Multivariate Data Analysis

(MGT513, BAT531, TIM711)

Lecture 13

Logit Choice Model

References

- LCG – Ch. 13 Logit Choice Models
- Discrete Choice Methods with Simulation
  - Ch. 2, Ch. 3, and Ch. 6

  https://eml.berkeley.edu/books/choice2.html
- R for Marketing Research and Analytics
  - Ch. 13: Choice Modeling

  Online version is available at the UNIST library

Agenda

1. Intro to Choice Models
2. Conjoint Analysis
3. Intro to Bayesian Statistics
4. Heterogeneity
5. Hierarchical linear model (R exercise)

Intro to Choice Models

Random Utility Model (RUM)

- A decision maker n faces a choice among J alternatives.
- The utility that decision maker n obtains from alternative j is $U_{nj}, j = 1, \ldots, J$.

$$U_{nj} = V_{nj} + \varepsilon_{nj} \;\; \forall j$$

  - V_{nj} (Deterministic component): representative utility / known part to the researcher
  - ε_{nj}(Stochastic/Random component): error / unknown part that is treated by the researcher as random

Random Utility Model (RUM)

- Decision rule: Choose alternative i if and only if $U_{ni} > U_{nj} \ \forall j \neq i$

$$
\begin{aligned}
P_{ni} &= \mathrm{Prob}(U_{ni} > U_{nj} \ \forall j \neq i) \\
&= \mathrm{Prob}\left(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \forall j \neq i\right) \\
&= \mathrm{Prob}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \ \forall j \neq i)
\end{aligned}
$$

Identification of Choice Model (RUM)

1. Only Differences in Utility Matter

- The absolute level of utility is irrelevant to both the decision maker's behavior and the researcher's model. If a constant is added to the utility of all alternatives, the alternative with the highest utility doesn't change.
- The decision maker chooses the same alternative with $U_{nj}\ \forall j$ as with $U_{nj} + k\ \forall j$ for any constant k.
- "A rising tide raises all boats."

Identification of Choice Model (RUM)

2. The Overall Scale of Utility Is Irrelevant

- Just as adding a constant to the utility of all alternatives does not change the decision maker's choice, neither does multiplying each alternative's utility by a constant. The alternative with the highest utility is the same no matter how utility is scaled.
- The model $U_{nj}^{0} = V_{nj} + \varepsilon_{nj} \;\; \forall j$ is equivalent to $U_{nj}^{1} = \lambda V_{nj} + \lambda \varepsilon_{nj} \;\; \forall j$ for any.
- To take account of this fact, the researcher must normalize the scale of utility.

Binary Choice

Linear specification of V_{nj} :

$$V_{nj} = \beta' x_{nj}$$

where

x_{nj}: predictors influencing the choice of alternative j

Choose alternative 1 over 2 if

$$U_{n1} > U_{n2}$$

$$\beta' x_{n1} + \varepsilon_{n1} > \beta' x_{n2} + \varepsilon_{n2}$$

$$(\varepsilon_{n2} - \varepsilon_{n1}) < \beta'(x_{n1} - x_{n2})$$

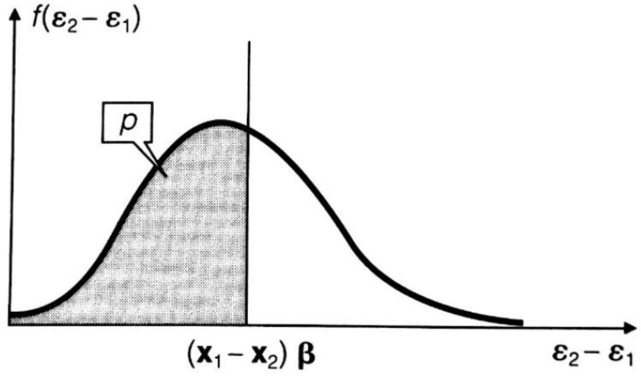
Binary Choice

Let $f(\varepsilon_{n2} - \varepsilon_{n1})$ be the probability density function of $(\varepsilon_{n2} - \varepsilon_{n1})$

The probability that the individual n chooses alternative 1:

$$P_{n1} = \int_{-\infty}^{\beta'(x_{n1}-x_{n2})} f(\varepsilon_{n2} - \varepsilon_{n1})\partial\,(\varepsilon_{n2} - \varepsilon_{n1})$$
$$= F(\beta'(x_{n1} - x_{n2}))$$

FIGURE 13.2
Probability that $u_1 > u_2$



Binary Probit Model

Assume that ε_{n1} and ε_{n2} follow normal distributions. Then $(\varepsilon_{n2} - \varepsilon_{n1})$ also follows a normal distribution.

$$P_{n1} = \Phi(\beta'(x_{n1} - x_{n2}))$$

No closed functional form for probability of choice

Binary Logit Model

Assume that each ε_{n1} and ε_{n2} is independently, identically distributed extreme value (Gumbel distribution / type I extreme value distribution / double exponential distribution).

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

Then $(\varepsilon_{n2} - \varepsilon_{n1})$ follows a logistic distribution.

$$f(\varepsilon_{n2} - \varepsilon_{n1}) = \frac{e^{-(\varepsilon_{n2}-\varepsilon_{n1})}}{(1 + e^{-(\varepsilon_{n2}-\varepsilon_{n1})})^2}$$

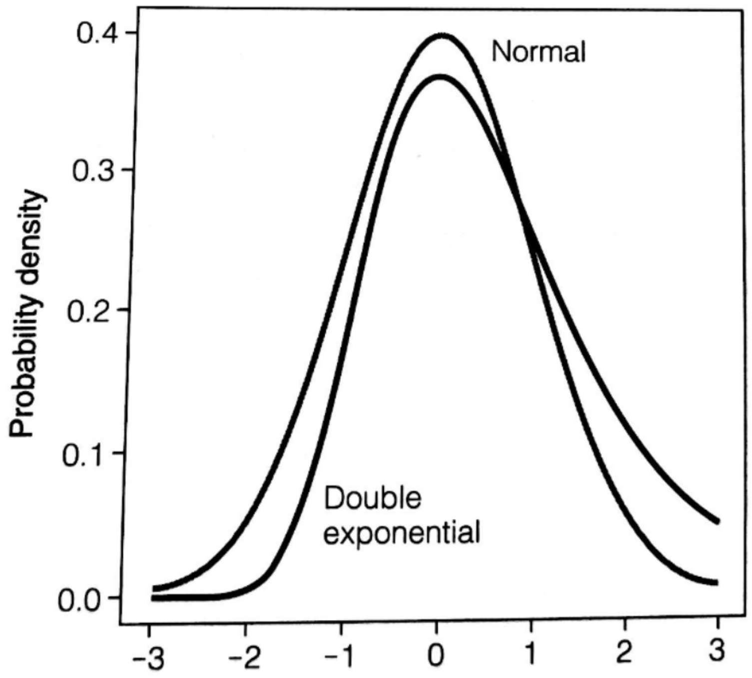
Then

$$P_{n1} = \int_{-\infty}^{\beta'(x_{n1}-x_{n2})} \frac{e^{-(\varepsilon_{n2}-\varepsilon_{n1})}}{(1 + e^{-(\varepsilon_{n2}-\varepsilon_{n1})})^2} \partial\left(\varepsilon_{n2} - \varepsilon_{n1}\right)$$

$$= \frac{e^{\beta'(x_{n1}-x_{n2})}}{1 + e^{\beta'(x_{n1}-x_{n2})}} = \frac{e^{\beta' x_{n1}}}{e^{\beta' x_{n1}} + e^{\beta' x_{n2}}}$$

Binary Logit Model

FIGURE 13.6
Comparison of the normal and double-exponential distributions



Difference between Logit and Probit

- Difference between estimates of α and β of two models is substantial – Scale differences
- Difference between choice probabilities of two models is not substantial

TABLE 13.4 Estimated values of α and β for the logit and probit models

Coeff	Logit	Probit
α	−2.94	−1.60
β	−0.20	−0.11

TABLE 13.5 Fitted choice probabilities for the logit and probit models at different levels of discount

Discount	Logit	Probit
0.00	0.0502	0.0542
0.05	0.1256	0.1469
0.10	0.2807	0.3104
0.15	0.5146	0.5242
0.20	0.7423	0.7310
0.25	0.8867	0.8792
0.30	0.9551	0.9578

Binary Logit Model

Odds:

$$0 \leq \frac{\theta(x)}{1 - \theta(x)} : \text{Odds of Success}$$

- If the probability of success is $\theta(x) = 0.25$, the odds of success= 0.25/(1 − 0.25) = 1/3: one success to each three failures.
- If the probability of success is $\theta(x) = 0.8$, the odds of success= 0.8/(1 − 0.8) = 4: four successes to one failure.

Binary Logit Model

Let's $x_{n1} - x_{n2} = x$ and rewrite the logistic model as:

$$P_1 = \theta(x) = \frac{e^{\beta' x}}{1 + e^{\beta' x}}$$

$$P_2 = 1 - \theta(x) = \frac{1}{1 + e^{\beta' x}}$$

- Logit link function (in a GLM framework)

$$\mathrm{logit}[\theta(x)] = \log\left[\frac{\theta(x)}{1 - \theta(x)}\right] = \beta' x$$

The log-odds follows a linear model.

Maximum Likelihood Estimation (MLE): Binary Logit Model

Let

$$Y_n = \begin{cases} 1 \text{ if choosing alternative } 1 \\ 0 \text{ if choosing alternative } 2 \end{cases}$$

$$P(Y_n = 1) = P_{n1} = \frac{e^{\beta' x_{n1}}}{e^{\beta' x_{n1}} + e^{\beta' x_{n2}}}$$

$$P(Y_n = 0) = P_{n2} = \frac{e^{\beta' x_{n2}}}{e^{\beta' x_{n1}} + e^{\beta' x_{n2}}} = 1 - P_{n1}$$

Likelihood function:

$$L = \prod_n P_{n1}^{Y_n} (1 - P_{n1})^{(1-Y_n)}$$

Maximum Likelihood Estimation (MLE): Binary Logit Model

The log-likelihood:

$$\ln(L) = \sum_{n} [Y_n \ln(P_{n1}) + (1 - Y_n) \ln(P_{n1})]$$

MLE: Choose β to maximize $\ln(L)$ - Use numerical optimization methods (i.e. Newton-Ralphson, BHHH, or BFGS)

Multinomial Logit Model

The joint density of the random vector $\varepsilon \ = < \varepsilon_{n1}, \ldots, \varepsilon_{nJ} >$ is denoted $f(\varepsilon_n)$. Using the density $f(\varepsilon_n)$, this cumulative probability can be rewritten as

$$P_{ni} = \text{Prob}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \ \forall j \neq i)$$

$$= \int_{\varepsilon} \ I(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \ \forall j \neq i) \, f(\varepsilon_n) d\varepsilon_n$$

The logit model is obtained by assuming that each ε_{nj} is independently, identically distributed extreme value. The distribution is also called Gumbel and type I extreme value distribution

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

The cumulative distribution is

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

Multinomial Logit Model

$$
\begin{aligned}
P_{ni} &= \text{Prob}(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \forall j \neq i) \\
&= \text{Prob}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \ \forall j \neq i)
\end{aligned}
$$

Since the ε's are independent, this cumulative distribution over all $j \neq i$ is the product of the individual cumulative distributions:

$$
P_{ni} | \varepsilon_{ni} = \prod_{j \neq i} e^{-e^{-(\varepsilon_{ni} + V_{ni} - V_{nj})}}
$$

Integrate $P_{ni} | \varepsilon_{ni}$ over ε_{ni}

$$
P_{ni} = \int (\prod_{j \neq i} e^{-e^{-(\varepsilon_{ni} + V_{ni} - V_{nj})}}) e^{-\varepsilon_{ni}} e^{-e^{-\varepsilon_{ni}}} \, d\varepsilon_{ni}
$$

Multinomial Logit Model

Logit probability (closed form expression)

$$P_{ni} = \int \left(\prod_{j \neq i} e^{-e^{-(\varepsilon_{ni}+V_{ni}-V_{nj})}}\right) e^{-\varepsilon_{ni}} e^{-e^{-\varepsilon_{ni}}}\, d\varepsilon_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}$$

Linear specification of V_{nj} : $V_{nj} = \beta' x_{nj}$

$$P_{ni} = \frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}}$$

Maximum Likelihood Estimation (MLE): Multinomial Logit Model

Let Y_{nj} denote a dummy variable which is equal to one if individual n made choice j and 0 otherwise.

The probability of the choice made for one individual n:

$$P_n = \prod_j P_{nj}^{Y_{nj}}$$

$$\ln(P_n) = \sum_j Y_{nj} \ln(P_{nj})$$

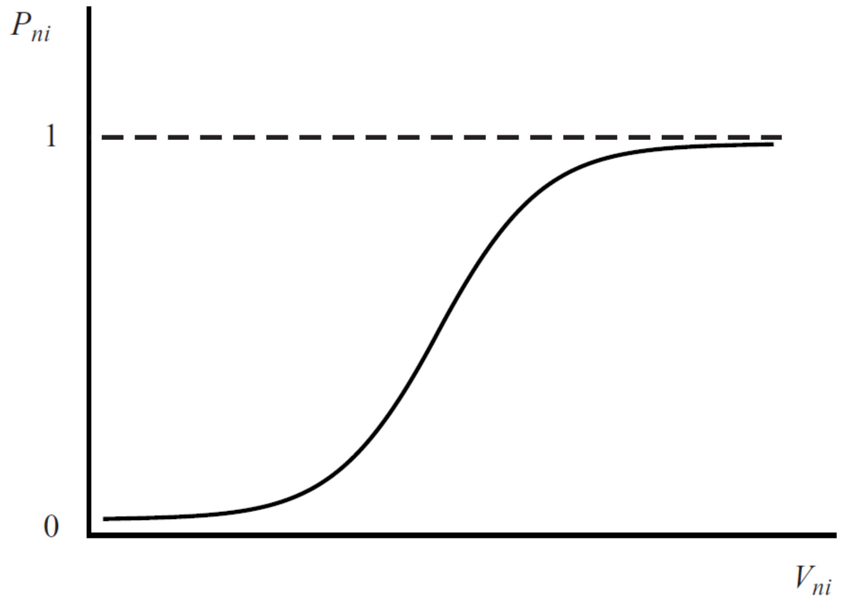
The log-likelihood function:

$$\ln(L) = \sum_n \ln(P_n) = \sum_n \sum_j Y_{nj} \ln(P_{nj})$$

MLE: Choose β to maximize $\ln(L)$ - Use numerical optimization methods (i.e. Newton-Ralphson, BHHH, or BFGS)

Properties of Logit Model

1. The choice probabilities are between 0 and 1 (as required for a probability)
2. The choice probabilities for all alternatives sum to 1: $\sum_{i=1}^{J} P_{ni} = \frac{\sum_i e^{V_{ni}}}{\sum_j e^{V_{nj}}} = 1$
3. S-shaped curve



Limitations of Logit Model

1. Taste variation

   Logit model cannot represent random taste variation.

2. Substitution Patterns

   The logit model implies proportional substitution across alternatives, given the researcher's specification of representative utility.

   IIA property(Independence from Irrelevant Alternatives)

$$\frac{P_{ni}}{P_{nk}} = \frac{e^{V_{ni}} / \sum_j e^{V_{nj}}}{e^{V_{nk}} / \sum_j e^{V_{nj}}} = \frac{e^{V_{ni}}}{e^{V_{nk}}} = e^{V_{ni} - V_{nk}}$$

Limitations of Logit Model

IIA property: “red-bus–blue-bus problem”

Transportation choice: Car (c) vs. Bus (b)

Red bus (rb), Blue bus (bb)

Assume that $P_c = P_{bb} = \frac{1}{2}$. Then $P_c/P_{bb} = 1$

If Red bus (rb) is introduced, the logit model predicts:

$P_{rb}/P_{bb} = 1$ and $P_c/P_{bb} = 1$. Then $P_c = P_{bb} = P_{rb} = \frac{1}{3}$

However, the correct Expectation should be:

$P_c = \frac{1}{2}$, $P_{bb} = P_{rb} = \frac{1}{4}$

Limitations of Logit Model

3. Penal data

If unobserved factors are independent over time in repeated choice situations, then logit can capture the dynamics of repeated choice, including state dependence. However, logit cannot handle situations where unobserved factors are correlated over time.

Dynamics associated with unobserved factors cannot be handled, since the unobserved factors are assumed to be unrelated over choices.

Elasticities

Elasticities: the percentage change in one variable that is associated with a one-percent change in another variable

Let z_{ni} be an attribute of alternative i.

The elasticity of P_{ni} with respect to z_{ni}

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

If representative utility (V_{ni}) is linear in z_{ni} with coefficient β_{z} ($V_{\mathrm{ni}} = \beta_{\mathrm{z}} \mathrm{z}_{ni}$)

$$E_{iz_{ni}} = \beta_{\mathrm{z}} \mathrm{z}_{\mathrm{ni}}(1 - P_{ni})$$

Elasticities

The cross-elasticity of P_{ni} with respect to a variable entering alternative j $(z_{n\mathrm{j}})$

$$E_{i,z_{nj}} = \frac{\partial P_{ni}}{\partial z_{nj}} \frac{z_{nj}}{P_{ni}}$$
$$= -\frac{\partial V_{nj}}{\partial z_{nj}} z_{nj} P_{nj}$$

If $V_{n\mathrm{i}} = \beta_{\mathrm{z}} \mathrm{z}_{n\mathrm{i}}$, then $E_{i,z_{nj}} = -\beta_z z_{nj} P_{nj}$

* A change in an attribute of alternative j changes the probabilities for all other alternatives by the same percent.

Model Significance

To assess the model

- Let LL_F and LL_R denote log-likelihood of a full model and a restricted model respectively. Then

$$LL_F - LL_R \sim \chi^2(\mathrm{df}_F - \mathrm{df}_R) \text{ as } n \to \infty$$

Goodness of Fit

Consider two types of *Information Criterion*

- $\mathrm{AIC} = -2LL + 2k$ by Akaike
- $\mathrm{SC} = -2LL + 2\ln(n)k$ by Schwartz

where k = number of parameters

- Smaller values of AIC and SC indicate better model fit
- Schwartz criterion is more conservative

Goodness of Fit

- Useful to explain how much uncertainty is explained by the model: McFadden suggests

$$\rho^2 = 1 - \frac{LL_F}{LL_0}$$

where LL_0 is the log-likelihood of the model with only intercept

- Unlikely R^2 in regression, it's unusual to see values of ρ^2 near 1.0
- A criterion: Good fit if $0.3 \leq \rho^2 \leq 0.5$

Mixed Logit

Mixed Logit can overcome the three limitations of standard logit by allowing for random taste variation, unrestricted substitution patterns, and correlation in unobserved factors over time.

The utility of person n from alternative j (linear utility specification)

$$U_{nj} = V_{nj}(\beta_n) + \varepsilon_{nj} = \beta_n{}' x_{nj} + \varepsilon_{nj}$$

Mixed Logit

Logit probability conditional on β_n

$$L_{ni}(\beta_n) = \frac{e^{\beta_n{}' x_{ni}}}{\sum_j e^{\beta_n{}' x_{nj}}}$$

Since β_n is unknown, we integrate $L_{ni}(\beta_n)$ over the distribution of β_n (mixing distribution $f(\beta)$)

Mixed logit probability

$$P_{ni} = \int L_{ni}(\beta) f(\beta) d\beta = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}}\right) f(\beta) d\beta$$

Mixed Logit

Heterogeneity specification via $f(\beta)$

1. Standard logit model

$f(\beta)$ degenerate at fixed parameters b

$$\begin{cases} f(\beta) = 1 \quad for \;\; \beta = b \\ f(\beta) = 0 \quad for \;\; \beta \neq b \end{cases}$$

$$P_{ni} = \frac{e^{b' x_{ni}}}{\sum_j e^{b' x_{nj}}}$$

Mixed Logit

Heterogeneity specification via $f(\beta)$

2. Latent class model

M segments (latent classes) with proportion s_m and parameter b_m

$$P_{ni} = \sum_{m=1}^{M} s_m \left(\frac{e^{b_m' x_{nj}}}{\sum_j e^{b_m' x_{nj}}} \right), \qquad \sum_m s_m = 1$$

Mixed Logit

Heterogeneity specification via $f(\beta)$

3. Mixed logit (Random Coefficients)

$f(\beta)$: continuous distribution (e.g. normal, lognormal, uniform, mixture of normals)

Normal distribution case:

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}}\right) \Phi(\beta | b, W) d\beta$$

where $\Phi(\beta | b, W)$ is the normal density with mean b and covariance W

Conjoint Analysis

Conjoint: Motivation

- How to learn what customers want?: ask direct questions about their preferences
  - What brand do you prefer?
  - What interest rate would you like?
  - What annual fee would you like?
  - What credit limit would you like?

- What kind of answers do you expect?

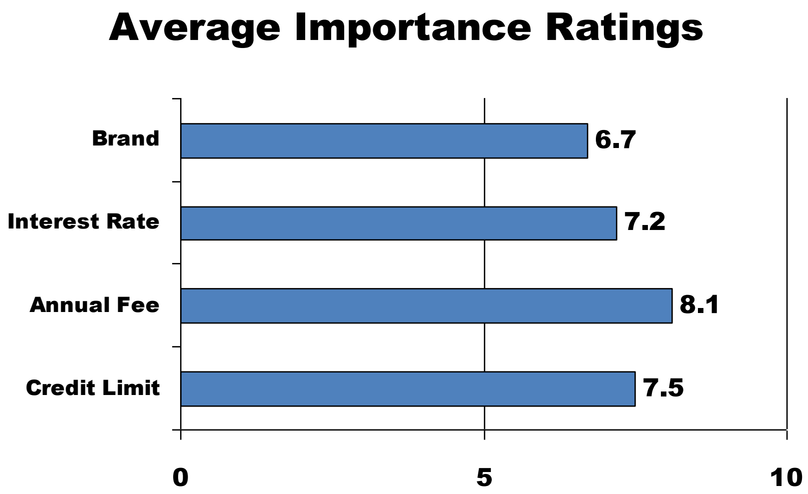
Conjoint: Motivation

- How to learn what customers want?: ask direct questions about their preferences

How important is it that you get the (brand, interest rate, annual fee, credit limit) that you want?

Not Important Very Important

0 1 2 3 4 5 6 7 8 9 10



What is **Cons**ider**joint**ly Analysis?

- A technique which requires respondents to evaluate different product/service bundles, each of which is characterized by different attributes (i.e., customers have to make tradeoffs among different attributes of a product/service)
- The basic outputs of conjoint analysis are:
  - Utility (i.e., part-worth, preference) each customer assigns to each level of each attribute
  - Numerical assessment of relative importance each customer attaches to different attributes of a product/service

Example: Coffee Maker

Assess how consumers evaluate the following levels of each of these product attributes

▪ Capacity:	4	8	10		cups
▪ Price:	$18	$22	$28		
▪ Brewing time:	3	6	9	12	minutes

Rating Tasks

Next we present to you descriptions of a series of Coffee Makers. Please rate your likelihood of purchasing each of these coffee makers on the following scale of 1 to 100.

Extremely Unappealing **Neutral** **Extremely Appealing**

|--------------------------|--------------------------|-------------------------|--------------------------|

1 **50** **100**

While evaluating these alternatives, remember that a configuration that is more appealing should be given a higher rating than those that are less appealing.

Coffee Maker # 1

Capacity: 4 cups

Brewing Time: 3 minutes

Price: $ 18

Your Rating ______

All Possible Coffee Maker Bundles To Be Rated

CAPACITY	4 cups			8 cups			10 cups		
PRICE	$18	$22	$28	$18	$22	$28	$18	$22	$28
BREWING TIME									
3 minutes	1	5	9	13	17	21	25	29	33
6 minutes	2	6	10	14	18	22	26	30	34
9 minutes	3	7	11	15	19	23	27	31	35
12 minutes	4	8	12	16	20	24	28	32	36

Different Types of Conjoint Analysis

- Conventional: non-metric (ranking), metric (rating)
- Self-explicated
- Adaptive: endogeneity issue
- Hybrid: a combination of self explicated and conventional conjoint
- Choice-based: have become popular

Choice-Based Conjoint Example

Assess how a parent-teen dyad evaluates the following levels of each personal computer attribute

Attributes	Levels
Computer brand	Dell HP
CPU brand	Intel AMD
CPU speed	3 GHz 5 GHz
Warranty	2-yr warranty No warranty
Price	$1,299 $1,799

Choice Tasks

Question 1: Please check (in the space □ provided) the PC you prefer the most.

Alternative 1: □		Alternative 2: □		Alternative 3: □
Dell PC		HP PC		Dell PC
Intel		Intel		AMD
3 GHz		5 GHz		3 GHz
2-yr warranty service		2-yr warranty service		No warranty service
$1799		$1299		$1799

Question 2: Please check (in the space □ provided) the PC you prefer the most.

Alternative 1: □		Alternative 2: □		Alternative 3: □
HP PC		Dell PC		HP PC
AMD		AMD		Intel
3 GHz		5 GHz		5 GHz
2-yr warranty service		2-yr warranty service		No warranty service
$1799		$1299		$1299

Multinomial Logit Model

$$Pr(\text{selecting } i) = \frac{e^{V_i}}{\sum_k e^{V_k}}$$

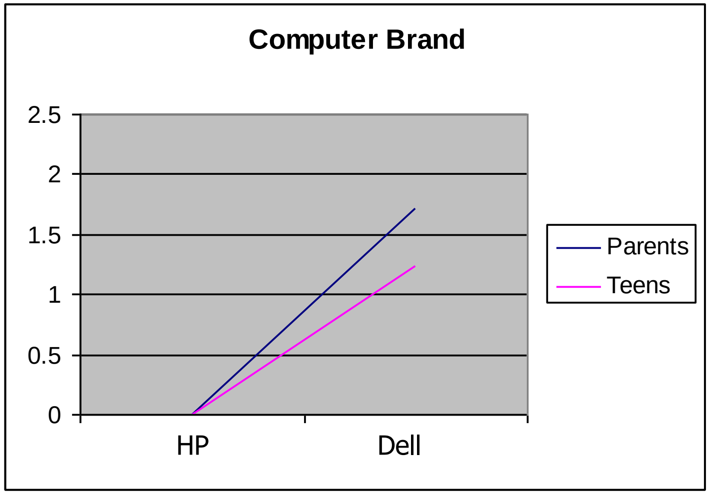
$$V_k = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i}$$

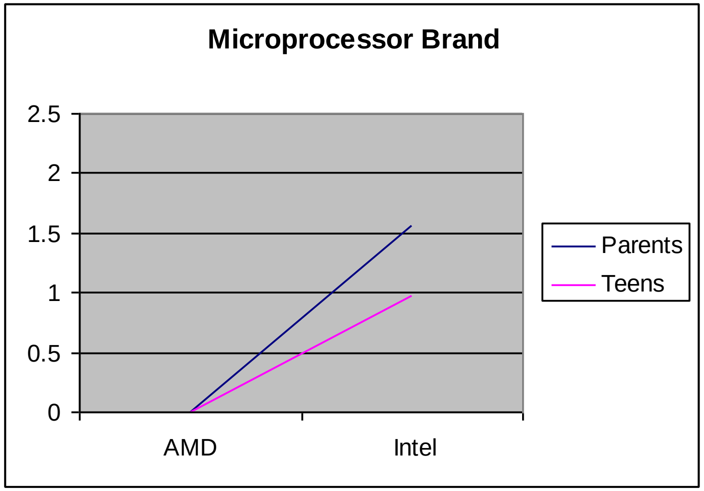
where

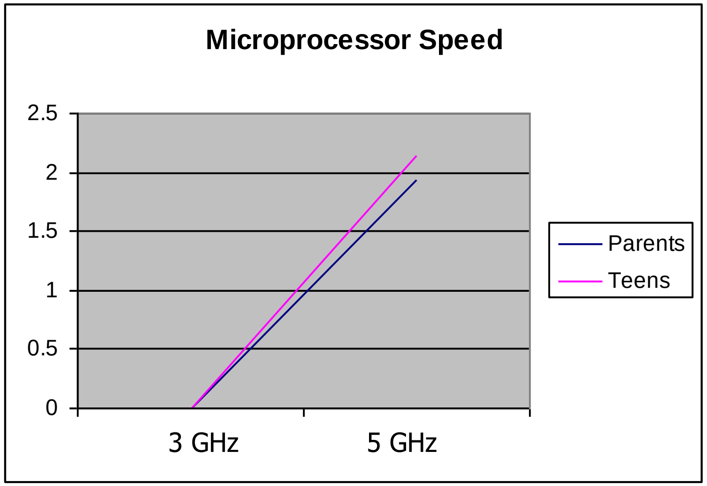
X_1 = 1 if computer brand is Dell and 0 otherwise
X_2 = 1 if microprocessor brand is Intel and 0 otherwise
X_3 = 1 if microprocessor speed is 5 GHz and 0 otherwise
X_4 = 1 if warranty is two-years and 0 otherwise
X_5 = 1 if price is $1799 and 0 otherwise

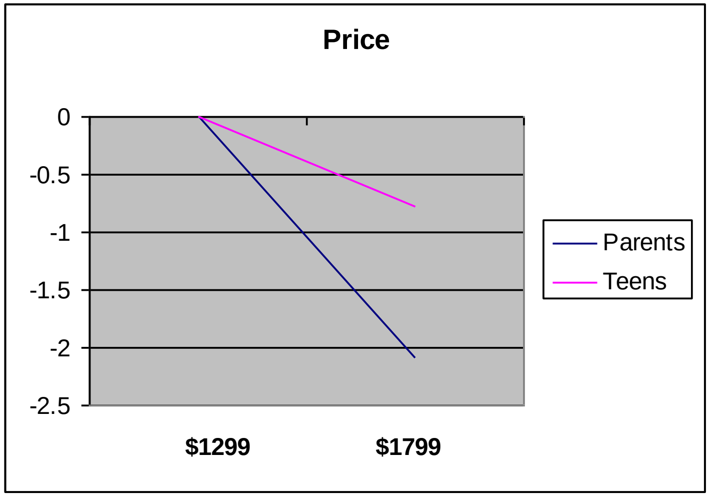
The baseline PC is HP; AMD; 3 GHz; no warranty; $1299

Utilities for Different PC Attributes: Heterogeneity









Intro to Bayesian Statistics

Frequentist (Classical) Approach

- Frequentist or classical approach
- Data y: an outcome of a random experiment
- Model: the data generating mechanism (i.e. the distribution of the data)
- Parameter θ: a quantity that characterizes the data generating mechanism
- Likelihood function:

$$l(\theta) = p(y|\theta) = \prod_i p(y_i|\theta)$$

Frequentist (Classical) Approach

Example: $N(y|\mu, \sigma^2)$

- Likelihood function: $\prod_i N(y_i|\mu, \sigma^2)$
- Parameter estimates:

$$\hat{\mu} = \bar{y} = \frac{\sum_i y_i}{n}, \qquad \hat{\sigma}^2 = s^2 = \frac{\sum_i (y_i - \bar{y})^2}{n-1}$$

- Sampling distribution

$$E(\bar{y}) \approx \mu, \qquad Var(\bar{y}) \approx s^2/n$$

- 95% confidence interval

$$\bar{y} \pm 1.96 s/\sqrt{n}$$

Bayesian Approach

- Data y: fixed information gathered
- Model Parameter θ: an unknown (random) quantity
- Three components
  - Likelihood function $[p(y|\theta)]$
  - Prior distribution$[p(\theta)]$: characterizes subjective beliefs (probabilities) about θ without data
  - Posterior distribution$[p(\theta|y)]$: characterizes the conditional probabilities of θ after data are taken into account

Bayes Theorem

Bayes Theorem:

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

Example: normal mean μ (known variance)

$$y \sim N(\mu, \sigma^2)\text{: Likelihood}$$
$$\mu \sim N(\mu_0, \tau_0^2)\text{: Prior}$$

$$p(\mu|y, \sigma^2) \propto p(y|\mu, \sigma^2)p(\mu)\text{: Posterior}$$
$$N(\mu_1, \tau_1^2) = N(\mu, \sigma^2)N(\mu_0, \tau_0^2)$$

$$\mu_1 = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \qquad \tau_1^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)^{-1}$$

MCMC Algorithms

Markov chain Monte Carlo (MCMC)

- A transition density to generate a sequence of draws θ_r

$$p(\theta_r \mid \theta_{r-1})$$

- The stationary or equilibrium distribution

$$p(\theta \mid y)$$

MCMC Algorithms

Markov chain Monte Carlo (MCMC)

- Gibbs Sampling
- Metropolis-Hastings Algorithm
- Hamiltonian Monte Carlo (HMC)
  - No-U-Turn Sampling (Stan)

Multiple Regression Example

$$y_i = x_i'\beta + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

$$Y = X\beta + \varepsilon$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Multiple Regression Example

Frequentist approach:

$$p(Y \mid \beta, \sigma^2) = N(Y \mid X\beta, \sigma^2 I_n)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) \right]$$

$$\hat{\beta} = (X'X)^{-1} X'Y$$

$$\hat{\sigma}^2 = \frac{1}{n}(Y - X\hat{\beta})'(Y - X\hat{\beta})$$

Multiple Regression Example

Bayesian approach:

$$p(Y \mid \beta, \sigma^2) = N(Y \mid X\beta, \sigma^2 I_n)$$

$$p(\beta \mid u_o, V_o) = MVN(\beta \mid u_o, V_o)$$

$$= (2\pi)^{-\frac{p}{2}} \mid V_o \mid^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} (\beta - u_o)' V_o^{-1} (\beta - u_o) \right]$$

$$p(\sigma^2 \mid r_o, s_o) = IG(\sigma^2 \mid r_0 / 2, s_o / 2)$$

Multiple Regression Example

Bayesian approach:

$$p(Y,\beta,\sigma^2) = N(Y \mid X\beta,\sigma^2 I_n) MVN(\beta \mid u_o, V_o) IG(\sigma^2 \mid r_o/2, s_o/2)$$

$$p(\beta,\sigma^2 \mid Y) = \frac{p(Y,\beta,\sigma^2)}{\int\int p(Y,\beta,\sigma^2)\, d\beta\, d\sigma^2}$$

$$\propto N(Y \mid \beta,\sigma^2) MVN(\beta \mid u_o, V_o) IG(\sigma^2 \mid r_o/2, s_o/2)$$

Multiple Regression Example

Bayesian approach:

$$p(\beta \mid Y, \sigma^2) \propto N(Y \mid \beta, \sigma^2) MVN(\beta \mid u_o, V_o) = MVN(\beta \mid u_n, V_n)$$

$$V_n = \left(\frac{1}{\sigma^2} X'X + V_o^{-1} \right)^{-1} \qquad u_n = V_n \left(\frac{1}{\sigma^2} X'Y + V_o^{-1} u_o \right)$$

$$p(\sigma^2 \mid Y, \beta) \propto N(Y \mid \beta, \sigma^2) IG(\sigma^2 \mid r_o, s_o) = IG(\sigma^2 \mid r_n, s_n)$$

$$s_n = s_o + (Y - X\beta)'(Y - X\beta) \qquad r_n = r_o + n$$

Multiple Regression Example

Bayesian approach: MCMC (Gibbs sampler)

do r=1 to R

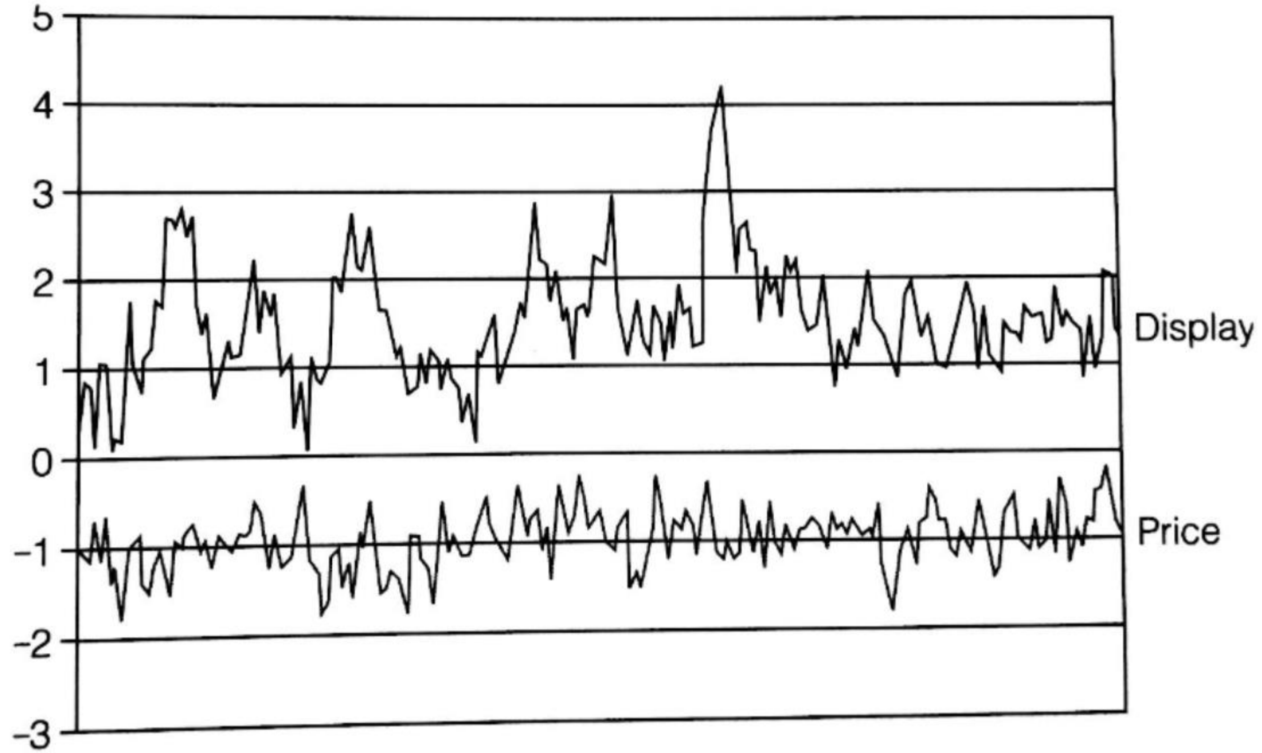
$$\beta_r \sim MVN(u_n, V_n \mid y, \sigma^2_{r-1})$$

$$\sigma^2_r \sim IG(r_n, s_n \mid y, \beta_r)$$

continue

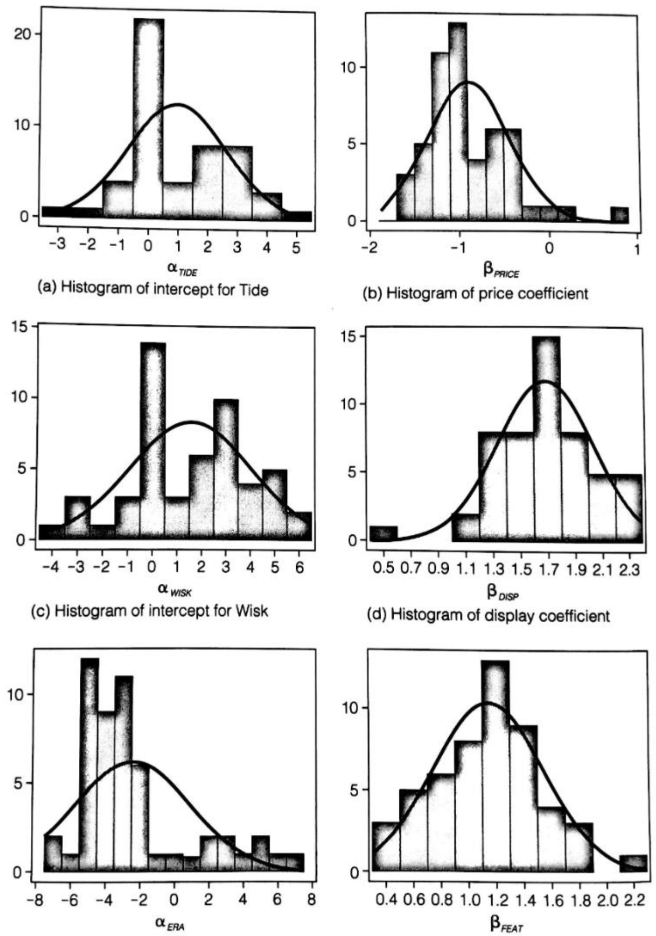
MCMC draws

FIGURE 13.8
Plot of coefficient values for price and display across 20,000 iterations (every 100th value is plotted)



MCMC draws

FIGURE 13.9 Frequency distribution histograms (across 52 panelists) of average parameter values for α_{TIDE}, α_{WISK}, α_{ERA}, β_{PRICE}, β_{DISP}, and β_{FEAT}



(a) Histogram of intercept for Tide

(b) Histogram of price coefficient

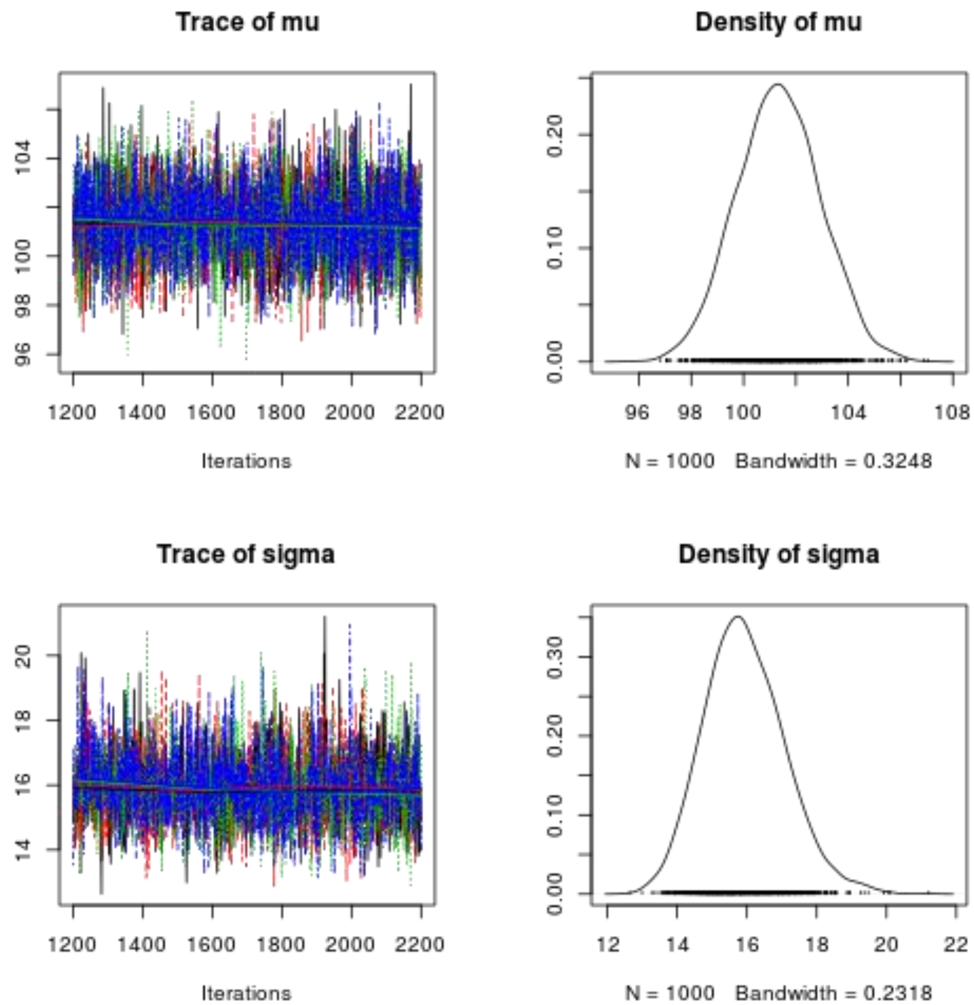
(c) Histogram of intercept for Wisk

(d) Histogram of display coefficient

(e) Histogram of intercept for Era

(f) Histogram of feature coefficient

Output: MCMC draws



Heterogeneity

How to Incorporate Heterogeneity

- Finite mixture model (Kamakura and Russell 1989)

$$u_{kit|s} = x'_{kit}\beta_s + \varepsilon_{kit}$$

Conditional probability

$$\Pr(y_{kit|s} = 1) = \frac{\exp(x'_{kit}\beta_s)}{\sum_{j=1}^{J}\exp(x'_{jit}\beta_s)}$$

Mixing distribution (discrete)

$$f_s = \frac{\exp(\lambda_s)}{\sum_{s'}\exp(\lambda_{s'})}; s = 1, \ldots, S$$

Likelihood function

$$l(u, \beta) = \prod_i f_s \prod_{t_i}\prod_j \Pr(y_{kit|s} = 1)^{y_{kit|s}}$$

UNIST
ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY
2 0 0 9

How to Incorporate Heterogeneity

- Classical random effect model

$$u_{kit} = x'_{kit}\beta_i + \varepsilon_{kit}$$

$$\Pr(y_{kit}=1) = \frac{\exp(x'_{kit}\beta_i)}{\sum_{j=1}^{J}\exp(x'_{jit}\beta_i)}$$

Mixing distribution (continuous)

$$\beta_i \sim MVN(\beta, D)$$

Likelihood function

$$l(\beta,\{\beta_i\}) = \prod_i \prod_{t_i} \prod_j \Pr(y_{kit}=1)^{y_{kit}} MVN(\beta_i \mid \beta, D)$$

$$l(\beta) = \prod_i \int \prod_{t_i} \prod_j \Pr(y_{kit}=1)^{y_{kit}} MVN(\beta_i \mid \beta, D) d_{\beta_i}$$

How to Incorporate Heterogeneity

- Hierarchical Bayes(HB) Model: similar to the classical random effect model, but

Likelihood function: $$l(\beta_i) = \prod_{t_i}\prod_{j} \Pr(y_{kit} = 1)^{y_{kit}}$$

Mixing distribution: $$\beta_i \sim MVN(\beta, D)$$

Prior distribution: $$\beta \sim MVN(u_o, V_o) \quad D_\beta \sim IW(f_o, F_o^{-1})$$

Joint posterior distribution:

$$p(\{\beta_i\}, \beta, D_\beta) \propto \left\{\prod_i l(\beta_i) MNV(\beta_i \mid \beta, D_\beta)\right\} MVN(u_o, V_o) IW(f_o, F_o)$$

MCMC-Hierarchical Multinomial Logit

Do r=1 to R

$$\bar{\beta} \sim MVN(u_n, V_n \mid \{\beta_{ir-1}\}, D_{\beta r-1})$$

Do i=1,n

$$\beta_i \sim P(\beta_i \mid \bar{\beta}_r, D_{\beta r-1}) \propto l(\beta_i) MVN(\bar{\beta}_r, D_{\beta r-1})$$

continue

$$D_\beta \sim IW(f_n, F_n \mid \{\beta_{ir}\}, \bar{\beta}_r)$$

continue

Hierarchical Bayes(HB) Model

$$\beta_i = \eta' w_i + \xi_i$$

$$W = \begin{bmatrix} w_1' \\ \vdots \\ w_m' \end{bmatrix}$$

Mixing distribution: $\beta_i \sim MVN(W\eta, D_\beta)$

Prior distribution: $\eta \sim MVN(u_o, V_o)$ $D_\beta \sim IW(f_o, F_o^{-1})$

Hierarchical linear model
(Hierarchical Bayes)

Logit R.R

Choice-based conjoint model example

1. Conditional multinomial logit model
2. Mixed multinomial logit model
3. Latent class multinomial logit model
4. Hierarchical Bayes multinomial logit model

Choice-based conjoint model example: Hierarchical Bayes multinomial logit model

- Typically: person i, alternative j, time t

Multinomial logit model:

$$u_{ijt} = x'_{ijt}\beta_i + \varepsilon_{ijt}$$

$$P(y_{ijt} = 1) = \frac{e^{x'_{ijt}\beta_i}}{\sum_j e^{x'_{ijt}\beta_i}}$$

Choice-based conjoint model example: Hierarchical Bayes multinomial logit model

- Typically:

$$\beta_i = \Delta' z_i + \nu_i$$

$$\nu_i \sim MVN(0, \Sigma)$$

  - Δ and Σ: hyperparameters
  - z_i includes an intercept
  - $z_i = \begin{cases} [1\ 0]\text{: no carpool} \\ [1\ 1]\text{: carpool} \end{cases}$

Choice-based conjoint model example: Hierarchical Bayes multinomial logit model

- In this example:

$$\beta_i = \mu + \Delta^{*\prime} z_i^* + \nu_i$$

$$\nu_i \sim MVN(0, \Sigma)$$

True $\mu = \begin{bmatrix} -1 \\ -1 \\ 0.5 \\ -1 \\ -2 \\ -1 \\ -2 \end{bmatrix}$

  - μ , Δ^* and Σ: hyperparameters
  - z_i^* does NOT include an intercept
  - $z_i^* = \begin{cases} 0\text{: no carpool} \\ 1\text{: carpool} \end{cases}$

True $\Delta^* = \begin{bmatrix} 1.5 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$