Lecture 9 - Discrete Choice and GLM

Paul Goldsmith-Pinkham

February 26, 2024

We are now going to generalize our estimation problem beyond linear models like linear (and quantile) regression, and consider more complex objective functions. This will initially be motivated by the binary choice model, but will be more generally applicable to a wide range of problems. This will lead to us covering a wide range of topics, including binary choice models, generalized linear models (GLMs), numerical estimation methods for non-linear models, inconsistency of non-linear models with many parameters, and the challenges of estimating models with multiple discrete choices.

Conceptually, we will be considering minimizing *objective functions* as a general case of minimizing *squares*.

Binary choice

Consider the following binary outcome problem: let Y_i denote if person i is a homeowner, and X_i includes three covariates: income, age and age^2 (plus a constant). How should we model the relationship between X and Y? Conceptually, a very general form would consider

$$Y_i = F_i(X_i),$$

where F_i could vary by individual. However, this doesn't seem like a very good model for considering estimands, such as "how much does homeownership increase with a 10k increase in income?"[1] In many ways, this is similar to the questions related to binscatter and other semiparametric models.

[1] Formally, this would look something like $E(dF_i(X_i)/dX_i|X_i)$, and we would need to make some assumptions on F_i to make progress. That's what we'll do now.

The potentially issues with blithly assuming a linear model for $F_i(X_i)$ becomes very apparent in the context of a binary dependent variable. Say we model this outcome using a linear regression (this is often called a linear probability model), assuming strong ignorability or just $E(\epsilon_i|X_i) = 0$:

$$E(Y_i|X_i) = Pr(Y_i = 1|X_i) = X_i\beta \qquad \rightarrow Y_i = X_i\beta + \epsilon_i \tag{1}$$

The problems with modeling Y in this way is twofold. First, since the outcome is binary, the error structure will be bimodal and unusual looking. To see this, consider $\varepsilon_i = Y_i - X_i\beta$, and consider how ε_i changes for $Y_i = 0$ vs 1. For a given X_i, it is exactly bimodal (like the outcome). One implication of this is that $V(Y|X) = X_i\beta(1 - X_i\beta)$, and you'll have pretty significant heteroskedasticity. This is solveable

using robust standard errors, but does mean that a normal approximation with the error is a poor one.

Seoncd, except under some special circumstances, it's very likely that the predicted values of Y_i will be outside of $[0, 1]$. What's an example where they will not be? Discrete exhaustive regressors! Why? Discrete exhaustive regressors are the one setting where you can guarantee that the model is correctly specified. When the model is misspecified, it is quite possible that the model will extrapolate in a way such that there will be values outside support.

How does this impact our causal estimates? If the model is correctly specified, we can generate counterfactual predictions of the outcome. If not, then we get a linear approximation that may be nonsensical.

Table 1: LPM model estimates

variable	linear est.	std.error
Intercept	0.0242	0.0410
age	0.0220	0.0017
age^2	-0.0002	0.0000
income /10k	0.0069	0.0007

Example 1 (LPM estimates of homeownership)
We estimate the linear model in Table 1. and note tthat if income were strictly ignorable, we could say that 10k increase in income leads to 0.69 p.p. increase in the probability of homeownership. But, the predicted probability of homeownership would range from 0.283 *to* 1.78*. Oops.*

Modeling discrete choice

There are two ways to think about how we think about this estimation problem. These are *not* mutually exclusive, and it is important to note that both of these approaches are very focused on the *model-based* aspect of estimating causal effects.

The first is a statistical view. How can we model the statistical process for Y_i better? In other words, can we fit the outcome model better? Consider $X_i\beta$ as the conditional mean of some process, what's the statistical model that fits with this? This is a case of what's termed "Generalized Linear Models" (GLM)

A second way to view this is as an structural (economic) choice problem. Most models of binary outcome variabels assume a latent index, on the utility of choosing Y_i:[2]

$$Y_i^* = X_i\beta + \varepsilon_i, \qquad Y_i = \begin{cases} 1 & Y_i^* > 0 \\ 0 & Y_i^* \leq 0. \end{cases} \tag{2}$$

[2] The careful reader will note that analogy to the Heckman model on treatment choice.

As we will now see, both approaches do arrive at a similar modeling conclusion, but the latter model will naturally accomodate choices.

A natural approach in either of these is to make a distributional assumption about ε_i. Two common assumptions:

1. ε_i is conditionally normally distributed (probit), such that $Pr(Y_i = 1|X_i) = \Phi(X_i\beta)$
2. ε_i is conditionally extreme value (logistic) such that $Pr(Y_i = 1|X_i) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$

Note that these are not, in the binary setting, deeply substantive assumptions. In Figure 1, we see that there are very minor differences in the thickness of the tails for a logit vs. normal error, but they're both symmetric and centered around zero.[3] One downside for probit models is that there's no closed form solution for Φ, the CDF for the normal distribution:

$$\Phi(X_i\beta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X_i\beta} e^{-t^2/2} dt \tag{3}$$

We will discuss later how to estimate β given these assumptions, but they will involve numerical optimization, as there is no closed form for β like in linear regression.

Example 1 (continued)

Consider now the same homeowner problem from Example 1, but estimated with logit. The β coefficients in Column 1 of Table 2 are hard to interpret. To see why, consider the derviative of the probability with respect to X_i:

$$\frac{\partial Pr(Y_i = 1|X_i)}{\partial X_i} = \beta\phi(X_i\beta) \qquad \textit{(Probit)}$$

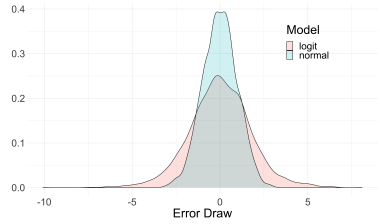
$$\frac{\partial Pr(Y_i = 1|X_i)}{\partial X_i} = \beta\frac{\exp(X_i\beta)}{(1+\exp(X_i\beta))^2} \qquad \textit{(Logit).}$$

In both cases, the effect of X_i changes, depending on the value of X_i. This is a problem for interpretation. The average derivative in Column 3 is a way to get around this, but it's not a perfect solution:

$$n^{-1}\sum_i \frac{\partial E(Y|X)}{\partial X} = n^{-1}\sum_i \beta\frac{\exp(X_i\beta)}{(1+\exp(X_i\beta))^2}$$

This will calculate the derivative for every value in the sample, and then average them. This is a way to get a sense of the average effect of X_i on Y_i. We see a much larger effect of income on homeownership in the logit model than in the linear model (Column 2).

Figure 1: Logit vs. Probit error terms



[3] Important caveat: these models only identify β up to scale. Why? The "true" model of ϵ could have variance σ^2 that is unknown. Consider if $F(X_i\beta) = \Phi(X_i\beta)$. If this were a general normal (rather than standardized with variance 1), we could just scale up the coefficients proportionate to σ and the realized binary outcome would identical. Hence, we normalize $\sigma = 1$ in most cases. This is *not* a meaningful assumption.

Table 2: Homeownership problem estimated with logit

term	(1) logit est.	(2) linear est.	(3) avg. deriv.
constant	-2.14	0.0242	-0.392
age	0.0903	0.022	0.0166
age^2	-0.0006	-0.0002	-0.0001
income/10k	0.0716	0.0069	0.0131

Figure 2: Linear vs. Logit model predictions

Example 1 (continued)
Figure 2 shows the predicted values of homeownership from the linear and logit models. The linear model is predicting values outside of the support of the outcome, and the logit model is not. This is one benefit of correctly specifying the model.

Generalized Linear Models (GLM)

We can generalize the intuition above, where we let the underlying distribution of ϵ be non-normal, and parameterize the mean of the distribution to be a function of $X_i\beta$. This is the idea behind Generalized Linear Models (GLM), originally formulated in Nelder and Wedderburn [1972].[4]

The overall setup of GLMs in broad strokes is to consider estimation of a **linear model** $X\beta$, which is linked to the conditional mean $E(Y|X)$ by **a link function** g: $E(Y|X) = g^{-1}(X\beta)$. The crucial underlying assumption for the underlying machinery is that Y, the outcome, is distributed by some member of the **exponential family** of distributions. This includes the normal, binomial, Poisson, and gamma distributions, among others.

Some simple examples of GLMs include:

1. Logit, with a link function $g^{-1}(X_i\beta) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$
2. Normal, with an identity link function $g^{-1}(X_i\beta) = X_i\beta$
3. Poisson, with a log link function $g^{-1}(X_i\beta) = \exp(X_i\beta)$

In essence, we can enforce a linear functional form to the *mean*, and allow the error distribution to fit the form of the data.[5]

[4] Interestingly, this is very common in non-economics fields, but much less common in economics.

[5] It's interesting to note the underlying machinery of GLMs is similar to many of the selection and discrete choice models we've discussed and discuss today. The linear index provides an extremely convenient parameterization of the mean, but also makes some particular assumptions about the substitutibilty of the covariates.

We will now discuss the Poisson regression case in more detail, as it tends to be underused in economics, and is a very important use case. A key takeaway in GLM, like with OLS, is that it is possible to correctly specify just the conditional mean function and then robustly estimation standard errors on parameters of that function, rather than fully specifying the distribution correctly.

Poisson Regression for non-negative outcomes

Consider an non-negative outcome $Y \geq 0$. There are a huge host of outcomes in economics and finance that are restricted to this support: investment, assets, wages, patent citations, output, and so on. We are often interested in the estimand of the partial effect $dE(Y|X)/dX$. If we estimate this conditional with linear regression (e.g. by assuming $Y_i = X_i\beta + \epsilon_i$), what are potential issues?

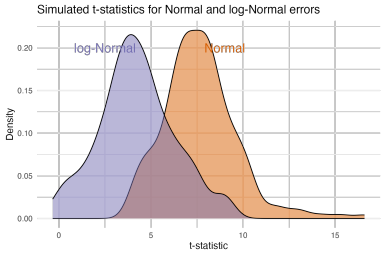
Mechanically, the error terms $\hat{\epsilon}_i = Y_i - X_i\hat{\beta}$ will be skewed, since Y_i is skewed. This is not on its own a huge issue, but it does suggest that the asymptotic approximation for $\hat{\beta}$ will be worse for a given n. This leads to highly influential outliers for OLS as well.[6]

[6] Note that one solution to this issue is to consider quantile regression instead!

Comment 1

Consider two outcomes, Y_1 and Y_2. In both cases, the true model is linear (with coefficient of 1) with respect to X, but the error term is Normal with mean zero and variance 1 in Y_1, and is log-Normal with mean zero and variance 1 in Y_2. If we simulate and estimate this model using linear regression, plotting the t-statistic of the coefficient on X for each model, we find much higher power for the model with Normal errors, rather than log-Normal errors. This reflects the lack of efficiency of OLS in the presence of non-Normal errors (but not a lack of consistency!). See Figure 3 for a visual representation of this.

Figure 3: Non-normal errors in linear regression



What are solutions to this issue? One commonly used approach is to estimate linear regressions on $\log(Y)$ instead of Y_i. This solves many of the outlier and skew issues,[7] but creates its own problems.

First, the parameters have a different interpretation. Note that the units for the outcome are different (log points). Often, these are interpreted in percentage points, since log differences are approximately equal to percentage changes.[8] This is useful, but can at times be confusing (e.g. what is the actual level effect? Sometimes a percentage effect can exaggerate or minimize a large level effect).

Second, what if $Y = 0$? This is a problem, as $\log(0)$ is undefined. One common solution is to use $\log(1 + Y)$ or $arcsinh(Y)$ [Manning and Mullahy, 2001, Ravallion, 2017, Bellemare and Wichman, 2020],

[7] In the ideal case, note that a log-Normal outcome will be exactly linear after using logs.

[8] Recall $\log(Y_1) - \log(Y_0) = \log(Y_1/Y_0) = \log(1 + \Delta Y/Y_0) \approx \Delta Y/Y_0$ for $\Delta Y/Y_0$ small.

which solves the second problem, but makes the first problem even worse! [Bellemare and Wichman, 2020, Aihounton and Henningsen, 2021, Cohn et al., 2022] Why these solutions? For one, they're both well-defined at $Y = 0$. Second, it has "similar" properties to the taking a log. Effectively, since the distance between $\log(1 + Y)$ and $\log(Y)$ was small as Y gets large, the hope is that the differences would "wash out." It turns out, thanks to work by Chen and Roth [2023], that neither of these solutions are a good idea and that these differences do not wash out.

The key point of Chen and Roth [2023] is that percentage effects are not well-defined for outcomes that are potentially zero-valued. That is in some ways obvious – there is no way to talk about the percent increase for something where the base-level is zero. Dividing by zero is infinite! But recall that part of the goal of using log outcomes was to approximate percentage changes in the outcome due to treatments. The main result of Chen and Roth [2023] shows that for *any other function* approximating log, but defined at zero, the results will be arbitrarily sensitive to changes in units (e.g. dollars to yuan).[9]

[9] This includes both $\log(1 + y)$ and $arcsinh(y)$.

What drives this effect? Effects close to zero and at zero. Most importantly, the *extensive margin* of moving from zero to non-zero has huge, and arbitrary, impacts on estimates on these types of rescaling. Put precisely, if you change the units of the outcome by a (e.g. $a = 100$, converting from cents to dollars), then the estimated effect will change by $\log(a)$ multiplied by the *extensive margin* effect. Note that this is fails scale equivariance, which is the property of OLS and quantile regression that usually makes a good estimator.[10]

[10] Note also the practical implication: if there is a big extensive margin effect, a large a has a big effect. In contrast, with a small a, then most effects will be close to zero (since they are extensive margin, and hence close to zero by definition).

This can have some really serious implications. Chen and Roth [2023] find that for half the papers they surveyed in the AER, the estimated effects would change by more than 100% if the units of the outcome were changed by 100 (e.g. dollars to Yen). This is a non-trivial effect!

The takeaway I want you to have: you should not be running a regression with $\log(1 + Y)$ or $arcsinh(Y)$ on the left-hand side![11] So what should you do if you have a zero in your left-hand side variable? Chen and Roth [2023] suggest other ways of considering these situations:[12]

[11] I have done this, historically, in my own work – we're all flawed creatures trying to inch towards better methodological implementations!

[12] These solutions are not perfect, but are motivated by a "trilemma" they prove: it is not possible to have an estimator that is simultaneously (1) an average of individual level treatment effects (2) invariant to rescaling of units and (3) point-identified without more assumptions about the joint distribution of the potential outcomes (beyond what we usually do in regression).

1. First, if you really need something interpretable as as a percentage effect (e.g. rescaling an ATE into percentage), you could estimate $\tau = E(Y_i(1) - Y_i(0))/E(Y_i(0))$, which scales the ATE by the baseline average. *This is the estimand targeted by Poisson regression.* There are also other normalizations one could consider. Instead of normalizing by $E(Y_i(0))$, if there is a pre-treamtent characteristic that is exogeneous, you could normalize by $E(Y_i(0)|W_i)$, e.g. the predicted baseline value given characteristic W_i. This captures

richer heterogeneity in the baseline characteristic, and may do a better job of reducing skewness.

2. Second, you could redefine the outcome in terms of functionals of the distribution, e.g. $\tilde{Y} = F_{Y^*}(Y)$. A prominent example is looking at the rank of an individual relative to the overall individual, as in Chetty et al. [2014].

3. If the goal is to consider trade-offs in some like concave preferences, then it is plausible to specify exactly the 'value' of a person at $Y = 0$, relative to positive Y, and then explicitly evaluate the parameter that way. This has the problem of losing scale-invariance, but at least the research is explicit about how they value these issues.

4. Finally, it is plausible to directly estimate the extensive and intensive effects separately. However, the intensive effect is only partially identified; we will explore this further in later lecutres.

See Table 3 for a full set of alternative estimators.

Description	**Parameter**	**Pros/Cons**
Normalized ATE	$E(Y(1) - E(Y(0))$	Pro: Percent interpretation Con: Does not capture decreasing returns
Normalized outcome	$E(Y(1)/X - Y(0)/X)$	Pro: Per-unit-X interpretation Con: Need to find sensible X
Explicit tradeoff of intensive/extensive margins	ATE for $m(y) = \begin{cases} \log(y) & y > 0 \\ -x & y = 0 \end{cases}$	Pro: Explicit tradeoff of two margins Con: Need to choose x; Monotone only if support excludes $(0, e^{-x})$
Intensive margin effect	$E\left[\log\left(\frac{Y(1)}{Y(0)}\right) \mid Y(1) > 0, Y(0) > 0\right]$	Pro: ATE in logs for the intensive margin Con: Partially identified

Table 3: Table 2 from Chen and Roth (2023)

Comment 2 (Poisson Regression)
Poisson regression is a good example of a way to estimate $E(Y_i(1) - Y_i(0))/E(Y_i(0))$. *This approach estimates* $\log(E(Y|X)) = X\beta$, *rather than* $E(\log(Y)|X)$. *You get a simple semi-elasticity measure for the parameters, and* Y *can be zero. What are the typical concerns?*

1. *If* $Y|X$ *is truly distributed Poisson, conditional on* X, *then* $Var(Y|X) = E(Y|X)$. *This just comes from the Poisson distribution's statistical properties, but feels like a restrictive model assumption. But, it's not relevant for the parameter estimates of* β. *The estimates are still consistent, and the standard errors for these estimates can be adjusted for misspecification using robust standard errors (e.g. sandwich covariance estimators). These will give correct coverage, obviating any concerns about the Poisson regression. It is* not *necessary to use a Negative Binomial regression.*

2. *As we will discuss shortly, in many non-linear models, if you include parameters, such as fixed effects, which cannot be consisitently estimated, then this will make* all *the estimates in the model inconsistent. This is different from linear models. This concern is less of an issue in Poisson regression, as fixed effects can be concentrated out (see* `PPMLHDFE` *in Stata and* `glmhdfe` *in R)*

3. *Individuals are often not sure how to do instrumental variables in Poisson regression, but it is feasible! See Mullahy [1997], Windmeijer and Santos Silva [1997].*

The benefits of using the Poisson model (instead of log(1+Y)) according to Cohn et al. [2022]: "We replicate data sets from six papers published in top finance journals that together study two count or count-like outcomes... We...estimate log1plus and Poisson regressions based on that specification, and compare the coefficients of interest. These coefficients differ markedly in all six cases and have different signs in three of the six, suggesting that inference about even the direction of a relationship is sensitive to regression model choice in real-world applications...in all five cases involving regressions with control variables, switching from a log1plus to Poisson regression results in a larger change in the coefficient of interest than omitting the most important control variable, generally by a wide margin."

Inconsistency in binary choice models

Consider estimating a panel fixed effects model with binary choice:

$$Y_{it} = \alpha_i + X_{it}\beta + \epsilon_{it}$$
$$Y_{it} = F(\alpha_i + X_{it}\beta)$$

where we are interested in the parameter β. If we have a short panel (e.g. few time periods), we cannot consistently estimate α_i. However, in the linear case, this does not affect estimation of β. More shockingly, however, is that for binary outcome case, the only model that consistently estimates β is a conditional Logit [Chamberlain, 1980, 2010].

More generally, if you have inconsistent fixed effects in your nonlinear models, this can cause serious issues (except in special cases like conditional). Often, the only way to get around these issues is by finding ways to "concentrate" or get around these nuisance parameters. Famous cases where this occurs include conditional logit, Poisson unit fixed effects, and partial likelihoods in the Cox proportional hazard model.

Multiple Choices

We'll now examine multiple discrete choice problems. Much of this discussion is very adjacent to industrial organization. However, many of these ideas are important for non-IO problems, such as multiple IVs and Roy models. Moreover, these tools are very promising in fields that have not yet used them.

Issues with choice problems that we'll discuss:

- Independence of Irrelevant Alternatives (IIA)
- Choice sets and consideration sets

Consider the following problem: we observe choices for individuals $Y_i = j$, $j \in \Omega = \{0, 1, \ldots, J\}$, where $J + 1 = |\Omega|$ is the total number of choices. Importantly, the order of the choices has no particular meaning. This could be red bus, blue bus and car as transportation choices, for example.

Given these sets of choices, we have different types of covariates we can observe. Some characteristics are choice specific (such as a price), while some are unit specific (such as a person's income). Often, we want to allow for the characteristics to vary by both dimensions. This includes allowing for a choice's characteristic to vary depending on the person (e.g. a unit specific coefficient on the choice's

characteristic), or allowing the person's characteristic to have differential effects on the choice of different goods. In total, we have three types of characteristics:

1. X_i (individual characteristics, invariant to choices),
2. X_j (choice characteristics)
3. X_{ij} includes individual-by-choice characteristics

We can write X_i as X_{ij} by interacting with choice fixed effects, and X_j can have i speicfic coefficients.[13]

[13] Note that when $J = 1$, we collapse down to binary choice.

Now recall there are two (non-exclusive) ways to think the discrete choice problem. The first is a statistical view: namely, how do we model the choice probabilities? In the binary choice problem, there is only one parameter that needs be known, conditional on X_i: $\pi(X_i) = Pr(Y_i = 1|X_i)$ With more than two choices, the dimensionality becomes more complicated. We now have $\pi_j(\mathbf{X}), j = 2, 3$ for 3 choices.

How should we parameterize how other choices' characteristics affect each other? Most of the models we will discuss will make very specific restrictions on how choices affect one another. These are not innocuous choices, as we'll see, but they provide a huge amount of additional structure that can be used to identify the parameters of interest.

The naive approach

If we want to estimate simple treatment effects, we could focus on binary outcomes. For exmaple: we have a randomly assigned treatment T, and J choices. What is the effect of T on $Pr(Y_i = j)$ under random assignment?

$$\tau_j = Pr(Y_i = j|T_i = 1) - Pr(Y_i = j|T_i = 0) \tag{4}$$

The downisde of this approach is that there's no information about the substitution patterns of individuals in this form. Concretely, if τ_2 is positive, is that because the share of individuals choosing $Y_i = 1$ decreases, the share of individuals choosing $Y_i = 0$ decreases, or both? Namely, what is the *substitution* pattern across the choices?[14]

[14] To put a statistical note on this, there are effectively two endogenous variables ($1(Y_i = 1)$ and $1(Y_i = 2)$), and we only have one randomly assigned variable (T). Hence, there's no way to simultaneously identify the effect on both.

Nonetheless, it is still very helpful to estimate these measures, and it's useful when faced with a lot of choices to focus on the effect on one margin. We will need more structure to estimate relative choice substitution across outcomes, and ask questions like "what is the effect of T on choosing j conditional on choosing j or k?"

Conditional logit

A second way to view the problem is as an structural (economic) choice problem (pioneered by McFadden [McFadden, 1972]). Consider a set of utilities U_{ij} (unobserved) such that

$$Y_i = \arg\max_{j \in \Omega} U_{ij}. \tag{5}$$

In other words, person i chooses j if it's the choice that maximizes the utility amongst all $J+1$ choices. Note the similarity to the Y_i^* in the binary case!

If we make the assumptions:

1. $U_{ij} = X_{ij}'\beta + \varepsilon_{ij}$
2. ϵ_{ij} are independent across choices and individuals, and distributed Type-I extreme value

then we get the McFadden conditional logit model:

$$Pr(Y_i = j|X_{ij}) = \frac{\exp(X_{ij}\beta)}{\sum_{k=0}^{J} \exp(X_{ik}\beta)}. \tag{6}$$

Comment 3
Note that if the characteristics X_{ij} only vary based on the individual (e.g. we can write $X_{ij}\beta$ as $X_i\beta_j$), then the effects across choices are relative *to each other. We can write our probability equation as*

$$Pr(Y_i = j|X_{ij}) = \frac{\exp(\alpha_j + X_i\beta_j)}{1 + \sum_{k=1}^{J} \exp(\alpha_k + X_i\beta_k)}. \tag{7}$$

This is the multinomial logit. *Once we allow for choice speicfic characteristics, then we need to write the probability following Equation (6).*

In many choice problems, a key parameter we're interested in is the price elasticity. The definition of the price elasticity is the percentage change in a market share of a good for a given percentage change in the price. Formally, the own-price elasticity is:

$$\epsilon_j = \frac{\partial Pr(Y_i = j|X_{ij})}{Pr(Y_i = j|X_{ij})} \frac{p_j}{\partial p_j} = \frac{\partial Pr(Y_i = j|X_{ij})}{\partial p_j} \frac{p_j}{Pr(Y_i = j|X_{ij})}. \tag{8}$$

We can *also* think about *cross*-price elasticities, e.g. how do market shares change when other goods shift their price:

$$\epsilon_{jk} = \frac{\partial Pr(Y_i = j|X_{ij})}{\partial p_k} \frac{p_k}{Pr(Y_i = j|X_{ij})}. \tag{9}$$

Note that with equation (6) as our probability model, we can estimate all these elasticities (assumign we have the data on prices, and we are willing to assume prices are exogeneous, a very strong assumption). But, this formulation creates issues.

A key issue with this formulation of the conditional logit model is that the cross-price elasticities are identical. Specifically, $\epsilon_{jk} = \epsilon_{lk}$, such that the effect of shifting price of a different good causes an identical proportionate shift in all choices' market share. You can see this by simply plugging in for $\frac{\partial Pr(Y_i=j|X_{ij})}{\partial p_k}$:

$$\begin{aligned}\epsilon_{jk} &= \underbrace{-\gamma Pr(Y_i = j|X_{ij})Pr(Y_i = k|X_{ij})}_{\frac{\partial Pr(Y_i=j|X_{ij})}{\partial p_k}} \times \frac{p_k}{Pr(Y_i = j|X_{ij})} \\ &= -\gamma Pr(Y_i = k|X_{ij})p_k,\end{aligned}$$

where γ is the coefficient on price in the conditional logit model. Note that this elasticity is not a function of j, and hence identical for all other products.[15]

[15] It's useful to note that the *levels* of the market share do vary by good, but the elasticity scaling makes the cross-price elasticities identical.

The canonical example of this is the "car, red bus and blue bus" example. Imagine a choice set where there are three choices for transportation: a car, and two busses: one red, and one blue. Presumably a person is purely indifferent between red and blue busses. Hence, a shift in the red bus price would presumably cause a bigger substitution from the blue bus than from car users, but the conditional logit (in this form) will not account for this.

How can we deal with the IIA issue? This is a problem of poor substitution patterns, which is an economics problem. In other words, economics gives us an intuition about the market substitution patterns, and we don't think identical cross-elasticities makes sense. It's also a statistical problem – there is a very strong statistical functional form we have assumed, which was analytically convenient but has somewhat perverse properties. We will now consider a few (but not all) solutions to the problem proposed in the literature.

Nested Logit and Correlated Multivariate Probit

One part of the IIA problem comes from the independence of ε across choices. Recall that the ε effectively rationalize the market shares beyond what we observe that is explained based on the covariates. Recall the blue and red bus case: getting two independent ε draws for the busses is not an intuitive view of bus demand. Instead, the blue and bus likely have highly correlated epsilon draws (if not identical), e.g. the unobserved latent demand for blue and red busses is correlated! The issue is exactly how to specify the correlation that preserves the ability to estimate the model.

With the nested Logit approach, you can specify sets (as the researcher), and allow correlation of the ε within these sets. The key is that the errors are uncorrelated across choice sets, which preserves the logit structure (see Goldberg [1995] for an example application), and the correlation *within* a nest is allowed to be correlated following a distinct similarity parameter. In essense, the similarlity parameter scales up and down the effect of the covariates within a nest: if the similarity is high, then the effect of the covariates is swamped by the random error, and the choices are highly correlated; if the similarity is low, the nest approaches the standard IIA setting. See Wen and Koppelman [2001] for a more recent discussion.

An alternative approach is to allow the covariance matrix of the error terms to be flexibly estimated by the data using a multivariate normal:

$$\epsilon_i = (\epsilon_{i0}, \epsilon_{i1}, \ldots, \epsilon_{iJ}) \sim \mathcal{N}(0, \Sigma) \tag{10}$$

where the researcher will then directly estimate Σ. Unfortunately, this problem gets hard with many choices (parameter space grows at rate $O(J^2)$). See McCulloch et al. [2000] and Geweke et al. [2003] for details and an application in the Bayesian setting, and Train (2009) for simulation discussions in the frequentist case.

Rather than directly target the distribution of the ε_{ij}, an alternative approach is to add more richness to the coefficients themselves. By adding more random variation in the loadings, it effectively creates a richer substitution pattern by adding more to the error term. Consider a slight extension of our previous model, with β_i varying by individual (in an unobserved way):[16]

$$\begin{aligned} U_{ij} &= X_{ij}\beta_i + \varepsilon_{ij} \\ U_{ij} &= X_{ij}\overline{\beta} + \nu_{ij}, \qquad \nu_{ij} = \varepsilon_{ij} + X_{ij}(\beta_i - \overline{\beta}) \end{aligned}$$

There are a number of ways to estimate this approach, but notice the key point – subtitution patterns are more richly modeled (and allowed) due to ν_{ij} varying by X_{ij}.

[16] Note that this random variation in preferences is usually viewed as *exogeneous*.

Example 2 (Random coefficients estimation example)
Let $J = 3$, and X_j be a scalar (e.g. price). We assume that

$$U_{ij} = X_j\beta_i + \varepsilon_{ij} \qquad \beta_i = (\overline{\beta} + \sigma\nu_i), \nu_i \sim \mathcal{N}(0,1). \tag{11}$$

Separate the utility of choosing j into

$$U_{ij} = \mu_{ij}(\overline{\beta}) + X_j\sigma\nu_i + \varepsilon_{ij} \tag{12}$$
$$\mu_{ij} = X_j\overline{\beta}. \tag{13}$$

We can write the probability of choosing j as:

$$Pr(Y_i = j|X, \overline{\beta}, \sigma) = \int \frac{\exp(X_j\overline{\beta} + X_j\sigma\nu_i)}{\sum_{k=0}^{J}\exp(X_j\overline{\beta} + X_j\sigma\nu_i)}\phi(\nu_i)d\nu_i \tag{14}$$

where $\phi(\cdot)$ is the Normal standard normal pdf.
This setup is often referred to as a "mixed logit" model (in contrast with the more common Berry Levinsohn Pakes approach, which we'll discuss later) [McFadden and Train, 2000]. The typical approach for estimating these models involves using Maximum Simulated Likelihood, or Method of Simulated Moments. McFadden and Train [2000] show that a straightforward approach to estimating this is to simulate the model S times, and then use the simulated data to approximate the integral:

$$\hat{E}(Pr(Y_i = j|X, \overline{\beta}, \sigma)) = \frac{1}{S}\sum_{s=1}^{S}\frac{\exp(X_j\overline{\beta} + X_j\sigma\nu_{is})}{\sum_{k=0}^{J}\exp(X_j\overline{\beta} + X_j\sigma\nu_{is})}. \tag{15}$$

Then, this probability can be used to form a log-likelihood function, and the model can be estimated using standard optimization techniques for maximizing log-likelihoods.
Note that an important piece in this setting is micro-level choice data (which we use to form the likelihood), and the lack of any unobserved heterogeneity that creates endogeneity and bias in our estimates. Without an additional error term, there's no need for an instrument here. This is a version of assuming exogeneity conditional on observables. Often, we will only observe market-level *shares of goods. Then, we'll need many markets in order to have sufficient independent variation to estimate parameters. We will discuss this next.*

The workhorse set of demand estimation models is known as BLP (Berry Levinsohn Pakes), named after the authors in Berry et al. [1995]. This model combines random coefficient estimation with unobserved market-good-level demand heterogeneity that is potentially endogeneous and correlated with price. In other words, not only are

individuals allowed to have random (independent) error, but there is a fixed unobserved error in demand for each good. This allows for a highly correlated set of demand choices within a market, and also creates unobserved demand heterogeneity that requires an instrument.

This model is often specified using the following utility function:

$$U_{ijm} = \delta_{jm} + \mu_{ijm} + \epsilon_{ijm}, \tag{16}$$

where $\delta_{jm} = X_j\beta + \xi_{jm}$ is the mean utility of choosing j in market m, μ_{ijm} is the random substition pattern specific to an individual (typically driven by the random coefficients on good characterstics as in Example 2), and ϵ_{ijm} is the individual specific logit error that is i.i.d. Often, this type of setting is used when only market-level data is available, and so the researcher observes the market shares of goods, but not the individual choices.[17]

[17] This is a common setting in many IO applications, where the researcher observes the market shares of goods, but not the individual choices. However, it's wonderful when you have more, and a host of papers using the micro data exist as well [Berry et al., 2004, Conlon and Gortmaker, 2023].

Under the standard logit distributional assumptions for ϵ_{ijm},

$$Pr(Y_{im} = j|X) = s_{jm}(\boldsymbol{\delta}_m, \theta) = \int \frac{\exp(\delta_{jm} + \mu_{ijm})}{\sum_{k \in J_m} \exp(\delta_{km} + \mu_{ikm})} f(\boldsymbol{\mu}|\theta) d\boldsymbol{\mu}_{im}. \tag{17}$$

The key insight in Berry et al. [1995] is to note that the vector of δ_{jm} in market m, δ_m, can be inverted from the market shares, s_m, and θ, the parameters of the random mixing coefficients. Once we know δ_m, we can define $\xi_{jm} \equiv \delta_{jm} - X_j\beta$, and define a conditional moment condition $E(\xi_{jm}|Z_{jm}) = 0$. This moment condition can be used to estimate β using GMM. Conlon and Gortmaker [2020] provide a very nice discussion of the algorithmic approach on how to do this, and provide a Python package to solve this problem.[18]

[18] Part of the reasoning for this is that the trick to invert the shares and recover δ_{jm} is a non-linear fixed point problem that needs to converge to a high degree of precision for successful estimation. Conlon and Gortmaker [2020] highlight the best approaches.

Conclusion

Underlying structure of discrete choice is valuable in IV settings. Much of this discussion centered on IO style applications. But this discussion shows up when thinking about Roy style models.[19] When we discuss instruments and individuals' choice to take up a policy or not, if the policy is multi-dimensional, this types of models play a huge role. Recall our discussion of propensity scores for treatment effects. If individuals choose between multiple treatment options, this maps directly into a discrete choice setting like what we've discussed today. Thinking carefully about the counterfactual pattern across will give guidance in more complicated IV settings.

[19] See Hull [2018] for an example.

There is also value in arbitraging IO methods in other fields. Many fields have discrete choice applications but have not adopted the tools. The cutting edge of IO tools is quite complex, but this type of

structure is very valuable when thinking about complicated choice patterns. Worthwhile to try to arbitrage these methods in fields that are less exposed to them (e.g. Koijen and Yogo [2019]).

References

Ghislain BD Aihounton and Arne Henningsen. Units of measurement and the inverse hyperbolic sine transformation. *The Econometrics Journal*, 24(2):334–351, 2021.

Marc F Bellemare and Casey J Wichman. Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics*, 82(1):50–61, 2020.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995.

Steven Berry, James Levinsohn, and Ariel Pakes. Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of political Economy*, 112(1):68–105, 2004.

Gary Chamberlain. Analysis of covariance with qualitative data. *The review of economic studies*, 47(1):225–238, 1980.

Gary Chamberlain. Binary response models for panel data: Identification and information. *Econometrica*, 78(1):159–168, 2010.

Jiafeng Chen and Jonathan Roth. Logs with zeros? some problems and solutions. *The Quarterly Journal of Economics*, page qjad054, 2023.

Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics*, 129(4):1553–1623, 2014.

Jonathan B Cohn, Zack Liu, and Malcolm I Wardlaw. Count (and count-like) data in finance. *Journal of Financial Economics*, 146(2): 529–551, 2022.

Christopher Conlon and Jeff Gortmaker. Best practices for differentiated products demand estimation with pyblp. *The RAND Journal of Economics*, 51(4):1108–1161, 2020.

Christopher Conlon and Jeff Gortmaker. Incorporating micro data into differentiated products demand estimation with pyblp. Technical report, NYU working paper, 2023.

John Geweke, Gautam Gowrisankaran, and Robert J Town. Bayesian inference for hospital quality in a selection model. *Econometrica*, 71 (4):1215–1238, 2003.

Pinelopi Koujianou Goldberg. Product differentiation and oligopoly in international markets: The case of the us automobile industry. *Econometrica: Journal of the Econometric Society*, pages 891–951, 1995.

Peter Hull. Estimating hospital quality with quasi-experimental data. *Available at SSRN 3118358*, 2018.

Ralph SJ Koijen and Motohiro Yogo. A demand system approach to asset pricing. *Journal of Political Economy*, 127(4):1475–1515, 2019.

Willard G Manning and John Mullahy. Estimating log models: to transform or not to transform? *Journal of Health Economics*, 20(4):461–494, 2001. ISSN 0167-6296. DOI: https://doi.org/10.1016/S0167-6296(01)00086-8. URL https://www.sciencedirect.com/science/article/pii/S0167629601000868.

Robert E McCulloch, Nicholas G Polson, and Peter E Rossi. A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of econometrics*, 99(1):173–193, 2000.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. 1972.

Daniel McFadden and Kenneth Train. Mixed mnl models for discrete response. *Journal of applied Econometrics*, 15(5):447–470, 2000.

John Mullahy. Instrumental-variable estimation of count data models: Applications to models of cigarette smoking behavior. *Review of Economics and Statistics*, 79(4):586–593, 1997.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384, 1972.

Martin Ravallion. A concave log-like transformation allowing non-positive values. *Economics Letters*, 161:130–132, 2017.

Chieh-Hua Wen and Frank S Koppelman. The generalized nested logit model. *Transportation Research Part B: Methodological*, 35(7): 627–641, 2001.

Frank AG Windmeijer and Joao MC Santos Silva. Endogeneity in count data models: an application to demand for health care. *Journal of applied econometrics*, 12(3):281–294, 1997.