Qualitative Response Models

Michael R. Roberts

Department of Finance
The Wharton School
University of Pennsylvania

January 21, 2009

Discrete Choice Framework

- The choice set must exhibit 3 characteristics:
  1. Alternatives must be mutually exclusive.
  2. Choice set must be exhaustive.
  3. # of alternatives must be finite.
- 1 and 2 can usually be satisfied with appropriate classifications.
- 3 is the defining feature of discrete choice models.

Random Utility Models

- Decision maker (agent, firm, person, etc.) i faces J alternatives (alts).
- Decision maker obtains a certain utility or profit from each alt
- Utility that agent i obtains from alt j is U_{ij}.
- Agent chooses alt that provides highest utility $U_{ij} > U_{ik} \forall j \neq k$.

Empirical Implimenation

- Utility for agent i, alternative j, U_{ij}, decomposed into two components:
  1. $V(x_{ij}, s_j, \theta)$ = **Indirect Utility** observed by researcher. Function of alternative attributes x_{ij}, agent attributes s_i, and parameters θ.
  2. ε_{ij} = unobservable *to researcher* factors affecting utility. Defined relative to reseracher's representation of choice situation (i.e., V).
- Unobserved components, ε_{ij}, are assumed random according to a distribution $f(\varepsilon_{ij})$:
- Unobserved vector of errors across alternatives is described by joint density $f(\varepsilon_i)$:

$$\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{iJ}) \sim f(\varepsilon_i)$$

Probability of Agent's Choice

- With f we can make probabilistic statements about agent's choice

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

- This last expression is a CDF

$$P_{ij} = \int_{\varepsilon} I(\varepsilon_{ik} - \varepsilon_{ij} > V_{ij} - V_{ik}, \forall j \neq k) f(\varepsilon_i) d\varepsilon_i$$

  This is a multidimensional (J) integral over the domain of ε_i (e.g., $\mathbb{R}^J$). **Note**: Independence implies $f(\varepsilon_i) = f(\varepsilon_{i1}) \times \cdots \times f(\varepsilon_{iJ})$.
- Different distributions $\Longrightarrow$ different models.
  1. ε_i i.i.d. extreme value $\Longrightarrow$ **logit** (Closed Form)
  2. ε_i i.i.d generalized extreme value $\Longrightarrow$ **nested logit** (Closed Form)
  3. ε_i multivariate normal $\Longrightarrow$ **probit**

Model Identification

- **Only Differences in Utility Matter** or **The level of utility doesn't matter**. The choice probability is:

$$\begin{aligned} P_{ij} &= Pr\left(U_{ij} > U_{ik}, \forall j \neq k\right) \\ &= Pr\left(U_{ij} - U_{ik} > 0, \forall j \neq k\right) \end{aligned}$$

  which depends only on the difference in utility not its absolute level. Similarly,

$$\begin{aligned} P_{ij} &= Pr\left(V_{ij} + \varepsilon_{ij} > V_{ik} + \varepsilon_{ik}, \forall j \neq k\right) \\ &= Pr\left(\varepsilon_{ij} - \varepsilon_{ik} > V_{ik} - V_{ij}, \forall j \neq k\right) \end{aligned}$$

  which also just depends on differences.
- In general, the only parameters that can be estimated are those that capture differences across alternatives.

Alternative-Specific Constants

- Assume

$$V_{ij} = x_{ij}\beta + k_j, \forall j$$

  k_j captures average effect on utility of all factors not in model (like intercept in linear regression).
- Including k_j forces ε_{ij} to have zero-mean
- "Only differences matter" $\implies$ only differences $k_j - k_k$ matter, not absolute level of each. Normalize one of the constants to zero.
- With J alternatives, $J - 1$ constants can be estimated.

Sociodemographic Variables

- Consider choosing between commuting via bus or car

$$\begin{aligned} U_c &= \alpha T_c + \beta M_c + \theta_c Y + \varepsilon_c \\ U_b &= \alpha T_b + \beta M_b + \theta_b Y + k_b + \varepsilon_c \end{aligned}$$

  where T=commute time, M=commute cost, Y=income.
- We can only estimate differences: $\theta_c - \theta_b$ (or vice versa).
- So, either normalize one θ to 0

$$\begin{aligned} U_c &= \alpha T_c + \beta M_c + \varepsilon_c \\ U_b &= \alpha T_b + \beta M_b + \theta_b Y + k_b + \varepsilon_c \end{aligned}$$

  where $\theta_b = \theta_b - \theta_c$, or interact alternative-specific variables

$$\begin{aligned} U_c &= \alpha T_c + \beta M_c / Y + \varepsilon_c \\ U_b &= \alpha T_b + \beta M_b / Y + \theta_b Y + k_b + \varepsilon_c \end{aligned}$$

Independent Error Terms

- Recall choice probability is J-dimensional integral:

$$P_{ij} = \int_{\varepsilon} I(\varepsilon_{ik} - \varepsilon_{ij} > V_{ij} - V_{ik}, \forall j \neq k) f(\varepsilon_i) d\varepsilon_i$$

- Can write in terms of $J-1$-dimensional integral

$$\begin{aligned} P_{ij} &= Pr(\tilde{\varepsilon}_{ijk} > V_{ik} - V_{ij}, \forall j \neq k) \\ &= \int_{\tilde{\varepsilon}} I(\tilde{\varepsilon}_{ijk} > V_{ij} - V_{ik}, \forall j \neq k) g(\tilde{\varepsilon}_{ij}) d\tilde{\varepsilon}_{ij}, \end{aligned}$$

1. $\tilde{\varepsilon}_{ijk} = \varepsilon_{ij} - \varepsilon_{ik}$ = difference in errors for alt's j and k
2. $\tilde{\varepsilon}_{ij} = (\tilde{\varepsilon}_{ij1}, ..., \tilde{\varepsilon}_{ijJ}) = J-1$-dim vector of error differences over all alternatives *except* j.
3. $g(\tilde{\varepsilon}_{ij})$ is the $J-1$-dimensional density of error differences.

- Since choice prob's can be expressed in terms of $g(\tilde{\varepsilon}_{ij})$, one dimension of $f(\varepsilon_i)$ is not identified and must be normalized.

Scale of Utility is Irrelevant

- Multiplying by a *positive* constant doesn't affect choice. Following two models are equivalent $\forall \lambda > 0$:

$$\begin{aligned} U_{ij}^0 &= V_{ij} + \varepsilon_{ij}, \forall j \\ U_{ij}^1 &= \lambda V_{ij} + \lambda \varepsilon_{ij}, \forall j \end{aligned}$$

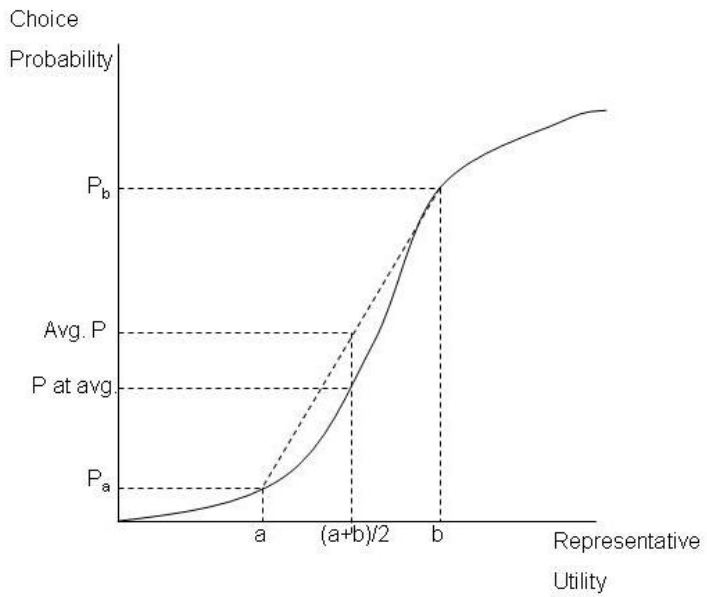
- Address by normalizing the variance of error terms.
  1. **i.i.d. errors**: The two models are equiv

$$\begin{aligned} U_{ij}^0 &= x_{ij}\beta + \varepsilon_{ij}^0, V(\varepsilon_{ij}^0) = \sigma^2 \\ U_{ij}^1 &= x_{ij}(\beta/\sigma) + \varepsilon_{ij}^1, V(\varepsilon_{ij}^1) = 1 \end{aligned}$$

     Normalizing constant important when comparing coeffs across models (e.g., probit and logit) or datasets where scale varies.
  2. **Heteroskedastic Errors**: Normalize one of the variances $\Longrightarrow$ normalize variance of error difference.
  3. **Correlated Errors**: Normalize the variance of one of the error differences.

Average Response

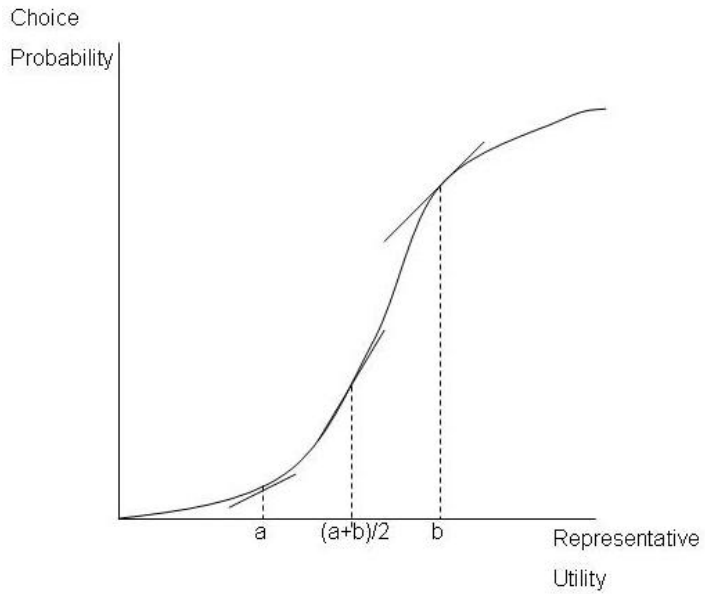
- *Linear* model $f \implies E[f(x)] = f(E[x])$ so we can insert aggregate or average values into model, e.g., $\bar{y} = \alpha + \beta\bar{x}$
- *Nonlinear* models $f : E[f(x)] \neq f(E[x])$



- Prob at avg utility can over- or under-estimate depending on where individual choice prob's are (convex or concave portion of curve).

Average Marginal Effects

- Derivative is small *a* and *b*, derivative of avg. large.



- **Solution**: To get aggregate outcome, average indiv probs. To get at average marginal effect, avg indiv MEs (APE).

A Regression Perspective

- Consider binary case: two mutually exclusive outcomes captured by $Y \in \{0, 1\}$

$$\begin{aligned} Pr(Y = 1|x) &= F(x, \beta) \\ Pr(Y = 0|x) &= 1 - F(x, \beta) \end{aligned}$$

- Assume $F(x, \beta) = x'\beta$.

$$E(Y|x) = 0 * Pr(Y = 0|x) + 1 * Pr(Y = 1|x) = Pr(Y = 1|x) = x'\beta$$

so the regression model is:

$$y = E(Y|x) + (y - E(Y|x)) = x'\beta + \varepsilon$$

Two Problems

1. Heteroskedastic errors

$$\begin{aligned} V(\varepsilon|x) &= E(\varepsilon^2|x) + (E(\varepsilon|x)^2) \\ &= E((y - x'\beta)^2|x) \\ &= E(y^2 - 2yx'\beta + (x'\beta)^2|x) \\ &= E(y^2|x) - E(y|x)2x'\beta + (x'\beta)^2 \\ &= x'\beta - 2(x'\beta)^2 + (x'\beta)^2 \\ &= x'\beta(1 - x'\beta) \end{aligned}$$

   (FGLS solves this.)
2. Predicted values not constrained to $[0, 1]$ implies
   - nonsense probabilities
   - negative variances

Solution to Unbounded Predictions

- Choose F:

$$\begin{aligned} lim_{z\to\infty} F(z) &= 1 \\ lim_{z\to-\infty} F(z) &= 0 \end{aligned}$$

- Examples:
  - $F(z) = \Phi(z) = \int_{-\infty}^{z} \phi(t)dt$ (Probit - symmetric)
  - $F(z) = \Lambda(z) = \frac{e^z}{1+e^z}$ (Logit - symmetric)
  - $F(z) = G(z) = exp[-exp(-z)]$ (Gumbel - asymmetric))
  - $F(z) = L(z) = 1 - exp[exp(z)]$ (Complementary Log Log - asymmetric))

  where ϕ is the standard normal density $(1/2\pi)^{0.5} exp(-0.5t^2)$.
- Little guidance on choice
- Asymmetry refers to ε not Y

Index Model

- Consumer weighs *unobservable* Marginal Costs and Benefits of decision

$$y^* = x'\beta + \varepsilon$$

where $x'\beta$ is **index function** and observation equation is

$$y = 1 \text{ if } y^* > 0$$
$$y = 0 \text{ if } y^* \leq 0$$

- Note:
  1. Variance of ε is unidentified (data depend only on sign)

$$y^* = x'\beta + \sigma\varepsilon \iff (y^*/\sigma) = x'(\beta/\sigma) + \varepsilon$$

  2. Zero threshold is irrelevant. Consider threshold a

$$Pr(y^* > a|x) = Pr((\alpha - a) + x'\beta + \varepsilon > 0|x)$$

- See Brock and Durlauf (2000) for examples.

Specification Concerns

- Yatchew and Griliches (1984) show that unlike *linear* regression,
  1. Ommitted variables: even if the ommitted variable is uncorrelated with the included ones, the estimated coefficient will be inconsistent
  2. Heteroskedastic regressors result in inconsistent MLEs and inapproriate covariance matrix
- MLE is also inconsistent if
  1. there is unmeasured heterogeneity
  2. the functional form of the index is nonlinear but assume linear
  3. the distributional assumption (e.g., normal, logit) is incorrect
- **Punchline**: specification errors are more serious in nonlinear setting

Error Distibution

- Model same as before: $U_{ij} = V_{ij} + \varepsilon_{ij}, \forall j$.
- Logit assumes that ε_{ij} i.i.d. extreme value (a.k.a. Gumbel, Type I Extreme Value) across agents i and alternatives j.
  - **Density** is

$$f(\varepsilon_{ij}) = e^{-\varepsilon_{ij}} e^{-e^{-\varepsilon_{ij}}}$$

  - **CDF** is

$$F(\varepsilon_{ij}) = e^{-e^{-\varepsilon_{ij}}}$$

  - **Variance** of this distribution is $\pi^2/6$, which implicity normalizes the **scale of utility**.
  - **Mean** $\neq 0$ But, irrelevant since only differences in utility matter and difference of two random vars with same mean is 0.

Error Difference Distribution

- If ε_{ij} i.i.d. extreme value, then $\tilde{\varepsilon}_{ijk} = \varepsilon_{ij} - \varepsilon_{ik}$ is logistic:
  - **CDF** is

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

    This is distribution for binary logit. Similar to normal but fatter tails.
- Key assumption is independence of errors. Ok if model is "well-specified" If errors are correlated then
  1. use different model to allow for corr
  2. respecify representative utility V to capture corr
  3. consider model only as approximation
- Violation of indep assumption less important for estimating avg preferences, than for forecasting substitution patterns.

Logit Choice Probabilities I

- McFadden (1974)

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

- Fix ε_{ij}'s, this is CDF of ε_{ik}'s evaluated at $V_{ij} - V_{ik} + \varepsilon_{ij}$
- Indep of ε's $\Longrightarrow$ CDF i

$$P_{ij}|\varepsilon_{ij} = \prod_{k \neq j} e^{-e^{-(\varepsilon_{ij} + V_{ij} - V_{ik})}}.$$

This is conditional joint CDF of all ε's except ε_{ij}, the conditioning variable.

Logit Choice Probabilities II

- To get unconditional density, integrate out ε_{ij}

$$P_{ij} = \int \left(\prod_{k \neq j} e^{-e^{-(\varepsilon_{ij} + V_{ij} - V_{ik})}} \right) e^{-\varepsilon_{ij}} e^{-e^{-\varepsilon_{ij}}} d\varepsilon_{ij} = \frac{e^{V_{ij}}}{\sum_j e^{V_{ij}}}.$$

- If representative utility V is linear in parameters then

$$P_{ij} = \frac{e^{x_{ij}\beta}}{\sum_j e^{x_{ij}\beta}}.$$

- Choice probs sum to 1.
- Relation between P and V is sigmoid shaped
  - Point at which increase in V has largest effect on prob of alternative being chosen is when prob ≈ 0.5. Small changes tip the balance.

Coefficients

- Consider choice between gas and electric heating system

$$\begin{aligned} U_g &= \beta_1 PP_g + \beta_2 OC_g \\ U_e &= \beta_1 PP_e + \beta_2 OC_e \end{aligned}$$

  where PP =purchase price, OC =operating cost.
- The sign of the coefficients indicate the effect on utility (expect β_1 and $\beta_2 < 0$).
- Ratio of coefficients, β_2/β_1 is willingness to pay for operating-cost reductions. E.g., $\beta_2 = -1.14$ and $\beta_1 = -0.20 \implies \beta_2/\beta_1 = \$5.70$. Pay \$5.70 more for a system whose annual OCs are \$1 less.
- Take total deriv of utility and set $= 0$.

$$dU = \beta_1 dPP + \beta_2 dOC = 0 \implies \partial PP/\partial OC = -\beta_2/\beta_1$$

Multinomial Logit

- Reconsider choice probabilities on previous slide

$$\begin{aligned} P_g &= \frac{e^{\beta_1 \beta_1 PP_g + \beta_2 OC_g}}{e^{\beta_1 \beta_1 PP_g + \beta_2 OC_g} + e^{\beta_1 PP_e + \beta_2 OC_e}} \\ &= \frac{1}{1 + e^{(\beta_1 \beta_1 PP_g + \beta_2 OC_g) - (\beta_1 PP_e + \beta_2 OC_e)}} \end{aligned}$$

- Consider a third option, oil heating

$$P_g = \frac{e^{\beta_1 \beta_1 PP_g + \beta_2 OC_g}}{e^{\beta_1 \beta_1 PP_g + \beta_2 OC_g} + e^{\beta_1 PP_e + \beta_2 OC_e} + e^{\beta_1 PP_o + \beta_2 OC_o}}$$

- In binomial or multinomial, if we have variables constant across alternatives, must normalize one of the alternative-specific coefficients to zero.

Scale Invariance

- Logit assumes type 1 extreme value with variance $\pi^2/6$.
- Utility is $U_{ij}^* = V_{ij} + \varepsilon_{ij}^*$, where $Var(\varepsilon_{ij}^*) = \sigma^2(\pi^2/6)$.
- Scale irrelevance $\implies$ divide by σ without changing behavior

$$U_{ij} = V_{ij}/\sigma + \varepsilon_{ij}$$

  where $\varepsilon_{ij} = \varepsilon_{ij}/\sigma$ and $V(\varepsilon_{ij}) = \pi^2/6$.
- The choice prob is:

$$P_{ij} = \frac{e^{V_{ij}/\sigma}}{\sum_j e^{V_{ij}/\sigma}} = \frac{e^{(\beta/\sigma)x_{ij}}}{\sum_j e^{(\beta/\sigma)x_{ij}}}$$

- Estimated coeffs = orig params scaled by σ $\implies$ careful when interpreting magnitudes (Ben-Akiva & Morikawa '90, Swait & Louviere '93)
- Can't identify scale σ $\implies$ normalize to 1.

Applicability of Logit Models

1. Logit can represent *systematic* (i.e., related to observables) taste variation, not *random* (i.e., unrelated to observables) taste variation
2. Logit implies proportional substitution across alternatives (i.i.a.)
3. Logit can handle state dependence and repeated choice but can't handle serially correlated errors.

Taste Variation - Observables

- Consider car choice

$$U_{ij} = \alpha_i SR_j + \beta_i PP_j + \varepsilon_{ij}$$

where SR = shoulder room=, PP = purchase price.

- Note parameters (α_i, β_i) are agent-specific. E.g.,

$$\alpha_i = \rho M_i, \beta_i = \theta / I_i$$

where $M = \#$ of family members, I = income. Then

$$U_{ij} = \rho(M_i SR_j) + \theta(PP_j / I_i) + \varepsilon_{ij}$$

- As long as taste varies with **observables**, no problem

Taste Variation - Unobservables

- Now let parameters (α_i, β_i) vary with unobservables

$$\alpha_i = \rho M_i + \mu_i, \beta_i = \theta / I_i + \eta_i$$

where $\mu_i =$ unobserved to researcher value of shouler room (e.g., size of people, frequency of traveling together), and $\eta_i =$ unobserved to researcher value of purchase price. Now

$$U_{ij} = \rho(M_i SR_j) + \theta(PP_j / I_i) + \tilde{\varepsilon}_{ij}$$

where $\tilde{\varepsilon}_{ij} = \mu_i SR_j + \eta_i PP_j + \varepsilon_{ij}$.

- New error term *can't* be i.i.d.
  1. μ_i and $\eta_i \implies$ corrrelation across alts.
  2. SR_j and $PP_j \implies$ heteroskedasticity across alts.
- Solution: Probit or Mixed Logit.

Substitution Patterns

- Increase in prob of choosing one alt $\Longrightarrow$ decrease in prob of choosing other alternatives (probs sum to 1).
- Logit model implies a specific substitution pattern. Can be seen as
  1. restriction on the ratios of probs, and/or
  2. restriction on the cross-elasticities of probs.

Independence of Irrelevant Alternatives (IIA)

- Ratio of logit probs for any 2 alts

$$\frac{P_ij}{P_ik} = \frac{e^{V_{ij}} / \sum_j e^{V_{ij}}}{e^{V_{ik}} / \sum_j e^{V_{ij}}} = \frac{e^{V_{ij}}}{e^{V_{ik}}} = e^{V_{ij}} - e^{V_{ik}}$$

- The ratio does **not** depend on any other alts.
- IIA can be viewed as a property of a properly specified model.

IIA: Red Bus-Blue Bus

- Travel by car or blue bus: $P_c = P_{bb} = 1/2 \implies P_c/P_{bb} = 1$.
- Introduce identical red bus, $\implies P_{rb} = P_{bb} \implies P_{rb}/P_{bb} = 1$.
- iia $\implies P_c/P_{bb} = 1$ since new alt has no effect on ratio.
- Only probs: $P_c/P_{bb} = 1$ and $P_c/P_{bb} = 1$ are $P_c = P_{bb} = P_{rb} = 1/3$.
- In real life, this is silly. P_c shouldn't change and $P_{rb} = P_{bb}$ if they're identical. So, $P_c = 1/2$ and $P_{rb} = P_{bb} = 1/4$
- IIA leads to overest of bus, underest of car.

Proportional Substitution

- Same idea in terms of cross-elasticities of logit probs.
- How does Δ in characteristics of alt j affect all prob of *other* alts?
- Elasticity of P_{ij} wrt variable in representative utility of alt k (see below for deriv):

$$\frac{\partial P_{ij}}{\partial z_{ik}} = E_{iz_{ik}} = -\beta_z z_{ik} P_{ik}$$

- Note cross-elasticity is same $\forall j$ since j does not enter the formula.
- Means improvement in attributes of 1 alt reduces probs $\forall$ other alts by same %.
- I.e., improvement in 1 alt draw proportionately from all other alts. (a.k.a., **proportionate shifting** and is manifest of iia).

Advantages & Tests of IIA

- With lots of alts, we can focus on subset.
- Two types of tests:
  1. Hausman and McFadden (1984): Parameter ests from a subset of alts are same as parameter ests from full set of alts.
  2. McFadden (1987) and Train et al. (1989)

Panel Data

- If unobserved factors affecting agents are independent over repeated choices logit is kosher.
- Dynamics related to observed factors easily accomodated:
  1. State dependence where agent's past choices affect current choice, or

$$V_{ijt} = \alpha y_i j(t-1) + \beta x_{ijt}, \text{ or}$$

$$V_{ijt} = \alpha \sum_{0}^{t} y_{ij(s)} + \beta x_i jt,$$

  where $y_{ij(t)} = 1$ if alt j chosen in period t.

  2. Lagged response to changes in attributes.

$$V_{ijt} = \beta x_{ij(t-1)}, \text{ or}$$

$$V_{ijt} = \beta x_{ij(t-1)} + \beta_2 x_{ij(t-2)} + \ldots,$$

- Dynamics related to unobserved factors cannot be handles because of indep assumption.

Dynamic Models

$$y_{it} = I(x_{it}'\beta + \alpha_i + \gamma y_{i,t-1} + \varepsilon_i t > 0)$$

- **Lagged effects** or **persistence** arises from three souces:
  1. serial correlation in ε
  2. **heterogeneity**, α_i (some individuals are more inherently more likely to choose or experience event for all time)
  3. **state dependence**, $y_{i,t-1}$ (occurence of past event or decision influences prob of current event or decision)

  (see Heckman (1978,1981))
- Example: Choice of restaurant
  1. Someone barfed outside Taco Bell yesterday, turning me off
  2. I don't like fast food
  3. I ate at Taco Bell yesterday and don't want the same thing twice
- Intial conditions have big impact on entire path in short panels

Consumer Surplus

- Agent's consumer surplus is defined as the utility, in \$, that the person receives from a choice situation.
- Consumer Surplus is:

$$CS_i = (1/\alpha_i) max_j(U_{ij} \forall j)$$

where $\alpha_i = dU_i/Y_i$ is marginal utility of income, and Y_i is income of agent i. (Division by MU inc translates utility into \$.)

- Observe V, not U so we can compute expected CS:

$$E(CS_i) = (1/\alpha_i) E_{\varepsilon_{ij}}[max_j(U_{ij} \forall j)]$$

- If U is linear in income, then α_i is constant wrt income &

$$E(CS_i) = (1/\alpha_i) \underbrace{\ln\left(\sum_{j=1}^{J} e^{V_{ij}}\right)}_{\text{Log-Sum Term}} + \underbrace{C}_{\text{Unknown Constant}}$$

Policy Analysis

- $E(CS_i)$ = avg CS in subpop of people with same V's as person i.
- Total CS in population is weighted sum of $E(CS_i)$ over sample
- Δ in CS due to change in alternatives is:

$$E(CS_i) = (1/\alpha_i)\left[\ln\left(\sum_{j=1}^{J^1} e^{V_{ij}^1}\right) - \ln\left(\sum_{j=1}^{J^0} e^{V_{ij}^0}\right)\right]$$

  where 0 and 1 superscripts refer to pre- and post-change
- To get α_i, can use coef on a price or cost variable. E.g., cost coef, β, should be < 0. $-\beta$ is amount utility rises due to \$1 inc in dec in cost, which is equiv to \$1 inc in income since person can spend dollar saved just as if he received an extra \$1 in income. $-\beta$ is inc in utility from \$1 inc in income (i.e., MU of inc).
- If MU income function of income see (e.g., McFadden (1999), Karlstrom (2000)).

Own Alternative Derivatives

- Change in prob that agent i chooses alt j given change in observed factor z_{ij} entering that alt is

$$\begin{aligned}\frac{\partial P_{ij}}{\partial z_{ij}} &= \frac{\partial \left(e^{V_{ij}} / \sum_k e^{V_{ik}}\right)}{\partial z_{ij}} \\ &= \frac{\partial V_{ij}}{\partial z_{ij}} P_{ij}(1 - P_{ij})\end{aligned}$$

- If V linear in parms then $\partial V_i j / \partial z_{ij} = \beta$
- Derivative is largest when $P_{ij} = 0.5$ and smaller when P_{ij} approaches 0 or 1.
- Intuition: Change matters most when choice probs indicate high degree of uncertainty.

Cross Alternative Derivatives

- Change in prob that agent i chooses alt j given change in observed factor z_{ik} entering a *different* alt is

$$\begin{aligned}\frac{\partial P_{ij}}{\partial z_{ik}} &= \frac{\partial \left(e^{V_{ij}} / \sum_k e^{V_{ik}}\right)}{\partial z_{ik}} \\ &= -\frac{\partial V_{ik}}{\partial z_{ik}} P_{ij} P_{ik}\end{aligned}$$

- If V linear in parms then $\partial V_i k / \partial z_{ik} = \beta$
- If z is desirable so that $\beta > 0$, then raising z increases prob of that alt but lowers prob of other alts.

Elasticities

- Elasticity of P_{ij} wrt z_{ij} entering the utility of alt i is:

$$\begin{aligned} E_{jz_{ij}} &= \frac{\partial P_{ij}}{\partial z_{ij}} \frac{z_{ij}}{P_{ij}} \\ &= \frac{\partial V_{ij}}{\partial z_{ij}} z_{ij}(1 - P_{ij}) \end{aligned}$$

- Elasticity of P_{ij} wrt z_{ik} entering the utility of alt k is:

$$\begin{aligned} E_{jz_{ik}} &= \frac{\partial P_{ij}}{\partial z_{ik}} \frac{z_{ik}}{P_{ij}} \\ &= -\frac{\partial V_{ik}}{\partial z_{ik}} z_{ik} P_{ik} \end{aligned}$$

- Cross-elasticity is same $\forall i \implies$ changing attribute of alt j changes prob $\forall$ other alts by same %.

Derivatives and Elasticities: Details

- Derivatives and elasticities are fxn of data (that's the P_ij, the i subscript) $\Longrightarrow$ choose value (e.g., mean, median). But, this choice can have big impact on deriv/elast.
- APE = average ME over each observation. Need to estimate SE and valid only in "large" samples.
- For dummy variables, d, compute differnce in probabilities at 0 and 1 holding all other variables fixed (at mean, median, etc.)
- Can use this technique for continuous variables to estimate change in prob for a large movement in x (e.g., 25th to 75th percentile, 1 SD below mean to 1 SD above, etc.)
- Do *not* multiply marginal effect - valid locally - times SD - big change.

Maximum Likelihood

- Assume random sample and exogenous explanatory vars.
- Probability of agent i choosing alt he actually choose is:

$$\prod_j (P_{ij})^{y_{ij}},$$

where $y_{ij} = 1$ if person i chose alt j, 0 otherwise. This reduces to prob of chosen alt since $y_{ij} = 0 \forall$ alts other than chosen one.

- (Log) Likelihood Function:

$$\prod_{i=1}^{N} \prod_j (P_{ij})^{y_{ij}}$$

$$\sum_{i=1}^{N} \sum_j y_{ij} ln P_{ij}$$

MLE Interpretations

- Can interpret the MLEs in several ways:
  1. Can be shown that MLEs of βs are those that make the predicted avg of each explanatory variable $=$ to observed avg in sample.
  2. This property $\Longrightarrow$ MLEs of alt-specific constants $=$ share of agents who choose alt. I.e., predicted shares $=$ actual shares.
  3. MLEs are values βs that make residuals $(y_{ij} - P_{ij})$ uncorrelated with explanatory variables.

Binary Case: Likelihood Function

- Recall log likelihood: $\sum_{i=1}^{N}\sum_{j} y_{ij} lnP_{ij}$.
- For binary case, we have $j \in \{0,1\}$ and from the regression approach earlier:

$$\begin{aligned} P_{i0} &= Pr(Y_i = 0|X) = F(X,\beta) \\ P_{i1} &= Pr(Y_i = 1|X) = 1 - F(X,\beta) \end{aligned}$$

- For binary case, log likelihood is:

$$\begin{aligned} lnL &= ln(Pr(Y_1 = y_1, ..., Y_n = y_n|x)) \\ &= \sum_{i=1}^{n} \left\{ y_i lnF(x_i'\beta) + (1 - y_i) ln[1 - F(x_i'\beta)] \right\} \end{aligned}$$

  - Note: Symmetric dist. $\implies 1 - F(x_i'\beta) = F(-x_i'\beta)$
  - Note: Let $q = 2y - 1$ $\implies lnL = \sum_i lnF(q_i x_i'\beta)$

Binary Case: Model Fit

- Lots of measures, little guidance
  1. χ^2 test of slope coefficient significance (like a regression F-test)
  2. Pseudo-$R^2 = 1 - (lnL/lnL_0)$ where ln L = log likelihood of model and ln L_0 = log likelihood of constant only model (not same as R^2 from linear regression)
  3. 2 x 2 table of percentage hits and misses

Observed Value	Predicted Value $\hat{Y} = 0$	Predicted Value $\hat{Y} = 1$
$Y = 0$		
$Y = 1$		

where $\hat{Y} = 1$ if $\hat{F} > F^*$ and F^* is a threshold (e.g., 50%).

Naming Convention: "Multinomial Logit"

- The **multinomial logit** often refers to models in which the data are individual specific (no variation across alts).
- Since only differences across alts matter, and individual variables don't vary across alts, we can't identify coefficients on these vars.

$$P_{ij} = \frac{e^{x_i\beta}}{\sum_j^J e^{x_i\beta}} = \frac{1}{J}$$

- Solution: Interact these vars with dummies that vary across alts. Equiv to letting parameters vary across alts.

$$P_{ij} = \frac{e^{x_i\beta_j}}{\sum_k e^{x_i\beta_k}}$$

- The **conditional logit** often refers to models in which the data var across agent's *and* alts. We can restrict coefs to be constant across alts.

Naming Convention: "Conditional Logit"

- The **conditional logit** often refers to models in which the data var across agent's *and* alts. We can restrict coefs to be constant across alts.

$$P_{ij} = \frac{e^{x_{ij}\beta}}{\sum_k e^{x_{ik}\beta_k}}$$

- Derivative wrt x_{im}

$$\frac{\partial P_{ij}}{\partial x_{im}} = [P_{ij}(I(j=m) - P_{im})]\,\beta, m = 1, ..., J$$

- Elasticities. The effect of attribute k on choice m on P_{ij}

$$\frac{\partial \ln P_{ij}}{\partial \ln x_{mk}} = x_{mk}\,[I(j=m) - P_{im}]\,\beta_k$$

- The distinction between multinomial and conditional is artificial.

Introduction

- Logit assumes iia, imposing proportional substitution.
- GEV relax iia, allowing for variety of substitution patterns.
- Key assumption: unobserved portion of utility $\forall$ alts is jointly GEV, allowing for corr over alts.
- Example of GEV:
  1. **Nested Logit**
  2. **Paired Combinatorial Logit (PCL)**
  3. **Generalized Nested Logit (GNL)**
- Advantage of GEV = choice probs in *close form*.

Substitution Patterns

- Nested logit works when set of alts can be partitioned into subsets (i.e., nests):
  1. IIA holds within nest. (I.e., for any 2 alts in same nest, ratio of probs is indep of attributes or existence of all other alts.)
  2. IIA *doesn't* holds across nests. (I.e., Ratio of probs can depend on attributes of other alts in the two nests.)
- E.g., Choices = {drive alone, carpool, bus, rail}. How would probs change if one choice were removed?

		With Alt Removed			
Alt	Orig	Alone	Carpool	Bus	Rail
Alone	.40	–	.45 (+.125)	.52 (+.30)	.48 (+.20)
Carpool	.10	.20 (+1)	–	.52 (+.30)	.48 (+.20)
Bus	.30	.48 (+.60)	.33 (+.1)	–	.40 (+.33)
Rail	.20	.32 (+.60)	.22 (+.1)	.35 (.70)	–

Substitution Patterns (Con't)

- To determine partition look at how probs change when removing option.
- Note:
  1. Bus and Rail probs always change by same amount
  2. Along and Carpool probs always change by same amount
- Note: Remove alone & carpool rises proportionately *more* (1.0) than prob of bus (.60) or rail (.60)
- IIA holds within a nest but not across nests. Remove alt outside the nest and the probs of alts within the nest all change proportionately the same.

Choice Probabilities I

- Partition alts into K nonoverlapping subsets (nests): $B_1, ..., B_K$.
- Utility is still: $U_{ij} = V_{ij} + \varepsilon_{ij}$
- Nested logit comes from assumption on $\varepsilon_i = (\varepsilon_i 1, ..., \varepsilon_i J)$:

$$exp\left(-\sum_{k=1}^{K}\left(\sum_{j\in B_k} e^{-\varepsilon_i/\lambda_k}\right)^{\lambda_k}\right)$$

- Marginal dist of each ε_{ij} is univariate extreme.
- ε_{ij} are correlated within nests, λ_k is measure of degree of independence among alts in nest k.
- $1 - \lambda_k$ is an indicator of correlation among alts in nest k. $\lambda_k = 1 \forall k \implies$ standard multinomial logit.

Choice Probabilities II

- Choice probability for alternative j in nest B_k:

$$P_{ij} = \frac{e^{V_{ij}/\lambda_k} \left(\sum_{s \in B_k} e^{V_{is}/lambda_k} \right)^{\lambda_k - 1}}{\sum_{l=1}^{S} \left(\sum_{s \in B_l} e^{V_{is}/lambda_l} \right)^{\lambda_l}}$$

- Denominator is just sum over all nests, sum over all alts in a nest.
- Consider two alts, $j \in B_k$ and $m \in B_l$.

$$\frac{P_{ij}}{P_{im}} = \frac{e^{V_{ij}/\lambda_k} \left(\sum_{s \in B_k} e^{V_{is}/lambda_k} \right)^{\lambda_k - 1}}{e^{V_{im}/\lambda_l} \left(\sum_{s \in B_l} e^{V_{is}/lambda_l} \right)^{\lambda_l - 1}}$$

- If two alts in same nest ($k = l$)

$$\frac{P_{ij}}{P_{im}} = \frac{e^{V_{ij}/\lambda_k}}{e^{V_{im}/\lambda_l}}$$

IIA in the Nested Logit

- Ratio for two alts in same nest is indep of other alts (factors in parentheses cancel).
- Ratio for two alts in different nests depends on attributes of *all alts* in the nests containing j and m. Doesn't depend on attributes of alts in nests other than those containing j and m.
- A form of IIA holds in Nested Logit, IIN = Independence from Irrelevant Nests. Drop an alternative from one nest and all alts in another nest change prob in same proportion.

λ_k

- λ_k can vary over nests k reflecting different correlation among unobserved factors within each nest. in paren cancel.
- For model to be consistent with utility maximizing behavior, $\lambda_k \in (0, 1) \forall k$.
- For $\lambda_k > 1$, model consistent with utility maximizing behavior for a *range* of explanatory vars.
- $\lambda_k < 0 \implies$ model inconsistent with utility maximizing behavior and implies improving attributes of an alt can dec prob of alt being chosen.
- λ_k can be specified as a fxn of demographic characteristics (e.g., $\lambda_k = exp(\alpha z_i)$).

Decomposition in Two Logits I

- WLOG observed components of util, V, can be expressed into 2 parts:
  1. W_{ik} constant across alts within nest k
  2. Y_{ij} varies across alts within nest k

$$U_{ij} = W_{ik} + Y_{ij} + \varepsilon_{ij}, j \in B_k$$

- This decomposition enables us to write nested logit prob as product of 2 standard logits.

$$P_{ij} = P_{ij|B_k} P_{iB_k}$$

where $P_{ij|B_k}$ = conditional prob of choosing alt j given nest B_k was chosen, P_{jB_k} is marginal (over alts in nest B_k) prob of choosing alt j in nest B_k.

Decomposition in Two Logits II

- Conditional ($P_{ij|B_k}$) and marginal (P_{iB_k}) distributions are logits:

$$P_{iB_k} = \frac{e^{W_{ik}+\lambda_k I_{ik}}}{\sum_{l=1}^{K} e^{W_{il}+\lambda_l I_{il}}}; \; P_{ij|B_k} = \frac{e^{Y_{ij}/\lambda_k}}{\sum_{j\in B_k} e^{Y_{ij}/\lambda_k}}$$

  where I_{ik}

$$I_{ik} = \ln \sum_{j\in B_k} e^{Y_{ij}/\lambda_k}$$

- Prob of choosing an alt in B_k is a logit over nests and includes all vars that vary over nests W_{ik} but *not* vars that vary over alternatives, Y_{ij}.
- Conditional prob of choosing alt j given an alt in B_k was chosen is also logit over alts in B_k and includes all vars that vary over alts in nest, Y_ij but *not* vars that vary over nests, W_{ik}.

Inclusive Value Term

- I_{ik} is the log of the denominator of conditional prob. Called **inclusive value** of nest B_k.
- $\lambda_k I_{ik}$ = expected utility that agent i receives from choice among alts in nest B_k
- Marginal prob (choice of nest) = *upper model*
- Conditional prob (choice of alt—nest) = *lower model*. Some people don't divide by λ_k in lower model (STATA).
- Intuition: Inclusive value term enters upper model as explanatory var because choosing nest B_k depends on
  1. Expected util regardless of chosen alt, W_{ik} plus
  2. Expected util from being able to choose bets alt in nest, $\lambda_k I_{ik}$.

Marginal Effects

- Change in attribute r in the utility function for alt K in

$$\frac{\partial \ln Prob(alt = j, nest = k)}{\partial x(}$$

Estimation

- MLE works but likelihood fxn not globally concave, like logit. Check ests by varying starting values.
- Can estimate sequentially $\Longrightarrow$ consistent but inefficient ests. Sequential est performed "bottom up."
  1. Estimate lower model. I.e., estimate separate logits on each nest.
  2. Use coef ests from (1) to compute inclusive value terms.
  3. Estimate upper model (choice of nest), with inclusive value entering as explanatory vars.
- Two problems with sequently estimation:
  1. SEs biased downward because of estimation error in IV terms. (can correct for this, Ben-Akiva and Lerman (1985)).
  2. Some parameters appear in several submodels and equality restrictions may be violated.
- Use FIML if possible

GEV More Broadly

- Can have more than two levels.
- Can have overlapping nests. Alts appear in more than one nest. (Cross-nested logits, ordered GEV, paired combinatorial logit, Generalized nested logit)