37902 Foundation of Advanced Quantitative Marketing

Session II

Plan of action for the class

- IIA & Elasticity
- Model fit measures
- Welfare calculation
- Breaking free of IIA
  - EBA
  - HEBA
  - Nested Logit (Introducing the "outside good")

1 IIA & Elasticity

- <u>IIA</u>: Independence from irrelevant alternatives

$$\frac{P_{ijt}}{P_{ikt}} = \frac{e^{\alpha_j + X_{jt}\beta}}{e^{\alpha_k + X_{kt}\beta}}$$

  – The ratio of the probabilities of two brands is unaffected by the entry of a new brand that looks a lot like one of them {Pepsi, Sprite} vs {Pepsi, Coke, Sprite}. In the former case, if the probability of a consumer buying each brand is 0.5, then the ratio of Pepsi to Sprite is 1. So even after introduction of Coke, the ratio will be preserved according to the logit model.

- <u>Elasticities</u>

  – Own Elasticity$=\frac{\partial P_{ijt}/\partial X_{jt}}{P_{ijt}/X_{jt}}$

$$\begin{aligned}
\partial P_{ijt}/\partial X_{jt} &= \frac{1}{\sum e^{\alpha_k + X_{kt}\beta}} \cdot \left[\frac{\partial(e^{\alpha_j + X_{jt}\beta})}{\partial X_{jt}}\right] - \frac{e^{\alpha_j + X_{jt}\beta}}{[\sum e^{\alpha_k + X_{kt}\beta}]^2} \cdot \frac{\partial(\sum e^{\alpha_k + X_{kt}\beta})}{\partial X_{jt}} \\
&= \beta P_{ijt} - \beta P_{ijt}^2 \\
&= \beta \cdot P_{ijt}(1 - P_{ijt})
\end{aligned}$$

  – Own Elasticity$=\frac{X_{ijt}}{P_{ijt}}\left[\beta P_{ijt}(1-P_{ijt})\right] = \beta X_{jt}(1-P_{ijt})$

  – Cross Elasticity$=\frac{(\partial P_{ijt}/\partial X_{kt})}{(P_{ijt}/X_{kt})} = \frac{-X_{kt}}{P_{ijt}} \cdot \beta P_{ijt} P_{ikt} = -\beta P_{ikt} X_{kt}$

  – So for all $j \neq k$, the cross elasticity will be the same. This is another property of the logit model.

2 Model fit measurement

a) The ρ^2 - metric

$$\rho^2 = 1 - \frac{LL(\hat{\beta})}{LL(0)}$$

$LL(\hat{\beta})$: Value of the log-likelihood at the converged set of parameters.

$LL(0)$: Likelihood corresponding to no parameters being estimated.

- What is $LL(0)$?

  Suppose there are four choice alternatives with no information about the X variables (feature, display, etc.), the predicted choice probability would be 0.25 for the chosen brand. Then $LL(0) = N \cdot ln(0.25)$ where N is the number of observations in the data. It can be argued however that we have more information in the choice data than the "chance" assumption of 0.25. In principle I can compute the share of each of the 4 brands in the data and use them to compute $LL(0)$. This is equivalent to estimating a logit model only with the intercept or intrinsic brand preference.

- The ρ^2 is loosely interpreted as an R^2 type metric although the analogy is not too clear. It does lie between 0 and 1 though. Like $\bar{R}^2$ there is also a $\bar{\rho}^2$ measure that accounts for the number (K) of estimated parameters,

$$\bar{\rho}^2 = 1 - \frac{LL(\hat{\beta}) - K}{LL(0)}$$

  Note that since $LL(\cdot)$ is negative, $\bar{\rho}^2 < \rho^2$.

b) The Akaike information criterion

To choose among nested models, a variety of measures exist, the AIC is one of them

$$AIC = -2L(\hat{\beta}) + 2K$$

Then pick model with smallest AIC.

c) HQ/BIC/CAIC

One problem with $\bar{\rho}^2$ and AIC is that as the number of observations goes up, one is more likely to pick models with more parameters. To avoid this over-parameterization, several other criteria have been offered to penalize for number of observations (N). We discussed only BIC in class

Hannan - Quinn (HQ) Criterion $= -2L(\hat{\beta}) + 2Kln(ln(N))$

Bayesian Information Criterion (BIC) $= -2L(\hat{\beta}) + Kln(N)$

Consistent Akaike Information Criterion (CAIC) $= -2L(\hat{\beta}) + K(ln(N) + 1)$

Criteria in **b)** and **c)** can be used in Nested Hypotheses Testing.

3 Welfare calculation

Often, we would like to compute a measure of consumer surplus. This is the utility in dollar terms that a person derives in a choice situation, one in which the consumer chooses the brand that gives the highest utility.

$$CS_{it} = \frac{1}{\beta_p} max_j(U_{ijt}, \forall\, j)$$

β_p : MU or marginal utility of income, usually the price parameter.

From the researcher's perspective

$$CS_{it} = \frac{1}{\beta_p} E(max_j(V_{ijt} + \epsilon_{ijt}))$$

For the logit model, Small & Rosen (1981) show

$$CS_{it} = \frac{1}{\beta_p} \ln(\sum_{j=1}^{J} e^{V_{ijt}}) + C'$$

C : Unknown constant as the absolute value of utility cannot be measured.

$\ln(\sum e^{V_{ijt}})$ is also referred to as the "log-sum" parameter.

4 Breaking free of the IIA

i) Tversky (1972 a, b)

Products are made up of subsets of discrete attributes. Each attribute has a value associated with it. To compute the probability that a consumer chooses a product, it is assumed that the consumer does the following.

- Choose an <u>attribute</u> based on its relative value.
- Eliminate all alternatives <u>not</u> containing that attribute.
- Go to the next attribute and repeat until there is only 1 alternative left.

How does this work?

$Z_1 = \{V_1,\, V_4,\, V_5\}$, alternative 1 has attributes 1, 4 and 5 of value V_1, V_4, V_5.

$Z_2 = \{V_2,\, V_4,\, V_6\}$

$Z_3 = \{V_3,\, V_5,\, V_6\}$

So how does consumer end up choosing alternative 1?

Alternative 1 can be chosen in 3 ways:

1) Consumer picks attribute 1
2) Consumer picks attribute 4
3) Consumer picks attribute 5

$$P(picking\, attribute\, 1) = \frac{V_1}{\sum_{j=1}^{6} V_j = V_1 + V_2 + V_3 + V_4 + V_5 + V_6}$$

$$P(picking\, attribute\, 4) = \frac{V_4}{\sum_{j=1}^{6} V_j}$$

But picking attribute 4 doesn't guarantee alternative 1 is chosen. Why? It shares the attribute with alternative 2. So

$$P(picking\, 1|picking\, 4) = \frac{V_1 + V_5}{V_1 + V_5 + V_2 + V_6}$$

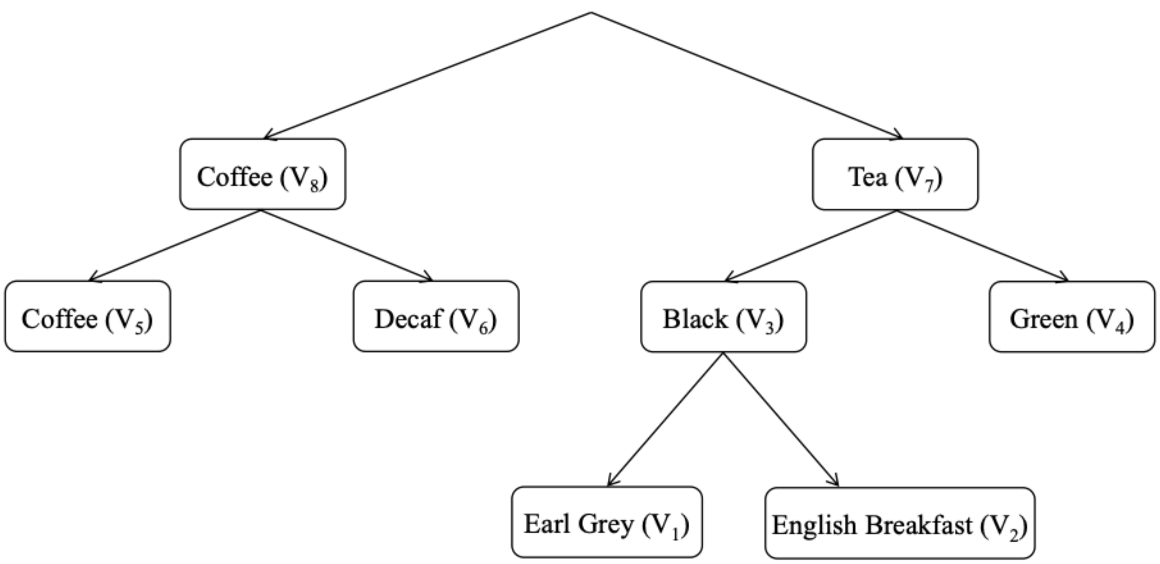
$$P(picking\, 5) = \frac{V_5}{\sum_{j=1}^{6} V_j}$$

$$P(picking\, 1|picking\, 5) = \frac{V_1 + V_4}{V_1 + V_4 + V_3 + V_6}$$

$$P(alternative\, 1) = \frac{V_1}{\sum_{j=1}^{6} V_j} + \frac{V_4}{\sum_{j=1}^{6} V_j} \cdot \frac{V_1 + V_5}{V_1 + V_5 + V_2 + V_6} + \frac{V_5}{\sum_{j=1}^{6} V_j} \cdot \frac{V_1 + V_4}{V_1 + V_4 + V_3 + V_6}$$

ii) Tversky & Sattath (1979)

Often, choices follow a "tree" structure, where each branch of the tree has some unique value to the consumer. Let's look at tea vs. coffee:



What is the probability of picking Earl Grey?

Hierarchy: Tea→Black→Earl Grey

Start at the bottom and work upwards.

$$\begin{aligned} P(Earl\, Grey) =& P(Earl\, Grey\,|\,Black\, Tea) \cdot P(Black\, Tea\,|\,Tea) \cdot P(Tea) \\ =& \frac{V_1}{V_1 + V_2} \cdot \frac{V_1 + V_2 + V_3}{V_1 + V_2 + V_3 + V_4} \cdot \frac{V_1 + V_2 + V_3 + V_4 + V_7}{\sum all\, attributes} \end{aligned}$$

$$where \sum = V_1 + V_2 + V_3 + V_4 + V_5 + V_6 + V_7 + V_8$$

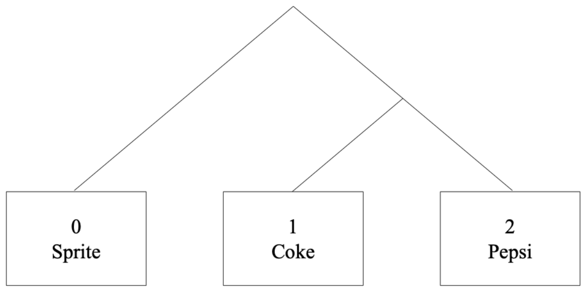
iii) Nested logit model

Back to our IIA example,

Initial market: {Coke, Sprite}

Pepsi enters: {Coke, Sprite, Pepsi}

Pepsi more similar to Coke, so expected to draw more share from Coke, a.k.a., Coke and Pepsi are similar, i.e., their utilities are correlated.



Under independence, $F(\epsilon_1, \epsilon_2) = e^{-e^{-\epsilon_1}} \cdot e^{-e^{-\epsilon_2}} = e^{-[e^{-\epsilon_1} + e^{-\epsilon_2}]}$. However, $F(\epsilon_1, \epsilon_2)$ in nested logit model is no longer $F(\epsilon_1) \cdot F(\epsilon_2)$.

Let's introduce a little bit of correlation

$$0 < \rho < 1$$

$$F(\epsilon_1, \epsilon_2) = e^{-[e^{-\frac{\epsilon_1}{\rho}} + e^{\frac{-\epsilon_2}{\rho}}]^{\rho}}$$

$$corr(\epsilon_1, \epsilon_2) \approx 1 - \rho^2$$

$$
\begin{aligned}
P(0) &= P(U_1 < U_0 \,\&\, U_2 < U_0) = P(V_1+\epsilon_1 < V_0+\epsilon_0 \,\&\, V_2+\epsilon_2 < V_0+\epsilon_0) \\
&= P(\epsilon_1 < (V_0-V_1)+\epsilon_0 \,\&\, \epsilon_2 < (V_0-V_2)+\epsilon_0) \\
&= \int_{-\infty}^{\infty} f(\epsilon_0)\left[\int_{-\infty}^{(V_0-V_1)+\epsilon_0}\int_{-\infty}^{(V_0-V_2)+\epsilon_0} f(\epsilon_1,\epsilon_2)d\epsilon_1 d\epsilon_2\right] d\epsilon_0 \\
&= \int_{-\infty}^{\infty} e^{-\epsilon_0}e^{-e^{-\epsilon_0}} \cdot e^{-\left[e^{-\frac{(V_0-V_1)+\epsilon_0}{\rho}}+e^{-\frac{(V_0-V_2)+\epsilon_0}{\rho}}\right]^{\rho}} d\epsilon_0 \\
&= \int_{-\infty}^{\infty} e^{-\epsilon_0}e^{-e^{-\epsilon_0}} \cdot e^{-e^{-\epsilon_0}\cdot\left[\underbrace{e^{\frac{-(V_0-V_1)}{\rho}}+e^{\frac{-(V_0-V_2)}{\rho}}}_{K}\right]^{\rho}} d\epsilon_0 \\
&= \int_{-\infty}^{\infty} e^{-\epsilon_0}e^{-e^{-\epsilon_0}} \cdot e^{-e^{-\epsilon_0}\cdot K} d\epsilon_0 \\
&= \int_{-\infty}^{\infty} e^{-\epsilon_0}e^{-(1+K)\cdot e^{-\epsilon_0}} d\epsilon_0 \\
&= \frac{1}{1+K}\underbrace{\int_{-\infty}^{\infty}(1+K)e^{-\epsilon_0}e^{(1+K)e^{-\epsilon_0}}d\epsilon_0}_{\text{same as logit }=1} \\
&= \frac{1}{1+\left[e^{\frac{-(V_0-V_1)}{\rho}}+e^{\frac{-(V_0-V_2)}{\rho}}\right]^{\rho}} \\
&= \frac{e^{V_0}}{e^{V_0}+\left[e^{V_1/\rho}+e^{V_2/\rho}\right]^{\rho}} \\
P(1) &= \frac{\left[e^{V_1/\rho}+e^{V_2/\rho}\right]^{\rho}}{e^{V_0}+\left[e^{V_1/\rho}+e^{V_2/\rho}\right]^{\rho}}\frac{e^{V_1/\rho}}{e^{V_1/\rho}+e^{V_2/\rho}}
\end{aligned}
$$

4.0.1 Estimating the nested logit model

What is observed

ID	Brand 1	Brand 2	Brand 3	Brand 4
1	0	0	0	1
1	0	1	0	0
1	0	1	0	0
2	1	0	0	0
2	1	0	0	0
2	0	1	0	0

What the model predicts

ID	P(Brand 1)	P(Brand 2)	P(Brand 3)	P(Brand 4)
1	P_{11}	P_{21}	P_{31}	P_{41}
1	P_{12}	P_{22}	P_{32}	P_{42}
1	...	...	...	...
2				
2				
2				

- Focusing on the first observation, we see that the consumer bought brand 4.
- What is the estimation task? To figure out $\{\alpha_1,\, \alpha_2,\, \alpha_3,\, \alpha_{4,}\, \beta_f,\, \beta_p, \rho\}$. There are two variables, feature and price in the data.
- For identification, we need to set one of the $\alpha's = 0$, Why? Adding or subtracting the same number from $V_{ijt}, \forall j$ will leave the choice probability unchanged, so there are 6 parameters to estimate $\{\alpha_1, \alpha_2, \alpha_3, \beta_f, \beta_p, \rho\}$ if we set $\alpha_4 = 0$.
- Since write

$$0 < \rho < 1$$

$$\rho = \frac{e^{\lambda}}{[1+e^{\lambda}]}$$

$$-\infty < \lambda < \infty$$

- so we write it in terms of the unconstrained parameter and estimate that.

- To choose the parameters, we want to do them like the logit, i.e., in such a way that the model's predictions match up as closely as possible to the observed data. In other words, e.g., we would like to make P_{41} as close to 1 as possible, P_{22} as close to 1 as possible, etc.

  IOW, we want to maximize the probability of the chosen brands. To accomplish that, we construct the likelihood function - the collection of the probabilities of the brand chosen on each occasion $P_{41}, P_{22}, P_{23,}$... To construct a simple entity to maximize, we recognize the independence of each observation and create the likelikhood function as:

$$L = \prod_{i=1}^{N} \prod_{t=1}^{T_i} \left[\prod_{j=1}^{J} P_{ijt}^{\delta_{ijt}} \right]$$

  $i = 1, 2, ..., N$ households; $j = 1, 2, .., J$ brands; $t = 1, 2, ..., T_i$ choice occasions for household i.

  where $\delta_{ijt} = 1$ if brand j is purchased at t by household i; 0 otherwise.

  Practically this is equivalent to maximizing the log of the above function

$$LL = ln(L) = \sum_{i=1}^{N} \sum_{t=1}^{Ti} \sum_{j=1}^{J} \delta_{ijt} ln(P_{ijt})$$

- So to estimate the model parameters, we use a maximization algorithm to pick values of $\Theta = \{\alpha_1, \alpha_2, \alpha_3, \beta_f, \beta_p, \lambda\}$ that maximizes $LL(\Theta)$.