R301 Notes on Lectures

Tobias Leigh-Wood

Lent 2023

Topic 4: The Mixed Logit Model

- Reading is important for mixed logit. Train is essential. (I HAVENT DONE THIS)
- Logit models suffer from independence of irrelevant alternatives. So the odds ratio can be just the ratio between two options — but this is not going to be true in general. Also logit models assume that everyone has the same partial effects and parameters.
- Mixed logit allows us to relax these assumptions.
- Consider following RUM model $Y_{ij} = v_j'\omega_i + \epsilon_{ij}$
  The parameter $\omega_i = \bar{\omega} + \tilde{\omega}$ The distribution of ω_i is $g(\omega|\bar{\omega}, \mathit{\Omega})$ Each parameter has an individual distribution and a joint distribution.
- Probability an individual chooses j
- $P(j|v,\eta) = \int_\omega L(j;v;\omega)g(\omega|\eta)d\omega = \int \exp^{v_j\omega} / \sum_{j\in\Sigma} \exp^{v_j\omega} g(\omega|\eta)d\omega$
- This has a logit core but then ω is a random variable which we condition on. We estimate the mean and covariance of g. $\omega = \bar{\omega} + \Lambda\zeta$ $\eta = (\bar{\omega}, \sum_\omega)$ is a vector of hyper parameters.
- Then we can get the log liklihood by taking logs of the joint probablity of P.
- If choices are independent then the variance of $g()$ is just a diagonal matrix with the variances of each choice.
- Mulitnomial logit means that the mean fully describes the distribution of the parameters.
- We can discretise $P(j|v,\eta)$ and let ω take on m values. This is referred to a latent class model.

- Because we get a mixture of L (often different) distributions we need to simulate the integrals.
- This is a specific form of parametric random parameter models.
  $f(y_i|x_i,\eta) = \int f(y_i|x_i;\gamma_i)g(\gamma|\eta)d\gamma$ The mixed logit model is a form of this more general type of model.
- But in general econometrics is moving toward models that don't make all of these distributional assumptions. I.e do we actually believe the erros are extreme value distributd?
- We have two types of model — Random Coefficient Mixed logit and Error Components Mixed Logit.
- **Random Coefficient Mixed logit** — Capture variation in preferences related to observed attributes, which may be interacted with observed characteristics.
  We get a composite error term that depends on ω $\zeta_i = V'\tilde{\omega}_i + \epsilon_i$ $\mathit{\Omega}$
  Tied together by constant taste parameters. In the lectures we look at the covariance between two choices. Probably useful to write that out.
- **Error components mixed logit** — Capture correltations amongst alternatives.
  $U_{ij} = v_j' + z_j'\mu_i + \epsilon_{ij}$
  μ_i is a vector of stochastic terms with zero mean. We let alternatives share this part of utility which induces correlation amongst alternatives. K different groups of alternatives. We can write the $Jx1$ composite error term $\zeta = z_j'\mu_i + \epsilon_j$ CHECK SUBSCRIPTS ON THIS....
- Example: Nested logit for 3G and 2G phones.
- Example: Introduction of electric car
  Logit: Electric car draws from all others cars proportionately.
  Mixed logit: Electric car draws proportionately from more similar cars — perhaps these cars are smaller so then the effect of electric cars on emissions is smaller.
- **Preferences for Electricity Supply**
  Varying preferences for electricity supply. Stated preference survey. We estimate the mean and standard deviation of different variables — in the lecture example we set all covariances to 0. We can allow this to vary.
  This gives us the preferences of a population — and then we can say what proportion of populatation has a positive coefficient and other similar questions.
- Other papers include Berry et al 1995.
-

Topic 5: Panel Data Models

- Back to linear models.
- Main difference will be correlated random effects. When we set this up we get a natural test for the random effects estimator.
- Asymptotic results generally based on small T and large N.
- Several advantages: more observatoins, discriminate between multiple hypothesis, control for unobserved heterogeneity at the individual level.
- Couple of examples – unions, production functions. Production function might have correlation between managerial quality, time invariant, and labour input, time varying.
- We restrict our models to homogeneous parameters. We don't let parameters change across time or individuals.
- Understand the price of estimators. Be able to talk about the trade off betweem them.
- Consider a composite error term $\alpha_i + \epsilon_{it}$. A pooled model sets $\alpha = \alpha_i$. If $\alpha \neq \alpha$ and uncorrelated with x_{it}, then the pooled OLS estimator (POLS) requires panel-corrected standard errors. If correlated with x_{it} then POLS is inconsistent.
- Assumptions for POLS:

$$
\begin{aligned}
y_{it} &= \beta x_{it} + v_{it} \\
v_{it} &= \alpha_i + \epsilon_{it}
\end{aligned}
$$

  We focus on two issues. Distribution of α_i and strict exogeneity of ϵ_{it}.

- Key factor is whether the random effects are correlated with x_{it}
- Decompose total sum of squares to the within sum plus the between sum. When we do fixed effects we just use the within variation and discard the between sum.
- We can think about this in a similar way to IV — decomposing the endogenous variation from the exogenous.
- Random effects takes a weighted average of both parts of the within and between sum of squares.
- Strict exogeneity means that all the errors are uncorrelated. $E(\epsilon_{it}|x_{i1}, x_{i2}, x_{i3}, ..., x_{iT}, \alpha_i) = 0$ It's easy to think about violations. Then we have weak exogeneity where ϵ_{it} is uncorrelated with all x_{it} up to time period t.

- Least squares dummy variable regression is same as the fixed effects estimator.
- With the within estimator we need to have strict exogeneity — Why? Because the transformation of the error means that ϵ
- **Random effects estimator**
- Form of GLS estimator because we get a covariance matrix that we need to estimate.
- Pooled OLS is consistent but inefficient under the random effects specification.
- Under our assumptions we get a block diagonal covariance matrix because of assuming 0 covariance between units in a time period.
- Apply a transformation to all individuals — $\Sigma^{-1/2}$.
- OLS — no error covariance across obsrvations for a given i. Fixed effects — GLS converges to the Fixed effect estimator.
  Next lecture
- Have a look at the between estimator.
- Lecture started by looking at coefficients and standard errors. Seemed to say looking at standard errors was a better way to go than talking about coefficients in the exams.
- Random effects: Critical assumption – zero correlation between α and x_{it}.
- Hausman test: tests equivalence of random effects and within estimator.

  $$k = \hat{\beta}_{RE} - \hat{\beta}_W$$

  Distributed chi squared. No correction for clustered standard errors – the errors used for $\hat{\beta}_w$ are iid.
- We can use a correction.
- Quasi demeaned variables we get a random effects estimator with a x_{it} that is still in the equation.
- **Correlated Random Effects**
- Strict exogeneity — (EMPHISIE THIS) - understand with and without conditioning on x
- $E[\alpha_i | x_{i1}, ..., x_{iT}] = \zeta + \bar{x}_i \delta$
- Therefore $E[y_{it}|x_i] = E[y_{it}|x_{it}, \bar{x}_i] = x_{it}\beta + \zeta + \bar{x}_i\delta + v$
- Ignoring $\bar{x}_i$ we have a random effects model. By including it we allow endogeneity in the model because we can control for it.

- We can then use pooled OLS on this model to consistently estimate all parameters.
- Test of δ is a test of the RE specification.
- Does anyone use different conditioning statistics? I.e. condition on the sd rather than mean.

Topic 8: Bayesian Inference

- Cornfield Bayes Theorem 1967, Greenberg chapters 1-4 for background readings.
- Combining evidence, parameter estimation, paratmeter heterogeneity
- Main topics: hierarchical models, model uncertainty, data augmentation, bayesian moddelling
- Have a look at preable post lectures.
- Key differences: Estimate posterior distribution, no test statistics, no estimators. Hypothesis testing is reversed — no longer testing D condition on H, now we test H condition on D.
- Laplace BT. $P(B|A) = P(A \cap B)/(P(A))$
- Classical perspective give a probability distribution on θ, in bayesian world θ is fixed and we just have a prob about it.
- Bayesian analysis asks what the probility of cancer is if i smoke.
- We observe probability of smoking if I have cancer. THen we can use bayes theorem to back out probility of cancer is if i smoke.
- Binomial distribution – estimating p with bayes.

  $$Pr(p|x) = \frac{P(x|p)P(p)}{P(x)}$$

  $Pr(p)$ is my prior probability. $Pr(p|x)$ is my posterior distribution.
  Bayes solved this and said the the posterior distribution given x is the beta dist

  $$Beta(X+1, N-X+1)$$

  X and N come from binomial.
- Bayes theorem for parameters. $p(\theta, y)$ denotes a posterior distribution. Prob of y conditioning on θ.

  $$f(y|\theta) = \prod f(y_i|\theta)$$

Then posterior distribution is

$$p(\theta|y) = f(y|\theta)p(\theta)/f(y)$$

$f(y)$ is the marginal liklihood. $f(y|\theta)$ is the liklihood function. $p(\theta)$ is the prior distribution.

- **Natural conjugate priors**
- A class of priors is a conjugate for a family of likelihoods if both prior and posterior are in the same class for all data y. Basically means we can keep the same distribution for priors given a distribution of the data.
- We start from a framework of estimating the probabilit distribution. Not separting it and using null hyopthesis to form probability.
- With binomial and beta distribution the mean of the posterior is a weighted average of the MLE estimator and the prior mean. So if we've got enough data then the priors won't matter. Your priors are being swamped by the data. Your posterior collapses to the distribution of the data as you increase the amount of data?.
- How do we use data to revise a prior probability about a hypothesis?
- $P(H_0) = p(w = 0)$ and $P(H_1) = p(w = 1)$. Posterior prob for h_0.

  $$P(H_0|y) = p(w = 0|y) = \frac{p(w = 0)p(y|w = 0)}{p(y)}$$

  We then can get the Bayes Factor, posterior odds (relative likelihood) of posterior probs.

  $$BF = \frac{P(H_0|y)}{P(H_1|y)} = \frac{p(w = 0)p(y|w = 0)}{p(w = 1)p(y|w = 0)}$$

- Posterior distribution is proportional to the prior and the liklihood component.
- One topic of contention is where we get our priors from.
- Questions on covid in the lecture slides are worth looking at

Topic 8: ii

- Multilevel Data, Hierarchical moddelling
- Readings: Stein's Paradox in statistics – interesting apparently. Casella – short and informative.

- Model Averaging. Everything we've done so far takes the model as given, then we find values for the parameters etc all conditional on the model. Bayesians don't like this. They want to make inference on the determinants of y unconditional on the model. We specify a functional form but then estimate many models.
- Let θ denote parameters. We estimate the posterior density. Model M_i described by a k by 1 binary vector γ Where the k_{th} element of γ is one (zero) indicating the inclusion of x_k in the model.
- What we might be interested in: Marginal Likelihood, posterior model uncertainty, posterior model odds, posterior parameter distributions
- How do we elicit prior distributions for: Space of models $p(m_j)$, parameters θ that appear in specific models, parameters θ that are common across models.
- For example:
$$p(M_j|\pi) = p(\gamma|\pi) = \prod \pi^{\gamma_i}(1-\pi)^{1-\gamma_i}$$
where π_i denotes the independent prior inclusion of variable x_i in M_j. This would be a simple benchmark prior we could use.
- For example: Doppelhofer and Weeks 2009 use a bayesian set up to see which variables are useful in describing GDP growth. They find posterior inclusion probabilities. These tells us whether we should include variables or not in the model.

Multi level data

- It's common for data to be organised into hierarchical levels.
- Sampling model should take into account the hierarchical nature of the data.
- Consider a population with J groups and n_j individuals per group. Each group has a parameter θ_j so we can get within group variation and between group – between is determined by the different js.
- How do we write a posterior distribution for these parameters? Quite complicated model but we can use Gibbs Sampling which approximates the posterior distribution
$$p(\mu_1, ..., \mu_j, \psi, \tau^2, \sigma^2|y_1, ..., y_j) \propto p(y_1, ..., y_J|\mu_1, ..., \mu_j, \psi, \tau^2, \sigma^2)\times p(\mu_1, ..., \mu_j|\tau^2, \sigma^2)\times p(\psi)p(\tau^2)p(\sigma^2)$$
- We can use this set up to talk about FE and RE in the classical panel data model — FE has constant mu's. Didn't understand this fully.
- If the sample size is small the estimated variance of the group may be large. Stein's paradox defines circumstances in which there are estimators better than the arithmetic average ((HAVE A LOOK AT THIS)).

- **Empirical Bayes**
- We need to assume a parametric setup so that we can have conjugate pairs – others our priors and posteriors won't match which wouldn't make sense.
- Empirical Bayes estimators the hyperparameters with data. (this is the Greene reading I think?)
- Bayesians don't like this because we are essentially estimating our priors from the data. We can then estimate posterior using our estimated hyperparameters. See slide 14.
- Baseball example: We have batting average after first 45 bats. We want to estimate the true mean. Morris 1983 utilizes Empirical Bayes. The EB means did better than just using arithmetic mean in predicting the true mean.
- Now link panel data models with Empirical Bayes. The random effects estimator is an inverse variance weighted average of the pooled estimator and the FE estimator. Compromise between the two models.
- If we go back to panel data notes we can see something similar where the random effects estimator is a weighted average of the pooled and FE estimators.

  Summary
- OLS treats α_i all the same as constant. If α_i has a distribution we can use RE. If α are all different but not random we use FE.
- In a Bayesian world all parameters are treated as random variables, so α_i has a distribution.
- For next week – Chapter 14 of Train has a good overview.

Topic 8:iii EM Algorithm

- Classical panel data distinction between RE and FE doesnt apply. The prior for RE is a normal assumption. The FE each one has it's own separate distribution, so each effect is different and not tied together. Bayesian FE is about how the priors are treated.
- EM algorithm for a binary choice problem. Data augemntation is the bayseian analog of the EM algo. Then we revist some discrete choice models.
- Train chapter 14 is critical for this lecture. Recomneded for the course.

- Looking for techniques to deal with missing data. How can we use maximisation techniques to do this
- Latent variables are a form of missing data. For example unobserved utility is a form of missing data but we observe choices which give us some information.
- We could create the marginal liklihood by integrating out the missing data.
- But we contruct the likelihood with missing data, called the complete data likelihood.
- Calculate the expected value of the complete data likelihood. And then iterate with θ, note that this determines likelihood of missing data as well.
- Now we re do it for binary response.
- Curse of dimensionality in random utility models
- as number of choices gets large it becomes more and more difficult to maximise the likelihood. Because we get multidimensional integrals. Bayesian approach does not help us.
- this is where data augmentation comes into it. Treat unobserved data the same way as we treat unobserved parameters. Prior over missing data becomes a function of θ.
- DA converts likelihood into deterministic function which is much easier to work with than a likelihood.
- Not covering Gibbs sampling this year.
- **Topic: iiii**
- Simulated maximum likelihood.
- Make sure I understand the trinomial probit model.
- Now we look at Bayesian Mixed Logit. Again we are faced with the issue of the posterior of the distribution being difficult to integrate out. We can simulate them out. OR we change the problem by augmenting the distribution with the unknown data. And then if

Topic 9: Machine learning approaches

- Elements of statistical learning is the bible!

- One culture assumes data is genrated from a stochasic process, the other assumes data generating process is unknown and we use the data to find the distribution.
- Random Forest is a particular form of regression. There will be hyper parameters, such as k in nearest neighbours or bins in a histogram.
- Decision tree just splits up x's into different bins. But it may fit the data wwell but id overfitted. SO we use a random forest. Build lots of trees for different samples and then just average.
- This is a bit like calculating a mean for each subgroup of the population. Because we are making an average.
- Causal inference in these models is the frontier of the research and is where we are heading.
- If add ages to the cef of wages we get something that is relatively similar to k means. i.e. $E(wages|sex, age) = g(x)$.
- We need to decide what variables to split for decision/logic trees.
- Density estimations. Problem of reconstructing the pdf using a set of given data. Essentially a histogram. Averaging over the points that are within a bin. We can then smooth this with a non para metric kernel estimator because the histograms has large discontinuities. Bin width is a hyper parameter, also known as a regulrisation parameter. Obviously as you decrease the bin size you start to overfit the data, changes might reflect idiosyncratic differnces rather than the true data generating process.
- k-nearest neighbours is one of the simplist non-parametric estimators. $y_i = f(x_i) + \epsilon$ where f() is the average of k-nearest data points.
- You can have local ks but most researchers assume a global value for k.
-