Estimating the preference for safety features in cars

Tool: Multinomial logit & Mixed logit

The Analytics Edge

Companies such as General Motors carry out conjoint studies with customers to understand the tradeoff (valuation) of different attributes that make up a product or a service. Using data on preferences for safety features in new vehicles and then building models of discrete choice, the company can obtain estimates on the valuation of attributes. This provides important information that can be used to infer the effect of introducing new products in the market, pricing the product and designing the product features.

The models incorporates the comparison of attributes across alternatives and can also be used to capture heterogeneity in the choice modeling process.

Which of the following packages would you prefer the most?

Choose by clicking one of the buttons below

Full Speed Range Adaptive Cruise Control	Adaptive Cruise Control	Full Speed Range Adaptive Cruise Control	NONE: I wouldn't purchase any of these packages
-	Navigation System with Curve Notification & Speed Advisor	Traditional Navigation System	
Traditional Back Up Aid	Rear Vision System	-	
Lane Departure Warning	-	-	
Front Collision Warning	-	Front Collision Warning	
Side Head Air Bags	Side Body Air Bags	Side Body & Head Air Bags	
Emergency Notification with Pictures	Emergency Notification	-	
-	-	Night Vision with Pedestrian Detection	
-	Head Up Display	Head Up Display	
Option Package Price: $3, 000	Option Package Price: $500	Option Package Price: $12, 000	
○	○	○	○

Analytics on safety feature data : R

Safety <- read.csv("safety data.csv")

Str (safety)

Summary (safety)

head (safety)

9500 observations of 112 variables

Each customer performs 19 choice tasks with a total of 500 customers

Case = Customer number (1 to 500)

No = Observation number (1 to 9500)

Task = Task number for a customer (1 to 19)

Each customer has a choice among four packages in each choice task. (Stated Preference data set)

For example customer 1 in the first choice task chooses Ch2

Attribute levels	Alternative 1 (Ch 1)	Alternative 2 (Ch 2)	Alternative 3 (Ch 3)	Alternative 4 (Ch 4)
GN	1	2	2	0
NS	4	1	4	0
BU	6	2	5	0
FP	3	2	3	0
RP	1	1	3	0
PP	1	1	3	0
TS	1	1	3	0
NV	1	1	1	0
MA	4	1	1	0
Price	2	2	2	
	level of attributes in package 1	Package 2	Package 3	All at zero level

CC	Cruise control	3 levels
GN	Go notifier	2 levels
NS	Navigation system	5 levels
BU	Backup aids	6 levels
FA	Front park assist	2 levels
LD	Lane departure	3 levels
BZ	Blind zone alert	3 levels
FC	Front collision warning	2 levels
FP	Front collision protection	4 levels
RP	Rear collision protection	2 levels
PP	Parallel park aids	3 levels
KA	Knee air bags	2 levels
SC	Side airbags	4 levels
TS	Emergency notification	3 levels
NV	Night vision system	3 levels
MA	Driver assisted adjustment	4 levels
LB	Low speed braking assist	4 levels
AF	Adaptive front lighting	3 levels
HU	Head up display	2 levels
Price	Price	11 levels

($500, 1000, 1500, 2000, 2500, 3000, 4000, 5000, 7500, 10000, 12000)

The variables from "segment" to "incomea" give demographic information

Ch 1 = 1 if package 1 is chosen, 0 o/w

Ch 2 = 1 if package 2 is chosen, 0 o/w

Ch 3 = 1 if package 3 is chosen, 0 o/w

Ch 4 = 1 if no package is chosen, 0 o/w

$$\text{Choice} = \begin{cases} \text{Ch 1} \\ \text{Ch 2} \\ \text{Ch 3} \\ \text{Ch 4} \end{cases}$$ Depending on which option is chosen

table (safety $ Choice)

Ch1	Ch 2	Ch3	Ch4
2165	2404	2136	2795

Segment = Segment of car population that individual has

Year = Year

Miles = Mileage

Night = % of time customer drives at night

Gender = Gender (Female or Male)

Age = Age bracket

Educ = Education level (4 year college, grade school, high school, postgrad, college 1-3 years, voc

region = resident of region (MW, NE, SE, SW, W)

Urban = resident of Rural, Suburban, Urban

Income = Income

which (colnames (safety) == "CC1")   4

which (colnames (safety) == "Price 4")   83

Columns 4 to 83 of the safety data frame contains the attribute levels for each of the 4 alternatives (choices) indexed by 1, 2, 3 and 4 respectively. Note the zero level in all cases indicate the attribute is not available with the fourth choice being the NONE option

In general higher levels for attributes correspond to more technically advanced features.

We can capture these variables either directly or by adding dummy variables that take a value of 1 or 0 depending on the level.

Developing a logit model for the data

```
S <- mlogit.data(subset(safety, Task <= 12),
       shape="wide", choice = "Choice",
       varying = c(4:83), sep = "",
       alt.levels = c("Ch1", "Ch2", "Ch3", "Ch4"),
       id.var = "Case")
```

This command takes the dataframe where we use the first 12 choice tasks for each customer in our training set (12 out of 19), indicates that the shape of the data frame is wide (where each row is an observation), indicates that the variable indicating the choice made is defined by "Choice", the separator of the variable name and the alternative name (this helps to guess variables and alternative names), varying = c(4:83) (helps indicate the variables indices that are alternative specific), alt.levels indicates the names of the alternatives and id.var indicates the name of the variable containing the individual index (here "Case")

head (S)
str (S)

This creates a data frame of the long format to which we can use the mlogit package (24000 = 6000 × 4 observations)

M <- mlogit (Choice ~ CC + GN + NS + BU + FA + LD + B7 + FC + FP + RP + PP + KA + SC + TS + NV + MA + LB + AF + MU + Price − 1, data = S)

Summary (M)

Log-likelihood at optimality $LL(\hat{\beta}) = -7567.9$

$$AIC = -2LL(\hat{\beta}) + 2(\text{Parameters}) = 15175.8$$

$$\text{Likelihood ratio} = 1 - \frac{LL(\hat{\beta})}{LL(0)}$$

$$= 1 - \left(\frac{-7567.9}{6000 \log\left(\frac{1}{4}\right)} \right)$$

$$= 1 - \left(\frac{-7567.9}{-8317.76} \right) = 0.09$$

Results indicate that

CC (cruise control), KA (knee air bags),

TS (emergency notification), MA (driver adjusted assesment), Price are very significant at 0.001 level.

As should be expected Price has a negative coefficient.

LD (lane departure) and SC (side air bags) are significant at 0.01 level.

The non price coefficients have positive signs indicating that these are valued (higher levels of safety features).

To check the correct predictions in sample simply based on the maximum predicted choice probability

```
ActualChoice <- Subset(safety, Task ≤ 12)[, "Choice"]
P <- predict(M, newdata = S)
PredictedChoice <- apply(P, 1, which.max)
```

Actual choice keeps track of actual observed choice

P predicts the choice probability for the given dataset

The apply function applies to the matrix P, across rows (indexed by 1), the function of which.max (identifies index that is maximum)

Willingness to pay

Ratio of β_j coefficients can be used to estimate the willingness to pay for a particular attribute

Suppose β_1 is the coefficient for attribute 1 and β_2 is the coefficient of attribute 2 (price)

Note that β_2 will be typically negative.

Since more the price lesser the utility.

Assume say β_1 is positive.

$$U = \beta_1 X_1 + \beta_2 X_2 + \cdots$$

Assume unit change (increase) in attribute 1

$$U = \beta_1 (X_1 + 1) + \beta_2 (X + \Delta) + \cdots$$

Then $\Delta = -\dfrac{\beta_1}{\beta_2}$ (Willingness to pay for unit increase in attribute)

Example: Willingness to pay for CC.
Cruise control.

$$WTP = \frac{0.1085}{0.2036} = 0.5329$$

table(Predicted Choice, Actual Choice)

Predicted \ Actual	Ch1	Ch2	Ch3	Ch4
1	616	273	226	356
2	222	647	235	320
3	194	258	629	309
4	339	352	324	700

Correct predictions $= \dfrac{616 + 647 + 629 + 700}{6000}$

$= \boxed{0.432}$

Assuming a complete random choice for predicting the choice would result in $\boxed{0.25}$.

```
Test <- mlogit.data(subset(safety, Task > 12),
        shape = "wide", choice = "Choice",
        varying = c(4:83), sep = "", alt.levels =
        c("Ch1", "Ch2", "Ch3", "Ch4"), id.var = "Case")

Test.Predict <- predict(M, newdata = Test)
Actual Choice <- subset(safety, Task > 12)[, "Choice"]
Predicted Choice <- apply(TestPredict, 1, which.max)
table(Predicted Choice, Actual Choice)
```

	Ch1	Ch2	Ch3	Ch4
1	376	120	93	176
2	119	435	97	227
3	118	124	356	190
4	181	193	176	517

Correct predictions $= \dfrac{(376 + 435 + 356 + 517)}{3500}$

$= \boxed{0.481}$

Mixed logit model

To capture additional features that multinomial logit cannot capture such as random taste variation.

Standard logit $\quad \tilde{U}_{ik} = \beta' x_{ik} + \epsilon_{ik}$

In mixed logit, $\tilde{\beta}$ is modeled as a random parameter. $\quad \tilde{U}_{ik} = \tilde{\beta}' x_{ik} + \tilde{\epsilon}_{ik}$

Standard logit

$$P(Y_i = k) = \frac{e^{\beta' x_{ik}}}{\sum_{l=1}^{K} e^{\beta' x_{il}}}$$

Mixed logit

$$P(Y_i = k) = \int \frac{e^{\beta' x_{ik}}}{\sum_{l=1}^{k} e^{\beta' x_{il}}} f(\beta)\, d\beta$$

where $f(\beta)$ is the density function of $\tilde{\beta}$.

Integral of logit probabilities

Mixed logit is computationally more challenging to solve due to the use of simulation optimization methods.

The problems are no longer convex in this setting and finding a global optimum might not

From an estimation perspective, the goal is to find the parameters Θ that define the density function $f(\beta|\Theta)$ where the functional form $f(\cdot)$ is given but parameters Θ are unknown.

To approximate the probability value given a particular Θ. Draw multiple β_r vectors from the distribution $f(\beta|\Theta)$

$$P(Y_i = k) \simeq \sum_{r=1}^{R} \frac{e^{\beta_r' x_{ik}}}{\sum_{l=1}^{K} e^{\beta_r' x_{il}}}$$

Plug this approximation into the log-likelihood objective function and estimate Θ value by doing optimization.

For mixed logit with repeated choice (panel data)

i = individual $\quad t$ = observation
k = alternative

$$P(Y_{i1} = k_1, \ldots, Y_{iT} = k_T)$$

$$= \int \prod_{t=1}^{T} \left(\frac{e^{\beta' x_{ik_i t}}}{\sum_{l=1}^{K} e^{\beta' x_{ilt}}} \right) f(\beta)\, d\beta$$

Panel data needs to account for the fact that the errors are correlated for the same individual over time.

Building more sophisticated models (even more)

```
M1 <- mlogit (Choice ~ CC + GN + NS + BU + FA + LD
      + BZ + FC + FP + RP + PP + KA + SC + TS + NV + MA
      + LB + AF + HU + PRICE - 1, data = S,
      rpar = c( CC = 'n', GN = 'n', NS = 'n', BU = 'n',
      FA = 'n', LD = 'n', BZ = 'n', FC = 'n', FP = 'n',
      RP = 'n', PP = 'n', KA = 'n', SC = 'n', TS = 'n',
      NV = 'n', MA = 'n', LB = 'n', AF = 'n', HU = 'n',
      PRICE = 'n'), panel = TRUE, print.level =
      TRUE )
```

This fits a mixed logit model where the coefficients are treated as random variables (This is captured with rpar (random parameters) argument where 'n' indicates it is modeled as a normal random variable; panel data captures the fact that we have multiple observations per individual.

Summary (M1)

This estimates for each random coefficient, a mean value and a standard deviation

Loglikelihood = −6531.3

P1 <- predict (M1, newdata = S)

PredictedChoice1 <- apply (P1, 1, which.max)

table (PredictedChoice1, Actual)

	Ch1	Ch2	Ch3	Ch4
1	530	243	179	288
2	203	552	207	256
3	173	220	537	236
4	465	515	419	905

$$\text{Correct predictions} = \frac{530+\cancel{352}\,552+537+905}{6000}$$

$$= \boxed{0.420}$$

Note that while the log-likelihood for the mixed logit model is better, in terms of predicting by simply looking at highest probability it is worse than MNL in this example.

TestPredict1 <- predict (M1, newdata = Test)
ActualChoice1 <- subset (Safety, Task > 12) [, "Choice"]
PredictedChoice1 <- apply (Testpredict, 1, which.max)
table (PredictedChoice1, ActualChoice1)

331	116	77	137
102	377	83	180
98	110	316	149
263	271	246	644

Correct predictions

$$= \frac{331+377+316+644}{3500}$$

$$= \boxed{0.4765}$$

The mixed logit model does a better job of predicting customers who are not interested in choosing any one of the offered options compared to MNL.

We can also compare out of sample (test) log-likelihood.

ProbPredict <- apply(TestPredict, 1, max)

MNL

sum(log(ProbPredict)) $\boxed{-2703.78}$

ProbPredict1 <- apply(TestPredict1, 1, max)

Mixed logit

Sum(log(ProbPredict1)) $\boxed{-3086.894}$