsilon_i|x_i, c_i \sim N(0, 1)$$

$$c_i|x_i \sim N(0, \tau^2)$$

c_i normalised such that $E(c) = 0$ and assumed independent of x.

$$\gamma c_i + \varepsilon_i|x_i \sim N(0, \gamma^2\tau^2 + 1)$$

$$\Rightarrow Pr(y_i = 1|x_i) = Pr(\gamma c_i + \varepsilon_i > -x_i'\beta|x_i)$$
$$= \Phi(-x_i'\beta/\sigma)$$

where $\sigma^2 = \gamma^2\tau^2 + 1$

Thus we see that this probit model of y on x consistently estimates β/σ

Since $\sigma = (\gamma^2\tau^2 + 1)^{1/2} > 1$ then $|\beta_j/\sigma| < |\beta_j|$ we have **attenuation bias** in estimated paramters. Attempting to normalise $\gamma^2\tau^2 + 1 = 1$ would imply $\gamma^2\tau^2 = 0$ and so there is no unobserved heterogeneity.

□

Proof. Unobserved heterogeneity does not affect partial effect estimates.
For given values of x_i and c_i the partial effect of interest is:

$$\frac{\partial Pr(y_i = 1|x_i, c_i)}{\partial x_{ij}} = \beta_j \phi(x_i'\beta + \gamma c_i)$$

Given $E(c) = 0$, evaluating at $c = 0$:
Ground truth: $\beta_j \phi(x_i'\beta$
Our estimate: $()\beta_j/\sigma)\phi(x_i'\beta\sigma)$
Thus probit does not give correct partial effects for $c = 0$

Estimating APE oo oo AA AA by averaging over c:

$$E_c[\frac{\partial Pr(y_i = 1|x_i, c_i)}{\partial x_{ij}}] = E_c[\beta_j \phi(x_i'\beta + \gamma c_i)]$$

But:

$$\begin{aligned}
E_c[\frac{\partial Pr(y_i = 1|x_i, c_i)}{\partial x_{ij}}] &= \frac{\partial}{\partial x_{ij}} E_c[Pr(y_i = 1|x_i, c_i)] \\
&= \frac{\partial}{\partial x_{ij}} E_c[E_{y|x,c}[y_i = 1|x_i, c_i]] \\
&= \frac{\partial}{\partial x_{ij}} E_{y|x}[y_i|x_i] \\
&= \frac{\partial}{\partial x_{ij}} Pr(y_i = 1|x_i) \\
&= \frac{\partial}{\partial x_{ij}} Pr(\gamma c_i + u_i > -x_i'\beta) \\
&= \frac{\partial}{\partial x_{ij}} \Phi(-x_i'\beta/\sigma) \\
&= \beta_j/\sigma \phi(x_i'\beta/\sigma)
\end{aligned}$$

For this measurement, the omitted heterogeneity is not a problem when independent of x. We can consistently estimate the APEs, conditional on normality of c and the probit model. The reason why the first PE was inconsistent was due to conditioning on the unobservable c_i, which we cannot do when estimating β.
But we can avoid this issue in estimating APEs by exploiting the law of iterated expectations to removing conditioning on c_i. □

Omitted heterogeneity in linear models is not a problem when independent of x. (subsumed by error term). Omitted heterogeneity in probit models also not a problem when independent of x. As ignoring it consistently estimates APEs.

Note:-

If c is correlated with x or otherwise dependent on x (i.e. $Var(c|x)$ is a function of x), then omission of c will mean we cannot get consistent APE estimates.

3 Multinomial Response Models

3.1 The Random Utility Model

A decision maker i faces a finite choice set Ω_J of dimension J. The decision maker would obtain a certain level of utility from each alternative. The utility that decision maker i derives from alternative j is U_ij, which is known to them but **unknown to the researcher**. Thus alternative j is chosen iff $U_{ij} > U_{ik}$ for all $k \neq j$.
The researcher doesn't observe the decision-maker's utility, only some attributes of the alternatives $\mathbf{v_j}$ (sometimes $\mathbf{v_{ij}}$) $\forall j$, and some attributes of the decision maker $\mathbf{x_i}$. The function is denoted $V_{ij} = V(\mathbf{v_j}, \mathbf{x_i}, \theta)$ where θ is a vector of unknown parameters. Since there are aspects of utility that are not observed, $V_{ij} \neq U_{ij}$, and we decompose utility as $U_{ij} = V_{ij} + \varepsilon_{ij}$, where ε_{ij} captures the unobserved aspects of utility. For example:

- $U_{ij} = \mathbf{v_j'}\omega + \varepsilon_{ij}$, where $\mathbf{v_j}$ is an $L \times 1$ vector of attributes for alternative j, and ω is an $L \times 1$ vector of parameters.
- $U_{ij} = \mathbf{x_i'}\beta + \alpha_j + \varepsilon_{ij}$, where $\mathbf{x_i}$ is an $K \times 1$ vector of attributes for decision maker i, β is a $K \times 1$ vector of parameters and α_j is a scalar alternative-specific constant (capturing average utility of alternative j).
- $U_{ij} = \frac{\mathbf{v_j'}\omega}{\mathbf{x_i'}} + \alpha_j + \varepsilon_{ij}$ denoting interactions between attributes and characteristics.

The researcher doesn't know ε_{ij} and therefore treats it as random, with the joint density of random vector $\varepsilon_\mathbf{i} = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})$ denoted $f(\varepsilon_\mathbf{i})$.

Note:-

How should we interpret ε_{ij}? It is not defined for a choice situation per se, rather it is defined relative to a researcher's representation of that choice situation. What does it mean for this to have a distribution?
Consider a population of people who face the same observed utility as person i. Among these people, the values of the unobserved factors differ. The density $f(\varepsilon_{ij})$ is the distribution of the unobserved portion of utility within the population of people who face the same observed utility as person i. It can also just be thought of as a subjective probability imposed by the researcher.

With this density, we can make probabilistic statements about choices. The probability that decision maker i chooses alternative j is:

$$\begin{aligned} P_{ij} &= P(U_{ij} > U_{ik} \forall k \neq j) \\ &= P(V_{ij} + \varepsilon_{ij} > V_{ik} + \varepsilon_{ik} \forall k \neq j) \\ &= P(\varepsilon_{ij} - \varepsilon_{ik} < V_{ik} - V_{ij} \forall k \neq j) \end{aligned}$$

This probability is the CDF of $\varepsilon_{ij} - \varepsilon_{ik}$ evaluated at $V_{ik} - V_{ij}$. Using the density of ε_{ij}, we can write this as:

$$\begin{aligned} P_{ij} &= P(\varepsilon_{ij} - \varepsilon_{ik} < V_{ik} - V_{ij} \forall k \neq j) \\ &= \int_\varepsilon I(\varepsilon_{ij} - \varepsilon_{ik} < V_{ik} - V_{ij} \forall k \neq j) f(\varepsilon_{ij}) d\varepsilon_i \end{aligned}$$

where $I(\cdot)$ is the indicator function. This is a multidimensional integral over the density of the unobserved portion of utility.

3.1.1 Identification

"Only differences in utility matter"
The absolute level of utility is irrelevant to both the decision maker and the researcher. If a constant is added to all utilities, the ordering of alternatives is unchanged. We can see this from the decomposition of the choice probability:

$$P_{ij} = P(\varepsilon_{ij} - \varepsilon_{ik} < V_{ik} - V_{ij} \forall k \neq j)$$

which only depends on differences. In general this means that the only parameters that can be estimated (identified) are those that capture differences across alternatives. This general statement takes several forms:

- **Alternative-specific constants**: α_j captures the average utility of alternative j. When they are included ε_{ij} has zero mean by construction. However, since only differences in utility matter, only differences in α_j matter, not the absolute level. To reflect this, the researcher must set the level of one constant. **With J alternatives, at most $J-1$ constants are identified, with one of the constants normalised to zero**[1]. It's irrelevant which constant is normalised to zero, just be careful of the interpretation of the other constants.
- **Number of independent error terms**. The choice probabilities take the form of a J dimensional integral over the density of the J error terms. However, recognising that only differences in utility matter, we can reduce the dimension by defining $\tilde{\varepsilon}_{ijk} = \varepsilon_{ij} - \varepsilon_{ik}$, giving:

$$P_{ij} = \int I(\tilde{\varepsilon}_{ijk} < V_{ik} - V_{ij} \forall k \neq j) g(\tilde{\varepsilon}_{ijk}) d\tilde{\varepsilon}_{ijk}$$

  This is a $J-1$ dimensional integral![2]

"The overall scale of utility is irrelevant "
Just as adding a constant to the utility of all alternatives doesn't change the choice, neither does multiplying each alternative's utility by a constant. To account for this, the researcher must normalise the scale of utility.
The standard way to do this is to normalise the variance of the error terms. When utility is multiplied by κ, the variance of ε_{ij} changes by κ^2: $Var(\kappa\varepsilon_{ij}) = \kappa^2 Var(\varepsilon_{ij})$. Therefore **normalising the variance of the error term is equivalent to normalising the scale of utility**.
When errors are assumed i.i.d., normalising is straightforward. We set the error variance to some number, usually something convenient. Since all errors have the same variance by assumption, normalising the variance of any of them sets the variance for them all. The original model becomes equivalent to $U_{ij}^1 = \mathbf{v_j'}\frac{\omega}{\sigma} + \varepsilon_{ij}^1$, with $Var(\varepsilon_{ij}^1) = 1$. The new coefficients reflect the impact of observed variables relative to standard deviation of the unobserved factors.

Example (Probit). $Var(\varepsilon_{ij}) = \sigma^2$, so we can set variance to 1 by dividing by σ.

Example (Logit). $Var(\varepsilon_{ij}) = \sigma^2\pi^2/6$, so we can normalise by setting variance to $\pi^2/6$ (and dividing by σ). Thus

$$Pr(y_i = j|\mathbf{v}) = \frac{e^{\mathbf{v_{ij}'}\omega/\sigma}}{\sum_{j=1}^{J} e^{\mathbf{v_{ij}'}\omega/\sigma}}$$

[1] The researcher could normalize to some value other than zero, of course; however, there would be no point in doing so, since normalizing to zero is easier (the constant is simply left out of the model) and has the same effect.
[2] Since $\varepsilon_\mathbf{i}$ has more elements than $\tilde{\varepsilon}_\mathbf{ij}$, g is consistent with an infinite number of different f's. One dimension of f is not identified, and must be normalised.

Question 1

Show that for a logit parametrisation which sets the normalised variance to 1.6, to convert the probit coefficients to the same scale as logit, they should be multiplied by $\sqrt{1.6}$.

Solution:-

$$U_{ij} = \mathbf{v'_{ij}}\omega + \varepsilon_{ij} \text{ where } Var(\varepsilon_{ij}) = 1.6$$
$$\Rightarrow U^1_{ij} = \mathbf{v'_{ij}}\frac{\omega}{\sqrt{1.6}} + \varepsilon^1_{ij} \text{ where } Var(\varepsilon^1_{ij}) = 1$$

Since probit sets the variance to 1:

$$\omega^p = \frac{\omega^l}{\sqrt{1.6}}$$
$$\Rightarrow \sqrt{1.6}\omega^p = \omega^l$$

3.2 Multinomial Probit

We write the conditional probability of choosing alternative j' as

$$Pr(y_i = j'|\mathbf{v_i}, \theta) = \int_{-\infty}^{V_{j'}-V_1} \cdots \int_{-\infty}^{V_{j'}-V_J} g(\tilde{\varepsilon}_{1j'} \cdots \tilde{\varepsilon}_{Jj'}, \Xi_{\varepsilon J-1}) d\tilde{\varepsilon}_{1j'} \cdots d\tilde{\varepsilon}_{Jj'}$$

g is a multivariate normal density of dimension $J-1$, $\tilde{\varepsilon}_{ij'} = \varepsilon_{ij'} - \varepsilon_{ij}$ as before, and $\Xi_{\varepsilon J-1}$ is the covariance matrix for the error differences.
There is a curse of dimensionality here, there are no closed form expressions for such high dimensional integrals, and for $J > 2$ we have to use simulation methods (SMLE).

Lemma 3.2.1 (Cholesky decomposition). Let A be a $K \times K$ matrix. We say that A possesses a Cholesky decomposition if and only if there exists a lower triangular $K \times K$ matrix L such that its diagonal entries are strictly positive real numbers and

$$A = LL'$$

Example (Trinomial probit).

$$y^*_j = \mathbf{x}'\beta + \varepsilon_j \quad j = 1, 2, 3$$

The deterministic component of utility (for alternative j) is given by $V_j = \mathbf{x}'\beta$ and $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)' \sim N(0, \Sigma)$. We can redefine using error difference to reduce dimensionality to 2:

$$Pr(y = 1|\mathbf{x}) = \int_{-\infty}^{V_1-V_2} \int_{-\infty}^{V_1-V_3} g(\tilde{\varepsilon}_{21}, \tilde{\varepsilon}_{31}, \rho) d\tilde{\varepsilon}_{21} d\tilde{\varepsilon}_{31}$$

where g is bivariate normal and ρ is the correlation between $\tilde{\varepsilon}_{21}$ and $\tilde{\varepsilon}_{31}$.
We can write the probability above as the product of conditionals as below:

$$\begin{aligned} Pr(y = 1|\mathbf{x}) &= Pr(\tilde{\varepsilon}_{21} < V_1 - V_2, \tilde{\varepsilon}_{31} < V_1 - V_3|\mathbf{x}) \\ &= Pr(\tilde{\varepsilon}_{21} < V_1 - V_2|\mathbf{x}) Pr(\tilde{\varepsilon}_{31} < V_1 - V_3|\mathbf{x}, \tilde{\varepsilon}_{21} < V_1 - V_2) \end{aligned}$$

The errors $\tilde{\varepsilon}$ are bivariate normal with covariance matrix Σ. Since Σ is positive definite, it

can be decomposed as $\Sigma = LL'$, where L is a lower triangular matrix.

$$L = \begin{bmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{bmatrix}$$

Using this decomposition, we can construct $\tilde{\varepsilon}$ using *independent* standard normal random variables e_1 and e_2:

$$\begin{bmatrix} \tilde{\varepsilon}_{21} \\ \tilde{\varepsilon}_{31} \end{bmatrix} = \begin{bmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

Thus we can represent the probability $Pr(y = 1|\mathbf{x})$ as:

$$Pr\left(\underbrace{\begin{bmatrix} -\infty \\ -\infty \end{bmatrix}}_{A} \leq \underbrace{\begin{bmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} e_1 \\ e_2 \end{bmatrix}}_{e} \leq \underbrace{\begin{bmatrix} V_1 - V_2 \\ V_1 - V_3 \end{bmatrix}}_{B} |\mathbf{x} \right)$$

where A and B are the respective lower and upper limits of the integrals given above. The two components of $Pr(y = 1|\mathbf{x})$ are:

$$\begin{aligned} Pr(\tilde{\varepsilon}_{21} < V_1 - V_2|\mathbf{x}) &= Pr(\ell_{11}e_1 < V_1 - V_2|\mathbf{x}) \\ &= Pr(e_1 < \frac{V_1 - V_2}{\ell_{11}}|\mathbf{x}) \\ Pr(\tilde{\varepsilon}_{31} < V_1 - V_3|\mathbf{x}, \tilde{\varepsilon}_{21} < V_1 - V_2) &= Pr(\ell_{21}e_1 + \ell_{22}e_2 < V_1 - V_3|\mathbf{x}, e_1 < \frac{V_1 - V_2}{\ell_{11}}) \\ &= Pr(e_2 < \frac{V_1 - V_3 - \ell_{21}e_1}{\ell_{22}}|\mathbf{x}, e_1 < \frac{V_1 - V_2}{\ell_{11}}) \end{aligned}$$

Going forward all probabilities are conditional on $\mathbf{x}$ and I define $b_1 = V_1 - V_2$ and $b_2 = V_1 - V_3$. From before:

$$Pr(y = 1) = Pr\left(e_1 < \frac{b_1}{\ell_{11}}\right) Pr\left(e_2 < \frac{b_2 - \ell_{21}e_1}{\ell_{22}} \middle| e_1 < \frac{b_1}{\ell_{11}}\right)$$

Suppose now we draw a random variable e_1^* from a truncated standard normal density with upper truncation point of b_1/ℓ_{11}. Then the probability can be rewritten as:

$$Pr(y = 1) = Pr\left(e_1 < \frac{b_1}{\ell_{11}}\right) Pr\left(e_2 < \frac{b_2 - \ell_{21}e_1^*}{\ell_{22}}\right)$$

The GHK simulator is the arithmetic mean of the probabilities given by the above for R random draws of e_1^*:

$$Pr^{GHK}(y = 1|\mathbf{x}) = \frac{1}{R}\sum_{r=1}^{R} \Phi\left\{\frac{b_1}{\ell_{11}}\right\} \Phi\left\{\frac{b_2 - \ell_{21}e_1^*}{\ell_{22}}\right\}$$

where Φ is the standard normal CDF.
If $\ell_{21} = 0$ then simulation methods are pointless, the errors are already uncorrelated

The advantage of this expression is the fact that the e's are independent normal distributed random variables and hence the probability can be equivalently expressed as a product of independent but conditioned univariate cumulative density functions.

3.3 Logit

In Multinomial Probit, for large J we need to solve a high dimensional integral. Logit is a highly tractable alternative, where the joint density $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})$ is assumed to be i.i.d. extreme value type I. The density for each unobserved component is

$$f(\varepsilon_{ij}) = e^{-\varepsilon_{ij}} e^{-e^{-\varepsilon_{ij}}}, \quad F(\varepsilon_{ij}) = e^{-e^{-\varepsilon_{ij}}}.$$

As we used before, $\text{Var}(\varepsilon_{ij}) = \pi^2/6$ [3].
The difference between two extreme value type I random variables is a logistic random variable, then $\tilde{\varepsilon}_{ijk} = \varepsilon_{ij} - \varepsilon_{ik}$ has density

$$F(\tilde{\varepsilon}_{ijk}) = \frac{e^{\tilde{\varepsilon}_{ijk}}}{1 + e^{\tilde{\varepsilon}_{ijk}}}$$

Using this density, we can write the choice probability as:

$$Pr(y_i = j|\mathbf{v_i}) = \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}}$$

Clearly all probabilities are positive and sum to 1.

3.3.1 Power and limitations of logit

- **Taste variation** - Logit can represent systematic taste variation (that is, taste variation that relates to observed characteristics of the decision-maker) but not random taste variation (differences in tastes that cannot be linked to observed characteristics).
- **Substitution patterns** - The logit model implies proportional substitution across alternatives, given the researcher's specification of representative utility (see later sections on IIA).

Independence of irrelevant alternatives (IIA)
For any two alternatives j and k, the ratio of the logit probabilities is:

$$\begin{aligned}\frac{P_{ij}}{P_{ik}} &= \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}} \frac{\sum_{k=1}^{J} e^{V_{ik}}}{e^{V_{ik}}} \\ &= \frac{e^{V_{ij}}}{e^{V_{ik}}} \\ &= e^{V_{ij} - V_{ik}}\end{aligned}$$

This ratio does not depend on any alternatives other than j and k. That is, the relative odds of choosing j over k is the same no matter what other alternatives are available or what the attributes of the other alternatives are. Since the ratio is independent from alternatives other than j and k, it is said to be independent from "irrelevant" alternatives.

Example (IIA and Three Buses). A car and blue bus are initially available. Assume $Pr_c = Pr_{bB} = 0.5$, thus $Pr_c/Pr_{bB} = 1$.
Now add a red bus to the choice set Ω_J, and assume the buses are perfect substitutes ($P_{rB} = P_{bB}$). Given IIA, $Pr_c/Pr_{bB} = 1$ holds before and after the introduction of the red bus.
$\Rightarrow Pr_c/Pr_{bB} = Pr_{rB}/Pr_{bB} = 1$
$\Rightarrow Pr_c = Pr_{rB} = Pr_{bB} = 1/3$
We expect the original probability Pr_c not to change when we introduce the red bus, such that $Pr_c = 0.5$ and $Pr_{bB} = 0.25$. We overestimate Pr_{rB} and Pr_{bB}, and underestimate Pr_c.

[3] The mean of this distribution is not zero, but since only differences in utility matter this is not a problem

Question 2

Let U_{ij} denote the unobserved utility for individual $i \in 1, \ldots, n$ who chooses from large gas cars (lgc), small gas cars (sgc) and small electric cars (sec) $j \in$ lgc,sgc,sec. A model which is additive in a linear index and error term is given by:

$$U_{ij} = \mathbf{x}'_{\mathbf{ij}}\beta + \varepsilon_{ij}$$

where ε_{ij} is an unobserved error term. The error terms are assumed to be i.i.d. extreme value type I. The choice probabilities are:

$$P_{lgc} = 0.66, \quad P_{sgc} = 0.33, \quad P_{sec} = 0.01$$

Suppose a subsidy for electric cars is introduced, and this subsidy raises P_{sec} to 0.1. What are the new values of P_{lgc} and P_{sgc}?

Solution:-

By the logit model, the probability for each of the gas cars would be predicted to drop by the same percentage:
The probability for the large gas car would drop by 10% of its original value, from 0.66 to 0.594 and the probability for the small gas car would drop by 10% of its original value, from 0.33 to 0.297.
In terms of absolute numbers, the increase in demand for the small electric car (0.09) is predicted to come twice as much from the large gas car (0.06) as from the small gas car (0.03).
This is precisely the notion of proportional substitution.

3.3.2 Logit variants

Conditional Logit

This is the normal logit we've seen before, *conditional* on the observed characteristics/attributes we get a choice probability:

Definition 3.3.1: Conditional Logit

$$\begin{aligned} U_{ij} &= V_{ij} + \varepsilon_{ij} \quad \text{where } \varepsilon_{ij} \sim \text{EV1} \\ V_{ij} &= \mathbf{v}'_{\mathbf{ij}}\omega \\ Pr(y_i = j|\mathbf{v}) &= \frac{e^V_{ij}}{\sum_{k=1}^J e^V_{ik}} \\ &= \frac{e^{\mathbf{v}'_{\mathbf{ij}}\omega}}{\sum_{k=1}^J e^{\mathbf{v}'_{\mathbf{ik}}\omega}} \end{aligned}$$

For example, when there are two alternatives, the conditional logit model is:

$$\begin{aligned} Pr(y_i = 1|\mathbf{v}) &= \frac{e^{\mathbf{v}'_{\mathbf{i1}}\omega}}{e^{\mathbf{v}'_{\mathbf{i1}}\omega} + e^{\mathbf{v}'_{\mathbf{i2}}\omega}} = \frac{e^{V_{i1}}}{e^{V_{i1}} + e^{V_{i2}}} \\ &= \frac{1}{1 + e^{\mathbf{v}'_{\mathbf{i2}}\omega - \mathbf{v}'_{\mathbf{i1}}\omega}} \end{aligned}$$

The analyst may observe both alternative specific and alternative invariant characteristics, models including both are referred to as **hybrid models**. This could be done by specifying an additively separable model: $V_{ij} = \mathbf{v'_{ij}}\omega + \mathbf{x'_i}\beta$, or alternatively allowing a bunch of interactions: $V_{ij} = \mathbf{v'_{ij}}\omega + \sum_k^K \mathbf{v_{ij}} \times \mathbf{x_{ik}}\delta_{jk}$.

3.3.3 Identification - Aside

A consumer facing the choice between J alternatives will choose the alternative in choice set Ω_J with the highest utility. As before, this implies only differences in utility matter.

Example.

$$U_{i1} = \alpha_1 + \mathbf{x'_i}\beta_\mathbf{1} + \varepsilon_{i1} \quad U_{i2} = \alpha_2 + \mathbf{x'_i}\beta_\mathbf{2} + \varepsilon_{i2}$$

With J alternatives, only J-1 ASC's are identified (as earlier). Similarly only $\beta_{12} = \beta_1 - \beta_2$ is identified. Normalisations: set $\alpha_1 = 0$ and $\beta_1 = 0$.

3.3.4 Elasticities

Below we derive own and cross elasticities for the logit model:

$$U_{ij} = V_{ij} + \varepsilon_{ij}, \quad P_{ij} = \frac{e^{V_{ij}}}{\sum_l e^{V_{il}}}$$

Own Partial effects - We first calculate the partial effect of a change in attribute k of alternative j for individual i on the probability of choosing alternative j:

$$\begin{aligned}
\frac{\partial P_{ij}}{\partial v_{ij}^k} &= \frac{\partial}{\partial v_{ij}^k}\left(\frac{e^{V_{ij}}}{\sum_l e^{V_{il}}}\right) \\
&= \frac{\sum_l e^{V_{il}} \frac{\partial V_{ij}}{\partial v_{ij}^k} e^{V_{ij}} - e^{V_{ij}} \frac{\partial V_{ij}}{\partial v_{ij}^k} e^{V_{ij}}}{\left(\sum_l e^{V_{il}}\right)^2} \\
&= \frac{\partial V_{ij}}{\partial v_{ij}^k}\left(\frac{e^{V_{ij}}}{\sum_l e^{V_{il}}} - \left(\frac{e^{V_{ij}}}{\sum_l e^{V_{il}}}\right)^2\right) \\
&= \frac{\partial V_{ij}}{\partial v_{ij}^k}(P_{ij} - P_{ij}^2) \\
&= \frac{\partial V_{ij}}{\partial v_{ij}^k} P_{ij}(1 - P_{ij})
\end{aligned}$$

If representative utility is linear in $\mathbf{v_{ij}}$, then $\frac{\partial V_{ij}}{\partial v_{ij}^k} = \omega_k$. Thus:

$$\frac{\partial P_{ij}}{\partial v_{ij}^k} = \omega_k P_{ij}(1 - P_{ij})$$

Own elasticity - Denote the elasticity of P_{ij} with respect to v_{ij}^k as $E_{jv_{ij}^k}$.

$$\begin{aligned}
E_{jv_{ij}^k} &= \frac{\partial P_{ij}}{\partial v_{ij}^k}\frac{v_{ij}^k}{P_{ij}} \\
&= \frac{\partial V_{ij}}{\partial v_{ij}^k} P_{ij}(1 - P_{ij})\frac{v_{ij}^k}{P_{ij}} \\
&= \frac{\partial V_{ij}}{\partial v_{ij}^k} v_{ij}^k(1 - P_{ij})
\end{aligned}$$

If representative utility is linear in $\mathbf{v_{ij}}$, then:

$$E_{jv_{ij}^k} = \omega_k v_{ij}^k (1 - P_{ij})$$

Cross partial effects - We now calculate the partial effect of a change in attribute h of alternative j for individual i on the probability of choosing alternative l:

$$\begin{aligned}
\frac{\partial P_{ij}}{\partial v_{ih}^k} &= \frac{\partial}{\partial v_{ih}^k}\left(\frac{e^{V_{ij}}}{\sum_l e^{V_{il}}}\right) \\
&= \frac{-e^{V_{ij}} e^{V_{ih}} \frac{\partial V_{ih}}{\partial v_{ih}^k}}{\left(\sum_l e^{V_{il}}\right)^2} \\
&= -\frac{\partial V_{ih}}{\partial v_{ih}^k}\left(\frac{e^{V_{ij}}}{\sum_l e^{V_{il}}}\right)\left(\frac{e^{V_{ih}}}{\sum_l e^{V_{il}}}\right) \\
&= -\frac{\partial V_{ih}}{\partial v_{ih}^k} P_{ij} P_{ih}
\end{aligned}$$

If representative utility is linear in $\mathbf{v_{ij}}$, then:

$$\frac{\partial P_{ij}}{\partial v_{ih}^k} = -\omega_k P_{ij} P_{ih}$$

Cross elasticity - Denote the elasticity of P_{ij} with respect to v_{ih}^k as $E_{jv_{ih}^k}$.

$$\begin{aligned}
E_{jv_{ih}^k} &= \frac{\partial P_{ij}}{\partial v_{ih}^k} \frac{v_{ih}^k}{P_{ij}} \\
&= -\frac{\partial V_{ih}}{\partial v_{ih}^k} P_{ij} P_{ih} \frac{v_{ih}^k}{P_{ij}} \\
&= -\frac{\partial V_{ih}}{\partial v_{ih}^k} v_{ih}^k P_{ih}
\end{aligned}$$

If representative utility is linear in $\mathbf{v_{ij}}$, then:

$$E_{jv_{ih}^k} = -\omega_k v_{ih}^k P_{ih}$$

Interpretation of cross elasticity

The cross-elasticity of demand is the same for all choices, i.e. the formula for the cross-elasticity doesn't contain anything relating to the j^{th} alternative.
If the k^{th} attribute of h changes, then the effect on substitution probabilities to other alternatives (here alternative j) is independent of the share originally with alternative j.

- An improvement in one alternative draws proportionately from all other alternatives.
- A worsening in one alternative propels proportionately to all other alternatives, i.e. increase in price, lost demand is redistributed in equal proportions.

Example. Assume there is a single attribute denoting cost v^c per alternative, $V_{ij} = \omega v_j^c$. Then for 3 alternatives we can display $E^L_{jv_h^c}$ (the cross elasticity for alternative j - rows - with respect to the cost of alternative h - columns) as:

$$E^L_{jv_h^c} = \begin{bmatrix}
(1 - P_{i1})\omega v_1^c & -P_{i2}\omega v_2^c & -P_{i3}\omega v_3^c \\
-P_{i1}\omega v_1^c & (1 - P_{i2})\omega v_2^c & -P_{i3}\omega v_3^c \\
-P_{i1}\omega v_1^c & -P_{i2}\omega v_2^c & (1 - P_{i3})\omega v_3^c
\end{bmatrix}$$

For $\omega < 1$ all own price (cross) elasticities are negative (positive).

4 The Mixed Logit Model

The mixed logit model is a highly flexible generalisation of the standard logit from last lecture. It resolves the three limitations of standard logit by allowing for random taste variation, unrestricted substitution patterns and correlation in unobserved factors over time.

4.1 Mixed Logit Setup

Mixed logit probabilities are the integral of standard logit probabilities over a density of parameters, i.e. it can be expressed in the form:

$$P_{ij} = \int L_{ij}(\omega) g(\omega) d\omega \tag{4.1}$$

where L_{ij} is the logit probability evaluated at parameters ω:

$$L_{ij}(\omega) = \frac{e^{V_{ij}(\omega)}}{\sum_{k=1}^{J} e^{V_{ik}(\omega)}}$$

and $g(\omega)$ is a density function. As before $V_{ij}(\omega)$ is a portion of utility which depends on the parameters ω ($U_{ij} = V_{ij} + \varepsilon_{ij}$). If utility is linear in ω, then $V_ij(\omega) = \omega' x_{ij}$ and the mixed logit probability takes the form:

$$P_{ij} = \int \frac{e^{\omega' x_{ij}}}{\sum_{k=1}^{J} e^{\omega' x_{ik}}} g(\omega) d\omega$$

Note:-

In the statistics literature, a *mixed function* is defined as the weighted average of several functions, and the density providing the weights is known as the mixing distribution. Mixed logit is a mixture of the logit function evaluated at different ω's, and the mixing distribution is $g(\omega)$.

4.1.1 Special Cases

Standard logit is a special case, where the mixing distribution $g(\omega)$ is degenerate at fixed parameters w: $g(\omega) = 1$ for $\omega = w$ and $g(\omega) = 0$ otherwise. I.e. it is a point mass at $\omega = w$. The choice probabilities are then given by the standard logit formula:

$$P_{ij} = \frac{e^{w' x_{ij}}}{\sum_{k=1}^{J} e^{w' x_{ik}}}$$

Discrete mixing distributions are also possible, with ω taking on a finite number of values. Suppose ω can take on M different values $w_1, w_2, \ldots, w_M$ with $Pr(\omega = w_m) = s_m$. This is known as the *latent class model* and the choice probabilities are given by:

$$P_{ij} = \sum_{m=1}^{M} s_m \frac{e^{w_m' x_{ij}}}{\sum_{k=1}^{J} e^{w_m' x_{ik}}}$$

This specification is useful if there are M segments in the population, each with its own choice behaviour. The share of each population in segment m is given by s_m.

4.1.2 Estimation

The researcher specifies a distribution for ω (i.e. a distribution for the taste parameters) and aims to estimate the parameters of this distribution[1]. For example if $\omega \sim N(\bar{\omega}, \Sigma_\omega)$, then the researcher estimates $\bar{\omega}$ and Σ_ω. Denote the parameters of the mixing distribution as θ, we now have choice probabilities $P_{ij} = \int L_{ij}(\omega) g(\omega|\theta) d\omega$ which is a function of θ, the parameters ω are integrated out. In this sense, the ω's are similar to the ε_{ij}'s in that both are random terms we integrate out to obtain the choice probabilities.
Note that the dimension of ω is $L \times 1$ with L attributes, maximisation requires the estimation of an L dimensional integral. This is typically done using simulation methods.

4.2 Elasticities

Mixed logit does not exhibit IIA or the restrictive substitution patterns of logit. The ratio of mixed logit probabilities depends on all the data, including attributes other than j and k:

$$\frac{P_{ij}}{P_{ik}} = \frac{\int \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}} g(\omega) d\omega}{\int \frac{e^{V_{ik}}}{\sum_{k=1}^{J} e^{V_{ik}}} g(\omega) d\omega}$$

The denominators of the logit formula are now inside the integrals, and therefore do not cancel as in standard logit. We can further compute elasticities, denote by $E_{jx_h^\ell}$ the elasticity of P_{ij} with respect to attribute h of alternative ℓ:

$$\begin{aligned}
E_{jx_h^\ell} &= \frac{\partial P_{ij}}{\partial x_h^\ell} \frac{x_h^\ell}{P_{ij}} \\
&= \frac{\partial}{\partial x_h^\ell} \left(\int \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}} g(\omega) d\omega \right) \frac{x_h^\ell}{P_{ij}} \\
&= \int \frac{\partial}{\partial x_h^\ell} \left(\frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}} \right) g(\omega) d\omega \frac{x_h^\ell}{P_{ij}} \\
\text{when } j \neq h: &= \int -\frac{e^{V_{ij}} \frac{\partial V_{ih}}{\partial x_h^\ell} e^{V_{ih}}}{\left(\sum_{k=1}^{J} e^{V_{ik}} \right)^2} g(\omega) d\omega \frac{x_h^\ell}{P_{ij}} \\
&= -\frac{x_h^\ell}{P_{ij}} \int \frac{\partial V_{ih}}{\partial x_h^\ell} L_{ij}(\omega) L_{ih}(\omega) g(\omega) d\omega \\
\text{when } j = h: &= \int \frac{\left(\sum_{k=1}^{J} e^{V_{ik}} \right) \frac{\partial V_{ij}}{\partial x_h^\ell} e^{V_{ij}} - e^{V_{ij}} \frac{\partial V_{ij}}{\partial x_h^\ell} e^{V_{ij}}}{\left(\sum_{k=1}^{J} e^{V_{ik}} \right)^2} g(\omega) d\omega \frac{x_h^\ell}{P_{ij}} \\
&= \frac{x_h^\ell}{P_{ij}} \int \frac{\partial V_{ij}}{\partial x_h^\ell} L_{ij}(\omega)(1 - L_{ij}(\omega)) g(\omega) d\omega
\end{aligned}$$

When $V_{ij} = \omega' x_{ij}$, $\frac{\partial V_{ij}}{\partial x_h^\ell} = \omega_h$ in the above. This elasticity is different for each alternative j. A 10% reduction for one alternative need not imply (as with logit) a 10% reduction for all other alternatives. Note that the percent change in probability depends on the correlation between $L_{ij}(\omega)$ and $L_{ih}(\omega)$, which is determined by the researchers specification of $g(\omega)$. For example, to represent a situation where an improvement in alternative j draws more proportionately from alternative i than alternative k, the researcher can specify an element of x that is positively correlated between i and j but uncorrelated or negatively correlated between k and j, with a mixing distribution that allows the coefficient of this variable to vary.

[1] There are 2 sets of parameters in the mixed logit model: the ω's which enter the logit formula and the parameters of the mixing distribution $g(\omega)$. Typically we care about estimating the latter.

4.3 Random coefficients

The mixed logit probability can be derived from utility-maximising behaviour in several ways that are are formally equivalent, but provide different interpretations. One interpretation is based on random coefficients, the decision-maker faces a choice among J alternatives and has a utility function of the form:

$$U_{ij} = \omega_i' x_{ij} + \varepsilon_{ij}$$

where ω_i now varies across decision-makers in the population with density $g(\omega)$. This density is a function of parameters θ that represent, for example, the mean and variance of the ω's in population. This specification is the same as for standard logit, except that ω is no longer fixed but varies across decision-makers. Of course the decision maker knows the value of their ω_i and ε_{ij}'s for all j, but the researcher only observes the x_{ij}'s. If the decision maker observed the ω_i's, then the model would be a standard logit. Thus the choice probability conditional on ω_i is the standard logit probability:

$$L_i j(\omega_i) = \frac{e^{\omega_i' x_{ij}}}{\sum_{k=1}^{J} e^{\omega_i' x_{ik}}}$$

However the researcher does not observe ω_i so cannot condition on it. To find the unconditional choice probability we integrate over the density of ω_i:

$$P_{ij} = \int L_{ij}(\omega) g(\omega) d\omega$$

which is exactly the mixed logit probability from equation 4.1.

Note:-

The researcher imposes $g(\omega)$ as before, and estimates its parameters θ. Typically we specify $g(\omega)$ as normal or lognormal: $\omega \sim N(\bar{\omega}, \Sigma_\omega)$ or $\ln \omega \sim N(\bar{\omega}, \Sigma_\omega)$. The researcher then estimates $\bar{\omega}$ and Σ_ω. The lognormal distribution is useful when the coefficient is known to have the same sign for every decision-maker, such as price that is known to be negative for everyone.

4.4 Error-components

Another interpretation of mixed logit is based on error-components that create correlations among the utilities for different alternatives. Utility is specified as:

$$U_i j = \omega' x_{ij} + \mu_i' z_{ij} + \varepsilon_{ij}$$

where x_{ij} and z_{ij} are vectors of observed variables relating to alternative j, ω represents fixed coefficients and μ_i represents a vector of random terms with zero mean. That is, the unobserved (random) portion of utility is $\eta_{ij} = \mu_i' z_{ij} + \varepsilon_{ij}$ which can be correlated across alternatives depending on the specification of z_{ij} [2].
With non-zero error-components, utility is correlated over alternatives:

$$Cov(\eta_{ij}, \eta_{ik}) = \mathbb{E}(\mu_i' z_{ij} + \varepsilon_{ij})(\mu_i' z_{ik} + \varepsilon_{ik}) = z_{ij}' \Sigma_\mu z_{ik}$$

where Σ_μ is the covariance of μ_i. Utility is correlated over alternatives even when the error components are independent such that W is diagonal.

Example. An analog to nested logit is obtained by specifying a dummy for each nest that equals for each alternative in the nest, and zero otherwise. With K non-overlapping nests, the error components are $\mu_i' z_{ij} = \sum_{k=1}^{K} \mu_{ik} d_{kj}$ where $d_{kj} = -1$ if alternative j is in nest k. For convenience let $\mu_{ik} \sim NID(0, \sigma_k^2)$. The random term μ_{ik} enters the utility of each nest,

[2] In the standard logit model $z_{ij} = 0$ for all i and j, such that there is no correlation in utility across alternatives. This lack of correlation is what results in IIA.

inducing correlation among these alternatives. It does not enter any od the alternatives in other nests, so there is no correlation across nests. The variance σ_k^2 captures the magnitude of the correlation within each nest.
The correlation within nests is given by:

$$\begin{aligned} Corr(\eta_{ij}, \eta_{ik}) &= \frac{Cov(\eta_{ij}, \eta_{ik})}{\sqrt{Var(\eta_{ij})Var(\eta_{ik})}} \\ &= \frac{Cov(\mu_k + \varepsilon_{ij}, \mu_k + \varepsilon_{ik})}{\sqrt{Var(\mu_k + \varepsilon_{ij})Var(\mu_k + \varepsilon_{ik})}} \\ &= \frac{Var(\mu_k)}{Var(\mu_k) + \sigma_\varepsilon^2} \\ &= \frac{\sigma_k^2}{\sigma_k^2 + \frac{\pi^2}{6}} \end{aligned}$$

Comparing error-components and random coefficients
Both specifications are formally equivalent, but provide different interpretations.
Random coefficients → *Error-components*: Random-coefficients can be written as $U_ij = \bar{\omega}' x_{ij} + \tilde{\omega}_i' x_{ij} + \varepsilon_{ij}$ where $\bar{\omega}$ is the mean of ω_i and $\tilde{\omega}_i$ the deviation. We can see this has error components defined by $z_{ij} = x_{ij}$ with $\mu_i = \tilde{\omega}_i$.
Error-components → *Random coefficients*: Error-components is denoted by $U_ij = \omega' x_{ij} + \mu_i' z_{ij} + \varepsilon_{ij}$, which is equivalent to random coefficients with fixed coefficients for variables x_{ij} and random coefficients with zero means for variables z_{ij}.

4.5 Variability in population

How can we learn about what proportion of the population have positive preferences for a particular attribute? I.e. given estimates of $\bar{\omega}$ and Σ_ω, how can we compute $Pr(\omega_j > 0)$?
Suppose $\omega \sim N(\bar{\omega}, \Sigma_\omega)$, then $\omega_j \sim N(\mathbb{E}[\omega_j], Var(\omega_j))$. We can replace $\mathbb{E}[\omega_j]$ and $Var(\omega_j)$ with their estimates $\hat{\omega}_j$ and $\hat{\sigma}_j^2$, then compute $Pr(\omega_j > 0)$ using the standard normal distribution:

$$\begin{aligned} z_j = \frac{\omega_j - \mathbb{E}[\omega_j]}{\sqrt{Var(\omega_j)}} &\sim N(0,1) \\ Pr(\omega_j > 0) &= Pr\left(\frac{\omega_j - \mathbb{E}[\omega_j]}{\sqrt{Var(\omega_j)}} > -\frac{\mathbb{E}[\omega_j]}{\sqrt{Var(\omega_j)}} \right) \\ &= Pr\left(z_j > -\frac{\mathbb{E}[\omega_j]}{\sqrt{Var(\omega_j)}} \right) \\ &= Pr\left(z_j > -\frac{\hat{\omega}_j}{\hat{\sigma}_j} \right) \\ &= \Phi\left(\frac{\hat{\omega}_j}{\hat{\sigma}_j} \right) \end{aligned}$$

5 Panel Data Models

5.1 Panel Data Setup

A general representation of a linear panel data model is given by:

$$y_{it} = \alpha_i + \beta_{1i}' x_{1it} + \beta_{2i}' x_{2i} + \beta_{3i}' x_{3t} + \varepsilon_{it} \quad \text{for } i = 1, \ldots, N \text{ and } t = 1, \ldots, T$$

α_i : individual specific effect, E.G. unobserved institutional effects in a cross country growth model

x_{1it} : $k_1 \times 1$ vector of measures varying over T and N, e.g. individual income

x_{2i} : $k_2 \times 1$ vector of measures varying over N, e.g. individual education

x_{3t} : $k_3 \times 1$ vector of measures varying over T, e.g. interest rates

Note that x_{2i} is distinct from α_i, the former represents *observed heterogeneity* across individuals, while the latter attempts to partially control for *unobserved heterogeneity* across individuals.

We can impose that the parameters β are identical across individuals to get the *homogenous slope panel model*, which we focus on for the rest of the lecture.

$$y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it}, \quad Var(\varepsilon_{it}) = \sigma_\varepsilon^2$$

$$x_{it} = \begin{bmatrix} x_{1it} \\ x_{2i} \\ x_{3t} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Or in matrix form:

$$y_i = \alpha_i \mathbf{1}_T + X_i \beta + \varepsilon_i$$

Definition 5.1.1: Strict exogeneity

$$\mathbb{E}[\varepsilon_{it} | x_{i1}, x_{i2}, \ldots, x_{iT}, \alpha_i] = \mathbb{E}[\varepsilon_{it}] = 0$$

This implies $Cov(x_{it}, \varepsilon_{is}) = 0 \quad \forall i, t, s$. Intuitively: every error is uncorrelated with every regressor at every time period.

Example (A violation of strict exogeneity). Consider a model where ε_{it} is correlated with future values of the regressors, e.g.

$$x_{i,t+1} = \varphi \varepsilon_{it} + e_{it+1}$$

A production function in which labour demand in period $t+1$ responds to unobserved productivity shocks in period t, represented by ε_{it}, would be an example of this.

Definition 5.1.2: Weak exogeneity

$$\mathbb{E}[\varepsilon_{it} | x_{i1}, x_{i2}, \cdots, x_{it}, \alpha_i] = 0$$

This implies $Cov(x_{it}, \varepsilon_{is}) = 0 \quad \forall s \geq t$. Intuitively: every error is uncorrelated with every regressor at or before the current time period.

Example (Programme Evaluation and Wages). Many studies have considered the impact of job training programmes on wages. A standard specification is:

$$\log(wage_{it}) = \alpha_i + x'_{it}\beta + \delta prog_{it} + \varepsilon_{it}$$

Consider a 2 period model where at t=1 nobody is treated and at t=2 a group of individuals enter the program, and wages are observed for both treated and untreated individuals. Do we have strict exogeneity?
Is there correlation between $prog_{it}$ and ε_{it} conditional on α_i? We might think not, an individual cannot react to wage shock and enlist in program in the same period.
Is there correlation between $prog_{it+1}$ and ε_{it}? This might be non-zero if future participation is based on current shocks to wages.

Note:-

A general dynamic panel model might be written as:

$$y_{it} = \tau y_{i,t-1} + \beta' x_{it} + \alpha_i + \mu_t + \varepsilon_{it}$$

where μ_t is a time specific effect. We have an identification problem here, since $y_{i,t}$ depends on the fixed term α_i, so does $y_{i,t-1}$ which implies non-zero correlation between $y_{i,t-1}$ and the error $\alpha_i + \mu_t + \varepsilon_{it}$. This is true even if ε_{it} is serially uncorrelated (but it is clear to see that the correlation grows with serial correlation in ε_{it}). To estimate this model we need to use *instrumental variables* or *GMM*.

Define the composite error $v_{it} = \alpha_i + \varepsilon_{it}$. We are thus interested in estimation of β in the model $y_{it} = x'_{it}\beta + v_{it}$. We consider three approaches to estimation under this setup: *pooled OLS*, *fixed effects*, and *random effects*.

5.2 Pooled OLS (POLS)

Definition 5.2.1

The pooled OLS estimator is given by:

$$\hat{\beta}_{POLS} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}x'_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}y_{it}\right) = \left(\sum_{i=1}^{N} X'_i X_i\right)^{-1}\left(\sum_{i=1}^{N} X'_i y_i\right)$$

where $\bar{x} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}$ and $\bar{y} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}$.
Assumptions:

$$\begin{aligned} &\text{(POLS1a)} \quad \mathbb{E}[x'_{it}\alpha_i] = 0 \quad \forall t \\ &\text{(POLS1b)} \quad \mathbb{E}[x'_{it}\varepsilon_{it}] = 0 \quad \forall t \\ &\text{(POLS2)} \quad \operatorname{rank}(\mathbb{E}[x_{it}x'_{it}]) = k \quad \forall i \end{aligned}$$

(POLS1) is equivalent to the standard exogeneity assumption, but on the composite error: $\mathbb{E}[x'_{it}v_{it}]$. POLS is consistent under (POLS1) and (POLS2), however the composite errors will be serially correlated on account of the common α_i term. Indeed this correlation does not decrease as $|t-s|$ increases, therefore inference using POLS requires robust standard errors.

5.3 Fixed Effects

In many cases we might expect the unobserved effect α_i to be correlated with the observed x_{it}, in which case POLS is inconsistent. We can instead use fixed effects in this case. Consider the model in matrix form from above:

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix} = \begin{bmatrix} 1_T & 0 & \dots & 0 \\ 0 & 1_T & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1_T \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} + \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

The *Least Squares Dummy Variable* (LSDV) estimator can be computed by running OLS on the stacked equations, giving N fixed effects. However, as N gets large, the inversion of a matrix of dimension $(N+k)\times(N+k)$ becomes computationally infeasible. The Fixed effects transformation by contrast only requires the inversion of a $k \times k$ matrix.

Definition 5.3.1

The fixed effects estimator is given by:

$$\begin{aligned} \hat{\beta}_{FE} &= \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) \right) \\ &= \left(\sum_{i=1}^{N} (Q_T X_i)'(Q_T X_i) \right)^{-1} \left(\sum_{i=1}^{N} (Q_T X_i)'(Q_T y_i) \right) \end{aligned}$$

where $Q_T = I_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T'$ and $\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}$.
Assumptions:

$$\begin{aligned} &\text{(FE1)} \quad \mathbb{E}[\varepsilon_{it}|X_i, \alpha_i] = 0 \quad \forall t \\ &\text{(FE2)} \quad \text{rank}(\mathbb{E}[X_i' Q_T X_i]) = k \\ &\text{(FE3)} \quad \mathbb{E}[\varepsilon_i \varepsilon_i'|X_i, \alpha_i] = \sigma_\varepsilon^2 I_T \end{aligned}$$

FE1:
Strict exogeneity conditional on the unobserved effects, however note that $\mathbb{E}[\alpha_i|X_i]$ is allowed to be any function of X_i. The cost of this assumption is that we lose the ability to estimate individual specific effects, as there is no way to distinguish between the time invariant unobserved an observed effects.
FE2:
If x_{it} contains an element that is constant over time for any i, then the corresponding column of $X_i' Q_T X_i$ will be zero, and the rank condition will not hold.
FE3:
Constant conditional variance and zero conditional covariance of the idiosyncratic error.

The matrix Q_T is an idempotent matrix with all off-diagonal elements equal to $-\frac{1}{T}$ and all diagonal elements equal to $1 - \frac{1}{T}$. $Q_T y_i$ simply subtracts individual means from y_i.
The FE for β is consistent (and unbiased) under (FE1) and (FE2) with both $N \to \infty$ and $T \to \infty$. Estimates of the fixed effects can be obtained using $\hat{\alpha}_i = \bar{y}_i - \bar{x}_i' \hat{\beta}_{FE}$. The FE estimate of α_i is unbiased and consistent for fixed N and $T \to \infty$, however we have an incidental parameters problem with increasing N as the number of fixed effects grows with N.

5.4 Random Effects

If we assume $\alpha_i \sim iid(0, \sigma_\alpha^2)$ independent of $\varepsilon_{it} \sim iid(0, \sigma_\varepsilon^2)$ and uncorrelated with x_{it}, we have a random effects or one-way error-components setup. We effectively put α_i into the error term, under the assumption that it is orthogonal to x_{it}, then accounts for the implied serial correlation with GLS.

Definition 5.4.1

The random effects estimator is a FGLS estimator, given by

$$\hat{\beta}_{RE} = \left(\sum_{i=1}^{N} X_i' \hat{\Omega}^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{N} X_i' \hat{\Omega}^{-1} y_i \right)$$

Assumptions:

$$\begin{aligned}
&\text{(RE1a)} \quad \mathbb{E}[\varepsilon_{it}|X_i, \alpha_i] = 0 \quad \forall t \\
&\text{(RE1b)} \quad \mathbb{E}[\alpha_i|X_i] = \mathbb{E}[\alpha_i] = 0 \quad \forall t \\
&\text{(RE2)} \quad rank(\mathbb{E}[X_i' \Sigma^{-1} X_i]) = k \\
&\text{(RE3a)} \quad \mathbb{E}[\varepsilon_i \varepsilon_i'|X_i, \alpha_i] = \sigma_\varepsilon^2 I_T \\
&\text{(RE3b)} \quad \mathbb{E}[\alpha_i^2|X_i] = \sigma_\alpha^2
\end{aligned}$$

RE1:
(RE1a) is the strict exogeneity assumption(**exactly the same as FE1**), while (RE1b) is the orthogonality assumption, this is implied by the assumption that x_{it} are fixed and $\mathbb{E}[\alpha_i|X_i] = 0$ or that α_i is independent of X_i. The important part of this assumption is $\mathbb{E}[\alpha_i|X_i] = \mathbb{E}[\alpha_i]$, that it equals zero is without loss of generality provided an intercept is included in x_{it}.
Why do we make (RE1), which is much stronger than (POLS1)? Random effects exploits serial correlation in the composite error in a GLS framework, to ensure it's consistent we need some form of strict exogeneity between the composite error and the regressors.
RE2:
For consistency we need the usual rank condition for GLS, we know GLS and FGLS are consistent under RE1 and RE2. However we have not yet exploited the unobserved effects structure of v_{it}.
RE3:
(RE3a) is constant conditional variance and zero conditional covariance of the idiosyncratic error (**exactly the same as FE3**), this is stronger than constant unconditional variance. I.e.: $\mathbb{E}[\varepsilon_{it}^2|X_i, \alpha_i] = \sigma_\varepsilon^2 \quad \forall t$ and $\mathbb{E}[\varepsilon_{it}\varepsilon_{is}|X_i, \alpha_i] = 0 \quad \forall t \neq s$.
(RE3b) is a homoskedasticity assumption on the unobserved effect. Taken together these imply the unconditional assumptions:

$$\begin{aligned}
&\mathbb{E}[\alpha_i|X_i] = 0 && \mathbb{E}[\alpha_i^2|X_i] = \sigma_\alpha^2 && \mathbb{E}[\alpha_i\alpha_j|X_i] = 0 && \forall i \neq j \\
&\mathbb{E}[\varepsilon_{it}|X_i] = 0 && \mathbb{E}[\varepsilon_{it}^2|X_i] = \sigma_\varepsilon^2 && \mathbb{E}[\varepsilon_{it}\varepsilon_{js}|X_i] = 0 && \forall i \neq j, t \neq s \\
&\mathbb{E}[\alpha_i\varepsilon_{jt}|X_i] = 0 && \forall i, j, t
\end{aligned}$$

We can now write the unconditional covariance matrix of v_{it} as:

$$\begin{aligned}\mathbb{E}[v_i v_i'] &= \mathbb{E}[(\alpha_i + \varepsilon_i)(\alpha_i + \varepsilon_i)'] \\ &= \mathbb{E}[\alpha_i \alpha_i'] + \mathbb{E}[\alpha_i \varepsilon_i'] + \mathbb{E}[\varepsilon_i \alpha_i'] + \mathbb{E}[\varepsilon_i \varepsilon_i'] \\ &= \sigma_\alpha^2 \mathbf{1}_T \mathbf{1}_T' + \sigma_\varepsilon^2 I_T \\ &= \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\varepsilon^2 + \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\varepsilon^2 + \sigma_\alpha^2 \end{bmatrix}\end{aligned}$$

Rather than depending on $T(T+1)/2$ unrestricted covariances (as in normal GLS), Σ only depends on 2 parameters, regardless of T. The correlation between composite errors does not depend on the size of the difference between t and s, and is given by $\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$.
$\mathbb{E}[v_i v_i']$ is written in terms of single individual, we can write the $NT \times NT$ covariance matrix for all individuals as:

$$\Omega = I_n \otimes \Sigma = \begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & \Sigma & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Sigma \end{bmatrix}$$

We can write this as block diagonal due to random sampling in the cross section (i.e. observations i and j are independent). This is convenient as we can invert Ω by focusing on inverting Σ.

Lemma 5.4.1 (Inverse of a block diagonal matrix)**.**

$$\begin{bmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_n \end{bmatrix}^{-1} = \begin{bmatrix} A_1^{-1} & 0 & \dots & 0 \\ 0 & A_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_n^{-1} \end{bmatrix}$$

Observe that by transforming the regression by premultiplying with $\Sigma^{-1/2}$, we can see $Var(\Sigma^{-1/2} v_i) = I_T$ and $\mathbb{E}[v_i v_j] = 0$ for $i \neq j$.

Lemma 5.4.2 (Matrix inversion lemma)**.** If a and b are nonzero scalars and P and Q are symmetric idempotent matrices such that $PQ = QP = 0$ and $P + Q = I$, then:

$$(aP + bQ)^{-1} = \frac{1}{a} P + \frac{1}{b} Q$$

Theorem 5.4.1. The GLS transformation

$$\Sigma^{-1/2} y_i = \Sigma^{-1/2} \mathbf{1_T} \lambda + \Sigma^{-1/2} X_i \beta + \Sigma^{-1/2} v_i$$

takes the form of a quasi-demeaning transformation, corresponding to an OLS regression of $y_{it}^* \equiv y_{it} - \theta \bar{y}_i$ on $x_{it}^* \equiv x_{it} - \theta \bar{x}_i$ where $\theta = 1 - \frac{\sigma_\alpha}{\sqrt{T\sigma_\varepsilon^2 + \sigma_\varepsilon^2}}$.

Proof. We know:

$$\Sigma = \sigma_\alpha^2 \mathbf{1}_T \mathbf{1}_T' + \sigma_\varepsilon^2 I_T$$

Define $P = \frac{\mathbf{1}_T \mathbf{1}_T'}{T}$ and $Q = I_T - \frac{\mathbf{1}_T \mathbf{1}_T'}{T}$. Clearly these are the residual maker and projection matrices onto the space spanned by the constant vector, and thus have all the required properties to apply the lemma.

$$\begin{aligned}
\Sigma &= \sigma_\alpha^2 \mathbf{1}_T \mathbf{1}_T' + \sigma_\varepsilon^2 I_T \\
&= T\sigma_\alpha^2 P + \sigma_\varepsilon^2 (P + Q) \\
&= (T\sigma_\alpha^2 + \sigma_\varepsilon^2) P + \sigma_\varepsilon^2 Q \\
\Rightarrow \Sigma^{-1} &= \frac{1}{T\sigma_\alpha^2 + \sigma_\varepsilon^2} P + \frac{1}{\sigma_\varepsilon^2} Q \\
&= \frac{1}{T\sigma_\alpha^2 + \sigma_\varepsilon^2} \left(P + \psi^{-1} Q\right) \quad \text{where } \psi = \frac{\sigma_\varepsilon^2}{T\sigma_\alpha^2 + \sigma_\varepsilon^2} \\
\text{note that } (P + \psi^{-1/2} Q)^2 &= P + \phi^{-1/2} PQ + \phi^{-1/2} QP + \phi^{-1} Q \\
&= P + \psi^{-1} Q \quad \text{since } PQ = QP = 0 \\
\text{thus } \Sigma^{-1/2} &= \frac{1}{\sqrt{T\sigma_\alpha^2 + \sigma_\varepsilon^2}} \left(P + \psi^{-1/2} Q\right) \\
&= \frac{1}{\sigma_\varepsilon} \psi^{1/2} \left(\psi^{-1/2} I_T + (1 - \psi^{-1/2}) P\right) \\
&= \frac{1}{\sigma_\varepsilon} \left(I_T - (1 - \psi^{1/2}) P\right)
\end{aligned}$$

Define $\theta = 1 - \frac{\sigma_\alpha}{\sqrt{T\sigma_\varepsilon^2 + \sigma_\varepsilon^2}}$. Note that P is a $T \times T$ matrix with every element equal to $\frac{1}{T}$. We apply the transformation to y_i:

$$\begin{aligned}
\Sigma^{-1/2} y_i &= \frac{1}{\sigma_\varepsilon} \left(I_T - \theta P\right) y_i \\
&= \frac{1}{\sigma_\varepsilon} \begin{bmatrix} 1 - \frac{\theta}{T} & -\frac{\theta}{T} & \dots & -\frac{\theta}{T} \\ -\frac{\theta}{T} & 1 - \frac{\theta}{T} & \dots & -\frac{\theta}{T} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\theta}{T} & -\frac{\theta}{T} & \dots & 1 - \frac{\theta}{T} \end{bmatrix} \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix} \\
&= \frac{1}{\sigma_\varepsilon} \begin{bmatrix} y_{i1} - \frac{\theta}{T} \sum_{t=1}^T y_{it} \\ y_{i2} - \frac{\theta}{T} \sum_{t=1}^T y_{it} \\ \vdots \\ y_{iT} - \frac{\theta}{T} \sum_{t=1}^T y_{it} \end{bmatrix} \\
&= \frac{1}{\sigma_\varepsilon} \begin{bmatrix} y_{i1} - \theta \bar{y}_i \\ y_{i2} - \theta \bar{y}_i \\ \vdots \\ y_{iT} - \theta \bar{y}_i \end{bmatrix}
\end{aligned}$$

□

Note that the GLS estimator itself is not implementable since σ_α^2 and σ_ε^2 are unknown. We instead estimate $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\varepsilon^2$ and perform FGLS. Note that this estimator may be inefficient for a number

of reasons, for example ε_{it} may be heteroskedastic over time or serially correlated. To combat this (and for fixed T and large N), a robust covariance matrix should be used.

Explanation. ***The random effects (equicorrelated) panel model may be viewed as a Pooled model, with the random effect subsumed within the error term***
In the random effects set up the composite error is independent of the regressors, so we could indeed run POLS with the composite error. Random effects performs a GLS transformation to account for the specific form of correlation in the composite error, and is subsequently efficient. □

Explanation. ***How might we generalise the FGLS estimator which follows from RE1-RE3?***
If the idiosyncratic errors are generally heteroskedastic and serially correlated across t, a more general estimator of Ω could be used:

$$\hat{\Omega} = N^{-1} \sum_{i=1}^{N} \hat{v}_i \hat{v}_i'$$

where $\hat{v}_i$ is the POLS residual. We would not use this normally since it has $T(T+1)/2$ estimated elements (v.s. 2 for RE), however if N is sufficiently large we can use this to perform FGLS instead. □

Explanation. ***Inference should be based on panel-robust standard errors (SE) that permit errors to be correlated over time for a given individual and allow variances and covariances to differ across i.***
Default SE assumes independence of model errors over t for a given i; this overestimates the benefit of an additional T resulting in downward biased SEs. In reality there is correlation across time for a given i so an extra T is less informative than if it was independent.
Ignoring heteroskedasticity in errors also leads to biased SEs, though this could be in either direction. □

5.4.1 RE as a nested model

We can see how the random effects estimator nests both the POLS and FE estimators by examining the limits of θ:
POLS: $\theta = 0$
When $\psi = 1$ (because $\sigma_\alpha = 0$) there is no error covariance across observations for a given i, so GLS reduces to OLS. This is because the fixed effects are absorbed into the constant term (since they are the same for all individuals), and the error term is just the idiosyncratic error.
FE: $\theta = 1$
When $\psi = 0$ (because σ_α or $T \to \infty$) GLS converges to FE as we give more weight to within individual sample means, and quasi-demeaning becomes full demeaning.
Case 1 $\sigma_\alpha \to \infty$: The effect of time-constant explanatory variables becomes harder to estimate due to noise in the unobserved component.
Case 2 $T \to \infty$: The unobserved α_i becomes observable, we know our estimator of $[\alpha, \beta]$ is consistent in T or n, thus:

$$y_{it} - x_{it}'\beta = \alpha_i + \varepsilon_{it}$$

becomes observable. The individual means give:

$$\bar{y}_i - \bar{x}_i'\beta = \alpha_i + \bar{\varepsilon}_i$$

and by LLN $\bar{\varepsilon}_i \to 0$, revealing α_i.

5.5 Comparing estimators

A useful way of comparing estimators is by examining how they use variation within and between individuals. We can decompose the total variation of a series as[1]:

$$\begin{aligned}
T_{yy} &= \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\bar{y})^2 \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\bar{y}_i+\bar{y}_i-\bar{y})^2 \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\bar{y}_i)^2 + \sum_{i=1}^{N}\sum_{t=1}^{T}(\bar{y}_i-\bar{y})^2 + 2\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\bar{y}_i)(\bar{y}_i-\bar{y}) \\
&= \underbrace{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\bar{y}_i)^2}_{\text{within}} + \underbrace{\sum_{i=1}^{N}\sum_{t=1}^{T}(\bar{y}_i-\bar{y})^2}_{\text{between}}
\end{aligned}$$

POLS: Minimises the total sum of squares, weighting within and between variation equally.
FE: Disregards between variation, only uses variation within individuals.
RE: Uses both within and between variation, but never weighting between more than within.

We can see a bias-variance tradeoff here, we can completely drop between variation (resulting in a higher variance estimate) to get unbiasedness through dropping α_i as in fixed effects, or we can use both within and between variation to get a lower variance estimate, but potentially with a bias from the unobserved effects. We can think of RE as optimally picking these weights under certain assumptions on the structure of the unobserved effects.
The fundamental differences come from how we treat the unobserved effects, as shown below.

	Constant $\alpha_i = \alpha$	Random $\alpha_i \sim (\alpha, \sigma_\alpha^2)$	Fixed $\mathbb{E}[\alpha_i \mid X_i] \neq 0$
POLS	**BLUE** **Consistent**	**Unbiased** **Consistent** **Inefficient**	**Biased** **Inconsistent**
FE	**Unbiased** **Consistent** **Inefficient**	**Unbiased** **Consistent** **Inefficient**	**BLUE** **Consistent**
RE	**Unbiased** **Consistent** **Efficient(?)**	**BLUE** **Consistent**	**Biased** **Inconsistent**

5.5.1 Hausman test

We can test whether the random effects specification is supported by the data. The basis for this is that RE is only consistent under the assumption that α_i is uncorrelated with the regressors,

[1] the final equality follows from $\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\bar{y}_i)(\bar{y}_i-\bar{y}) = \sum_{i=1}^{N}\left((\bar{y}_i-\bar{y})\sum_{t=1}^{T}(y_{it}-\bar{y}_i)\right) = \sum_{i=1}^{N}(\bar{y}_i-\bar{y})(T\bar{y}_i - T\bar{y}_i) = 0$

whereas FE is always consistent regardless of the conditional distribution of α_i. We thus test on the difference between the RE and FE estimators, which should be zero if RE is correctly specified. We want to test: $H_0 : \mathbb{E}[\alpha_i|X_i] = 0$ against $H_1 : \mathbb{E}[\alpha_i|X_i] \neq 0$. We can summarise our estimators:

Estimator	Under H_0	Under H_1
POLS	**Consistent**	**Inconsistent**
FE	**Consistent**	**Consistent**
RE	**Efficient**	**Inconsistent**
Between	**Consistent**	**Inconsistent**

Define :

$$\hat{\kappa} = \hat{\beta}_{RE_s} - \hat{\beta}_{FE_s}$$

Where $\hat{\beta}_{RE_s}$ is the subcomponent of $\hat{\beta}_{RE}$ relating to time-varying regressors. Under H_0, $\hat{\kappa} \xrightarrow{p} 0$ and $Cov(\hat{\kappa}, \hat{\beta}_{RE_s}) = 0$ (???), thus $Var(\hat{\kappa}) = Var(\hat{\beta}_{FE_s}) - Var(\hat{\beta}_{RE_s})$ (???). The Hausman test statistic is then:

$$H = \hat{\kappa}' \left(Var(\hat{\kappa})\right)^{-1} \hat{\kappa} \sim \chi^2(k_s)$$

where k_s is the number of time-varying regressors.

Note:-

Problems with the Hausman test

- This form of the Hausman test is invalid if α_i or $\varepsilon_i t$ are not i.i.d.
- There is no guarantee that Var($\hat{\kappa}$) is positive definite.

A moment based test of the RE model can be used given that the model is over-identified. We know FE is consistent under RE1 (since RE1a = FE1) but we can test whether the extra moment condition from RE1b is satisfied by an overid test.

Example. Suppose $\hat{\beta}_{FE} = 0.168$ and $\hat{\beta}_{RE} = 0.119$ with standard deviations 0.019 and 0.014 respectively, is there evidence of correlation between α_i and X_i?

$$H = \frac{(0.168 - 0.119)^2}{0.019^2 - 0.014^2} \approx 14 > \chi^2_{.05}(1) = 3.84$$

so the random effects model is rejected. The statistic H seems inflated because we have used default standard errors (which are greatly downward biased). This is a signal that a more general form of the Hausman test should be used.

5.6 Correlated Random Effects

We can extend the random effects model to allow for correlation between the unobserved effects and the regressors. Given random sampling in the cross section, an analyst starts with the following conditional mean assumption:

$$\mathbb{E}[\varepsilon_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}, \alpha_i] = 0 \quad \forall t$$

Further we now impose the additional moment condition[2]:

$$\mathbb{E}[\alpha_i|X_i] = \psi + \bar{x}_i'\delta$$

[2] we are now overidentified

where $\alpha_i = \psi + \bar{x}_i'\delta + \nu_i$ and $\mathbb{E}[\nu_i|X_i] = 0$. Note that previously we were only using $\alpha_i = \psi + \nu_i$, now we are allowing for a linear relationship between α_i and $\bar{x}_i$. **Thus when $\delta = 0$, we have the standard random effects model**.
Given this, we can find an expression for $\mathbb{E}[y_{it}|X_i]$:

$$\begin{aligned} y_{it} &= \alpha_i + \beta' x_{it} + \varepsilon_{it} \\ &= \psi + x_{it}'\beta + \bar{x}_i'\delta + \nu_i + \varepsilon_{it} \\ \Rightarrow \mathbb{E}[y_{it}|X_i] &= \psi + x_{it}'\beta + \bar{x}_i'\delta \end{aligned}$$

Assumptions:

$$\mathbb{E}[\nu_i|X_i] = 0 \quad \mathbb{E}[\varepsilon_{it}|X_i] = 0 \quad \forall t$$

This allows us to keep the modelling structure of RE (with α_i randomly distributed), but allows for some correlation between α_i and x_{it}. By adding $\bar{x}_i$ we control for some correlation between α_i and x_{it}, the remainder ν_i is uncorrelated with x_{it}.
Under the assumptions above, the estimating equation is given by:

$$\begin{aligned} y_{it} &= \psi + x_{it}'\beta + \bar{x}_i'\delta + \nu_i + \varepsilon_{it} \\ &= \psi + x_{it}'\beta + \bar{x}_i'\delta + u_{it} \end{aligned}$$

and we can just use POLS to consistently estimate all parameters, including δ. Further we can include time constant variables, for example if we begin with

$$y_{it} = \alpha_i + g_t'\theta + z_i'\zeta + x_{it}'\beta + \varepsilon_{it}$$

we can use the CRE estimating equation

$$y_{it} = g_t'\theta + z_i'\zeta + x_{it}'\beta + \psi + \bar{x}_i'\delta + u_{it}$$

which is algebraically equivalent to a fixed effects transformation! Thus we get the FE estimates of θ and β (the coefficients on the time varying covariates). This equivalence implies an interesting interpretation of the FE estimator: it controls for the average level $\bar{x}_i$ when measuring the partial effects of x_{it} on y_{it}. It would not be necessary to control for this average level of $\bar{x}_i$, as with RE, if $\delta = 0$. A test of $H_0 : \delta = 0$ is then a test of the RE specification.

6 Fundementals of Bayesian Inference

Classical approaches studies thus far mostly rely on distributions of estimators and test statistics over hypothetical repeated samples. There is no conditioning on the observed data.
In contrast, Bayesian inference is based on the posterior distribution of the parameter of interest, given the observed data. These distributions are exact in finite samples, distributions are derived conditional on the observed data.
Classical hypothesis testing measures support for the data, $Pr(D|H_0)$, while Bayesian measures the support for the hypothesis in the data $Pr(H_0|D)$.

6.1 Bayes Rule

Definition 6.1.1: Bayes Rule

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

where $P(A|B)$ is the likelihood, $P(B)$ is the prior, and $P(A)$ is the marginal likelihood. $P(B|A)$ is the posterior.

A prior probability is an initial value of the probability of an event, we update this with data to get the posterior probability.
We can write this for parameters of a model, let $p(\theta)$ be the prior density on some unknown parameter θ. Let $p(\theta|y)$ be the posterior density of θ. The probability of observing y conditional on θ is given by the likelihood:

Definition 6.1.2: Likelihood Function

$$f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

It is the conditional probability of observing the data given the parameter.

Bayes rule gives us

$$p(\theta|y) = \frac{f(y|\theta)p(\theta)}{f(y)}$$

However we note that the denominator is not a function of θ, so we can write

Definition 6.1.3: Posterior Distribution

$$\begin{aligned} p(\theta|y) &= \frac{f(y|\theta)p(\theta)}{f(y)} \\ &\propto f(y|\theta)p(\theta) \end{aligned}$$

6.2 Conjugate Priors

A class of priors is conjugate for a family of a likelihoods if both prior and posterior are in the same class for all data y.

Definition 6.2.1

If τ is a class of sampling distributions $p(y|\theta)$, and ω is a class of prior distributions for θ, then ω is conjugate for τ if:

$$p(\theta|y) \in \omega \quad \forall p(y|\theta) \in \tau, p(\theta) \in \omega$$

Intuitively what this means is that when we get some data, updating only involves updating the parameters of the distribution, not changing the distribution itself.

Example (Bernoulli distribution and Beta priors). Suppose we have a Bernoulli pdf we can write the likelihood in terms of the mean parameter as:

$$p(y|\theta) = \theta^y(1-\theta)^{1-y}$$

This suggests that a conjugate prior might be given by

$$p(\theta|\tau) \propto \theta^{\tau_1}(1-\theta)^{\tau_2}$$

This expression can be normalised if both τ's are greater than -1, and we get the beta distribution:

$$p(\theta|\alpha,\beta) = K(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where the normalisation constant can be found by solving:

$$1 = \int K(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta$$
$$\Rightarrow K(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

If we multiply the beta density by the Bernoulli likelihood we obtain a beta density. Consider N *i.i.d.* Bernoulli trials, then the likelihood is:

$$\begin{aligned} p(\theta|y,\alpha,\beta) &\propto \left(\prod_{i=1}^{N}\theta^{y_i}(1-\theta)^{1-y_i}\right)\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\sum_{i=1}^N y_i+\alpha-1}(1-\theta)^{N-\sum_{i=1}^N y_i+\beta-1} \\ &\sim \text{Beta}(\sum_{i=1}^{N} y_i+\alpha, N-\sum_{i=1}^{N} y_i+\beta) \end{aligned}$$

6.2.1 Beta distribution

Definition 6.2.2: Beta distribution

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

where Γ is the gamma function.

This is a super useful distribution, and it nests many other common distributions. For example, if $\alpha = \beta = 1$ we get the uniform distribution. Indeed any Beta with $\alpha = \beta$ is symmetric.
When we have $\alpha > \beta$ ($\alpha < \beta$) the distribution is skewed to the left (right).

6.2.2 Examples

Definition 6.2.3: Diriclet distribution

$$\mathrm{f(x;}\ \alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}x_i^{\alpha_i-1}$$

Question 1

Show that the Diriclet distribution is the conjugate prior for the multinomial distribution.

Solution:-

The likelihood for the multinomial distribution is given by:

$$p(y|\theta) = \theta_1^{\sum_{i=1}^{N}\mathbf{1}(y_i=1)}\theta_2^{\sum_{i=1}^{N}\mathbf{1}(y_i=2)}\dots\theta_k^{\sum_{i=1}^{N}\mathbf{1}(y_i=k)}$$

We write the Diriclet prior as:

$$p(\theta|\alpha) = K(\alpha)\prod_{i=1}^{k}\theta_i^{\alpha_i-1}$$

with

$$K(\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}$$

As before we can derive:

$$p(\theta|y,\alpha) \propto \left(\prod_{i=1}^{k}\theta_i^{\sum_{i=1}^{N}\mathbf{1}(y_i=i)}\right)\left(\prod_{i=1}^{k}\theta_i^{\alpha_i-1}\right) = \prod_{i=1}^{k}\theta_i^{\sum_{i=1}^{N}\mathbf{1}(y_i=i)+\alpha_i-1}$$

Question 2

Show that the gamma distribution is the conjugate prior for the poisson distribution.

Solution:-

The likelihood for the poisson distribution is given by:

$$p(y|\theta) = \frac{e^{-\theta}\theta^y}{y!}$$

We write the gamma prior as:

$$p(\theta|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$$

As before we can derive:

$$\begin{aligned} p(\theta|y,\alpha,\beta) &= \left(\frac{e^{-\theta}\theta^y}{y!}\right)\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\right) \\ &\propto \theta^{\sum_{i=1}^{N}y_i+\alpha-1}e^{-\theta(N+\beta)} \\ &\sim \mathrm{Gamma}(\sum_{i=1}^{N}y_i+\alpha, N+\beta) \end{aligned}$$

6.2.3 Mean and variance of the posterior

Consider the bernoulli and beta example. The mean of a beta distribution is given by:

$$\mathbb{E}[\text{Beta}(\alpha, \beta)] = \frac{\alpha}{\alpha + \beta}$$

Thus the mean of the posterior is given by:

$$\begin{aligned}
\mathbb{E}[\theta|y] &= \frac{\sum_{i=1}^{N} y_i + \alpha}{N + \alpha + \beta} \\
&= \frac{N}{N + \alpha + \beta}\bar{y} + \frac{\alpha + \beta}{N + \alpha + \beta}\frac{\alpha}{\alpha + \beta} \\
&:= w\frac{\alpha}{\alpha + \beta} + (1 - w)\bar{y}
\end{aligned}$$

Clearly this is a weighted average of the prior mean and the sample mean. As $N \to \infty$ the weight on the prior mean goes to zero. That is, as $N \to \infty$ the mean of the posterior approaches the MLE estimate of θ.
We can also examine the variance, note that the variance of a beta distribution is

$$Var[\text{Beta}(\alpha, \beta)] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Thus the posterior variance is:

$$Var[\theta|y] = \frac{(\sum_{i=1}^{N} y_i + \alpha)(N - \sum_{i=1}^{N} y_i + \beta)}{(N + \alpha + \beta)^2(N + \alpha + \beta + 1)}$$

We can see that $Var[\theta|y, \alpha, \beta] \to 0$, showing that the posterior distribution concentrates around the MLE as $N \to \infty$.

Claim 6.2.1. The posterior mean is a weighted average for both the gamma prior with a poisson likelihood and the Diriclet prior with a multinomial likelihood.

Proof. **Gamma prior with poisson likelihood:** Note that the mean of the gamma distribution is:

$$\mathbb{E}[\text{Gamma}(\alpha, \beta)] = \frac{\alpha}{\beta}$$

Thus the mean of the posterior is given by:

$$\begin{aligned}
\mathbb{E}[\theta|y] &= \frac{\sum_{i=1}^{N} y_i + \alpha}{N + \beta} \\
&= \frac{N}{N + \beta}\bar{y} + \frac{\beta}{N + \beta}\frac{\alpha}{\beta} \\
&:= w\frac{\alpha}{\beta} + (1 - w)\bar{y} \quad \text{where } w = \frac{\beta}{N + \beta}
\end{aligned}$$

Diriclet prior with multinomial likelihood: Note that the mean of the Diriclet distribution is:

$$\mathbb{E}[\text{Diriclet}(\alpha)] = \frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i}$$

Thus the mean of the posterior is given by:

$$\begin{aligned}\mathbb{E}[\theta|y] &= \frac{\sum_{i=1}^{N} \mathbf{1}(y_i = i) + \alpha_i}{N + \sum_{i=1}^{k} \alpha_i} \\ &= \frac{N}{N + \sum_{i=1}^{k} \alpha_i}\bar{y} + \frac{\sum_{i=1}^{k} \alpha_i}{N + \sum_{i=1}^{k} \alpha_i}\frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i} \\ &:= w\frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i} + (1-w)\bar{y} \quad \text{where } w = \frac{\sum \alpha_i}{N + \sum \alpha_i}\end{aligned}$$

□

6.3 Exchangability

Skipped in 2024

6.4 Parameter Uncertainty

Suppose we want to make inference on a population parameter μ, from a classical perspective μ is fixed, and we construct a confidence interval through the sampling distribution of the random variable, say $\bar{y}$. We can make a statement about $\bar{y}$:

$$Pr(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{y} \leq \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

the pivot this around $\bar{y}$ (inverting the test) to get a confidence interval for μ:

$$Pr(\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Given that μ is fixed, the randomness comes from $\bar{y}$. The interpretation of this is that (1- α) of the time, the interval will contain μ.
Our confidence interval is thus

$$\bar{y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

By contrast a Bayesian makes a probability statement based on the posterior of the parameter μ give the actual data: $Pr(\mu|y)$.

Definition 6.4.1: Posterior Odds

The posterior odds of one model vs another is given by:

$$\underbrace{\frac{Pr(M_1|y)}{Pr(M_2|y)}}_{\text{Posterior Odds}} = \underbrace{\frac{Pr(y|M_1)}{Pr(y|M_2)}}_{\text{Bayes Factor}} \times \underbrace{\frac{Pr(M_1)}{Pr(M_2)}}_{\text{Prior Odds}}$$

We use the Bayes factor (the relative likelihood) to update the prior odds to get the posterior odds.

7 Hierarchical Models for Combining Data

Hierarchical refers to the situation where there may be a natural structure to the data, e.g.: Individuals within regions within countries. In exploiting this we can manage instances where the parameter space is large and we may wish to reduce the dimension through distributional assumptions. For example, when we considered random effects, α_i was potentially high dimensional, but the distributional assumption that $\alpha_i \sim N(0, \sigma_\alpha^2)$ allowed us to reduce the dimension of the parameter space.

7.1 Multilevel Data

We now have new interpretations of within and between variability:

- **Within variability** Variation of individual-level data around individual time means in a panel data model. Variation of individual-level data around village means in a two-level model.
- **Between variability** Variation of individual time means around the overall mean in a panel data model. Variation of village means around the overall mean in a two-level model.

Considering a population with J groups and n_j individuals per group:

$$\begin{aligned} \{y_1, \ldots, y_{n_j}\} &\overset{\text{iid}}{\sim} p(y_j|\theta_j) \quad \text{within group} \\ p(\theta_1, \ldots, \theta_J) &= \prod_{j=1}^{J} p(\theta_j|\phi) \quad \text{between groups} \\ \phi &\sim p(\phi) \quad \text{prior} \end{aligned}$$

Within group parameters $\theta_j = (\mu_j, \sigma^2)$, between group parameters $\phi = (\psi, \tau^2)$. $p(\theta_j|\phi)$ describes heterogeneity between group means, while $p(y_j|\theta_j)$ describes within group variability. We assume the within group sampling variability is constant across groups. We can use Gibbs Sampling[1] to approximate the posterior distribution

$$p(\mu_1, \ldots, \mu_J, \sigma^2, \psi, \tau^2|y_1, \ldots, y_J)$$

7.1.1 Posterior Distributions

We can write this posterior distribution using Bayes rule:

$$\begin{aligned} p(\mu_1, \ldots, \mu_J, \sigma^2, \psi, \tau^2|y_1, \ldots, y_J) \propto & \underbrace{p(y_1, \ldots, y_J|\mu_1, \ldots, \mu_J, \sigma^2, \psi, \tau^2)}_{\text{likelihood}} \\ & \times \underbrace{p(\mu_1, \ldots, \mu_J|\psi, \tau^2)}_{\text{posterior for } \mu} \times \underbrace{p(\psi)p(\tau^2)p(\sigma^2)}_{\text{priors}} \end{aligned} \tag{7.1}$$

Note that there would normally be a large set of parameters to estimate here, by imposing the hierarchical structure we only need to estimate within gropu means (governed by the priors).

$$= \underbrace{\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} p(y_{ij}|\mu_j, \sigma^2)\right]}_{A} \underbrace{\left[\prod_{j=1}^{J} p(\mu_j|\psi, \tau^2)\right]}_{B} \underbrace{p(\psi)p(\tau^2)p(\sigma^2)}_{C}$$

[1] Train has a concise intro to Gibbs and other MCMC

- A is the conditional likelihood, since we have independence across both groups and individuals it can be expressed as 2 products.
- B is the prior for the parameters, by independence across groups we write this as a product.
- C is the prior for the hyperparameters

Explanation. ***How does the form of B relate to the distinction between FE and RE in the classical panel data model?***
RE can be represented by the full hierarchical model, we set the hyperprior mean to $\bar{y}$ (the empirical Bayes approach) and get estimates equal to classical RE estimates. The RE model involved a hyperprior that gives a distribution with a common mean for the μ_j's.
The FE model does not have a hyperprior, ie C is not included. For example each fixed effect μ_j has a prior distribution with mean 0 and variance ∞. □

If the sample size is small, the estimated variance of the group means $\bar{y}_j$ - the estimator of μ_j - will be large. It might be that the data supports some degree of pooling or shrinkage across groups to get a better estimate of μ_j[2].

7.1.2 Empirical Bayes

The form of the posterior in (7.1) can cause computational problems given the number of parameters and potentially small sample size. An Empirical Bayes approach represents an approximation to full Bayesian model.
Consider the following hierarchical model for group means:

$$\begin{aligned} \bar{y}_j|\mu_j &\overset{\text{iid}}{\sim} N(\mu_j, \sigma^2) \quad j = 1, \dots, J \quad (\sigma^2 \text{ known}) \\ \mu_j &\overset{\text{iid}}{\sim} N(\psi, \tau^2) \quad j = 1, \dots, J \quad (\tau^2 \text{ and } \psi \text{ unknown}) \end{aligned}$$

We can write the posterior distribution for μ_j as:

$$p(\mu_j|\bar{y}_j, \psi, \tau^2) = \frac{f(\bar{y}_j|\mu_j)p(\mu_j|\psi, \tau^2)}{\int f(\bar{y}_j|\mu_j)p(\mu_j|\psi, \tau^2)d\mu_j}$$

Here we have assumed σ^2 is known, so no prior is required, however there is no prior for ψ and τ^2. We can construct the posterior using estimates of these hyperparameters from the data:

$$p(\mu_j|\bar{y}_j, \hat{\psi}, \hat{\tau}^2) \sim N(\hat{S}\bar{y}_j + (1 - \hat{S})\hat{y}_j, (1 - \hat{S})\hat{\sigma}^2)$$

The EV estimator of the mean of the posterior distribution is a weighted average of the sample mean and the prior mean, where the weight is the shrinkage factor $\hat{S}$:

7.1.3 Panel Data

Consider the following unobserved linear panel data model:

$$y_{it} = \mu + \delta_i + \omega_{it}$$

The individual unobserved effects are $\mu_i = \mu + \delta_i$. Let us assume:

$$\begin{aligned} var(y_{it}) &= var(\delta_i) + var(\omega_{it}) \\ &= \tau^2 + \sigma^2 \end{aligned}$$

[2]See Stein's paradox which defines situations in which there are estimators better than the arithmetic average.

Here we're making the same assumption as in RE, that $\delta_i t$ and ω_{it} are independent. This gives us a Bayesian representation for the RE estimator:

$$y_{it}|\mu_i \sim N(\mu_i, \sigma^2) \tag{7.2}$$
$$\mu_i \sim N(\mu, \tau^2) \tag{7.3}$$
$$\mu_i|y \sim N(\hat{\mu}_i, \sigma^2 + \tau^2) \tag{7.4}$$

Note that unlike in RE we need to make assumptions on the full distributions from the outset, we can's just use moments as before.
(7.2) is the likelihood, we assume normal data here. (7.3) is the normal prior for individual effects, and (7.4) is the posterior distribution for μ_i. $\hat{\mu}_i$ is the RE estimator and the mean of the posterior distribution for μ_i. It is also an EB estimator.

Claim 7.1.1. The product of two univariate Gaussian PDFs is proportional to a Gaussian PDF.

Proof. Let $f(x)$ and $g(x)$ be two Gaussian PDFs with means μ_f and μ_g and variances σ_f^2 and σ_g^2 respectively:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\left(-\frac{(x-\mu_f)^2}{2\sigma_f^2}\right) \quad \text{and} \quad g(x) = \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left(-\frac{(x-\mu_g)^2}{2\sigma_g^2}\right)$$

Their product is:

$$f(x)g(x) = \frac{1}{2\pi\sigma_f\sigma_g} \exp - \left(\frac{(x-\mu_f)^2}{2\sigma_f^2} + \frac{(x-\mu_g)^2}{2\sigma_g^2}\right)$$

Examine the term in the exponent:

$$\begin{aligned}
\frac{(x-\mu_f)^2}{2\sigma_f^2} + \frac{(x-\mu_g)^2}{2\sigma_g^2} &= \frac{(\sigma_g^2(x-\mu_f)^2 + \sigma_f^2(x-\mu_g)^2)}{2\sigma_f^2\sigma_g^2} \\
&= \frac{(\sigma_g^2+\sigma_f^2)x^2 - 2(\sigma_g^2\mu_f + \sigma_f^2\mu_g)x + \sigma_f^2\mu_g^2 + \sigma_g^2\mu_f^2}{2\sigma_f^2\sigma_g^2} \\
&= \frac{x^2 - 2\frac{\sigma_g^2\mu_f+\sigma_f^2\mu_g}{\sigma_g^2+\sigma_f^2}x + \frac{\sigma_f^2\mu_g^2+\sigma_g^2\mu_f^2}{\sigma_g^2+\sigma_f^2}}{2\frac{\sigma_f^2\sigma_g^2}{\sigma_g^2+\sigma_f^2}}
\end{aligned}$$

This is a quadratic in x, and is thus also a Gaussian function. To pin down the new parameters we just need to complete the square and compare with the standard form of the Gaussian PDF. Since a term ε can be added that is independent of x to complete the square, we can write the exponent wlog as:

$$\frac{\left(x - \frac{\sigma_g^2\mu_f+\sigma_f^2\mu_g}{\sigma_g^2+\sigma_f^2}\right)^2}{2\frac{\sigma_f^2\sigma_g^2}{\sigma_g^2+\sigma_f^2}} + \varepsilon$$

Thus we have:

$$\mu_{fg} = \frac{\sigma_g^2\mu_f + \sigma_f^2\mu_g}{\sigma_g^2+\sigma_f^2} = \frac{\mu_f/\sigma_f^2 + \mu_g/\sigma_g^2}{1/\sigma_f^2 + 1/\sigma_g^2} \quad \text{and} \quad \sigma_{fg}^2 = \frac{\sigma_f^2\sigma_g^2}{\sigma_g^2+\sigma_f^2}$$

□

We can now derive (with some abuse of notation) the posterior distribution for μ_i:

$$\begin{aligned} p(\mu_i|y) &\propto p(y|\mu_i)p(\mu_i) \\ &= N(\mu_i, \sigma^2)N(\mu, \tau^2) \end{aligned}$$

Using 7.1.1 we can write the mean of the posterior distribution as:

$$\mathbb{E}[\mu_i|y] = \frac{\frac{\mu}{\tau^2} + \frac{\mu_i}{\sigma^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}}$$

We can write our finite sample estimate $\hat{\mu}_i$ as:

$$\begin{aligned} \hat{\mu}_i &= \frac{\frac{\bar{y}_i}{\hat{\sigma}^2} + \frac{\bar{y}}{\hat{\tau}^2}}{\frac{1}{\hat{\sigma}^2} + \frac{1}{\hat{\tau}^2}} \\ &= \frac{\hat{\sigma}^2\bar{y}_i + \hat{\tau}^2\bar{y}}{\hat{\sigma}^2 + \hat{\tau}^2} \\ &= (1 - \hat{S})\bar{y}_i + \hat{S}\bar{y} \quad \text{where} \quad \hat{S} = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\tau}^2} \end{aligned}$$

For $0 < \hat{S} < 1$, $\hat{\mu}_i$ is a compromise between the pooled estimator $\hat{S} = 1$ and the (individual means) FE estimator $\hat{S} = 0$.

Explanation. ***What happens when*** $\tau^2 \to 0$***?***
As $\tau^2 \to 0$, there is no longer any variation between individuals, so it is optimal to use the fully pooled model. Clearly we can see:

$$\lim_{\tau^2 \to 0} \hat{S} = 1 \quad \Rightarrow \quad \lim_{\tau^2 \to 0} \hat{\mu}_i = \bar{y}$$

that our estimator of μ_i becomes the population mean.
What happens when $\tau^2 \to \infty$***?***
As $\tau^2 \to \infty$, the individual means are too noisy to be informative, so it is optimal to use the individual means estimator (FE). This is analogous to RE → FE when $\sigma_\alpha^2 \to \infty$¿

$$\lim_{\tau^2 \to \infty} \hat{S} = 0 \quad \Rightarrow \quad \lim_{\tau^2 \to \infty} \hat{\mu}_i = \bar{y}_i$$

□

Explanation. ***How does*** (7.1) ***differ from the RE distribution?***
The distribution we considered in RE was $\alpha_i \sim i.i, d, (0, \sigma_\alpha^2)$, however the mean is unimportant since any mean would be absorbed by the constant terms.
We are making the same *i.d.d.* assumption here, with similar assumptions on the variance. The difference is that now we require the full distribution, not just the moments. □

How do Bayesians conceptualise fixed versus random effects estimators when all effects are random? The classical fixed vs RE dichotomy is not relevant. Here the distinction is:

- RE: Hierarchical prior $\mu_i \sim N(\mu, \tau^2)$
- FE: Non-hierarchical independence prior for each μ_i

7.2 Model Averaging

The problem: Estimation and inference on the determinants of y, where the set of regressors is large. Let θ denote the parameters of the regressors included in a model, we can estimate the posterior density as $p(\theta|y, M_j)$, where M_j is the jth model. Suppose there are K potential regressors, model M_j is described by a $K \times 1$ binary vector γ_j where $\gamma_{jk} = 1$ if regressor k is included in model M_j, and $\gamma_{jk} = 0$ otherwise.
The model space M is thus the set of all 2^K possible models. *How do we account for model uncertainty in making unconditional (on the space of models) inference on any given element of θ?*

7.2.1 Marginal Likelihood, Prior and Posterior Odds and the Bayes Factor

Marginal likelihood (integrated over the parameter space):

$$\ell(y|M_j) = \int_\theta p(y|\theta, M_j)\ell(\theta|M_j)d\theta$$

For two models M_i and M_j the posterior odds is given by:

$$\frac{p(M_i|y)}{p(M_j|y)} = \frac{p(M_i)}{p(M_j)}\frac{\ell(y|M_i)}{\ell(y|M_j)}$$

Using Bayes theorem we can write the posterior model probabilities as:

$$p(M_j|y) = \frac{p(M_j)\ell(y|M_j)}{\ell(y)} \propto p(M_j)\ell(y|M_j)$$

If the set of models is exhaustive, we can write:

$$p(M_j|y) = \frac{p(M_j)\ell(y|M_j)}{\sum_{i=1}^{2^K} p(M_i)\ell(y|M_i)}$$

We can compute the posterior distribution of θ given model M_j as:

$$p(\theta|y, M_j) = \frac{\ell(y|M_j, \theta)p(\theta|M_j)}{p(y|M_j)}$$

And the unconditional posterior distribution of θ is:

$$p(\theta|y) = \sum_{j=1}^{2^K} p(\theta|y, M_j)p(M_j|y)$$

We can thus think of the posterior model probabilities as the weights $p(M_j|y)$ that we should attach to the posterior distributions $p(\theta|y, M_j)$ when averaging over the model space.

7.2.2 Prior

How do we get priors for:

- The model space $p(M_j)$
- Elements of θ that are model=specific (EG regression parameters)
- Elements of θ that are common to all models (EG variance parameters)

Here we will focus on $p(M_j)$. Recall that a linear model is described by a set of binary variables responsible for including/excluding regressors. An independent Bernoulli prior for a given model $p(M_j|\pi)$ can be written as:

$$p(M_j|\pi) = p(\gamma|\pi) = \prod_{k=1}^{K} \pi_k^{\gamma_{jk}} (1-\pi_k)^{1-\gamma_{jk}}$$

where π_k is the independent prior probability that regressor k is included in the model. For example a uniform prior across all elements, $\pi = 0.5$ implies the number of covariates should be large. Note the limitation of this approach, we would expect the presence of some regressors to be correlated with the presence of others.

Definition 7.2.1: Prior Inclusion Probability

$$p(\gamma_{jk} = 1|y) = \sum_{j=1}^{2^K} \mathbb{1}(\gamma_{jk} = 1|y, M_j) p(M_j|y)$$

Note that this is essentially a counting rule, we are summing the probabilities of a model being selected which includes $\gamma_{jk} = 1$. This can be thought of as a measure of how important regressor k is in the model space.