Lecture 5: Logit Model II

ResEcon 703: Topics in Advanced Econometrics

Matt Woerman
University of Massachusetts Amherst

Agenda

Last time

- Logit Model

Today

- More Logit Properties
- Multinomial Logit Model Example in R
- Adamowicz et al. (1994)

Upcoming

- Reading for next time
  - Greene textbook, Chapters 7.1-7.2.1 and 14.1-14.6
- Problem set
  - Problem Set 1 is posted, due September 24

Logit Model Recap

Simple assumption about the joint distribution of unobserved utility

$$\varepsilon_{nj} \sim \text{i.i.d. type I extreme value (Gumbel) with } Var(\varepsilon_{nj}) = \frac{\pi^2}{6}$$

Choice probabilities have a closed-form expression

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

Some properties

- Coefficients are marginal utilities, not marginal effects
- Taste variation is limited to observed characteristics
- Coefficients are estimated relative to the scale parameter

More Logit Properties

Independence of Irrelevant Alternatives

The ratio of any two choice probabilities is

$$\frac{P_{ni}}{P_{nk}} = \frac{e^{V_{ni}}}{e^{V_{nk}}}$$

This ratio only depends on characteristics of alternatives i and k, so it is considered to be independent of irrelevant alternatives (IIA)

When IIA holds, you can estimate consistent parameters using only a subset of alternatives for each decision maker

- When the choice set is too large to be computationally feasible, you only have to consider a subset of alternatives
- When you only care about a subset of the choice set, you can ignore decision makers who choose the other alternatives

Red-Bus/Blue-Bus Problem

Two travel modes to commute to work: car and blue bus

- For simplicity, assume choice probabilities are equal

$$P_c = P_{bb} = \frac{1}{2} \quad \Rightarrow \quad \frac{P_c}{P_{bb}} = 1$$

Now suppose a red bus is introduced with all characteristics identical to the blue bus except the color

- Assuming the commuter does not care about the color of the bus

$$P_{rb} = P_{bb} \quad \Rightarrow \quad \frac{P_{rb}}{P_{bb}} = 1$$

But from IIA, the ratio of car and blue bus is not changed by the introduction of the red bus

- The choice probability for all three modes must be equal

$$P_c = P_{bb} = P_{rb} = \frac{1}{3}$$

Should the introduction of a red bus change the probability of driving?

Proportional Substitution

From last time: The cross-elasticity of P_{ni}, the choice probability of alternative i, with respect to z_{nj}, an observed factor of alternative j, is given by (assuming linearity of representative utility)

$$E_{iz_{nj}} = -\beta_z z_{nj} P_{nj}$$

This cross-elasticity depends only on features of alternative j and not on features of alternative i

- This cross-elasticity is the same for every alternative other than j!

When an attribute of one alternative changes, all other choice probabilities are changed by the same percentage (not percentage points)

- That is, substitution to other alternatives is proportional to their original choice probabilities

Proportional Substitution Example

Hummer H2

Cadillac Escalade

Smart Pure EV

Suppose Hummer lowers the price of the H2 to gain more market share

- Will that attract a greater proportion of Excalade drivers of Pure EV drivers?
- The logit model says that substitution to the H2 will be proportionally equal for these very different vehicles!

Panel Data

If we observe panel data for a discrete choice problem, we can add a time index to our random utility model and logit choice probabilities

$$U_{njt} = V_{njt} + \varepsilon_{njt} \quad \Rightarrow \quad P_{nit} = \frac{e^{V_{nit}}}{\sum_j e^{V_{njt}}}$$

We can estimate this model just as in the cross-section

- We can also include lagged or future variables to capture "dynamics"
- We can even include previous choices as explanatory variables to represent behavioral factors like habit formation

The logit assumption still has to hold

$$\varepsilon_{njt} \sim \text{i.i.d. type I extreme value (Gumbel) with } Var(\varepsilon_{njt}) = \frac{\pi^2}{6}$$

- But the unobserved characteristics of a decision maker that affect choice are unlikely to be independent over time

Consumer Surplus

The logit model gives a closed-form expression for consumer surplus that depends on α_n, the marginal utility of income for individual n

$$CS_n = \frac{1}{\alpha_n} \max_j (U_{nj})$$

We do not observe U_{nj}, but we know it in expectation

$$E(CS_n) = \frac{1}{\alpha_n} E\left[\max_j (V_{nj} + \varepsilon_{nj})\right]$$

If we further assume utility is linear in income, we get

$$E(CS_n) = \frac{1}{\alpha_n} \ln\left(\sum_{j=1}^{J} e^{V_{nj}}\right) + C$$

So the change in consumer surplus due to changes in alternatives is

$$\Delta E(CS_n) = \frac{1}{\alpha_n}\left[\ln\left(\sum_{j=1}^{J^1} e^{V_{nj}^1}\right) - \ln\left(\sum_{j=1}^{J^0} e^{V_{nj}^0}\right)\right]$$

Market-Level Data

The logit model can also be estimated from market-level data

- You observe the price, market share, and characteristics of every cereal brand at the grocery store, and you want to estimate the structural parameters of consumer decision making that explain those purchases

When aggregated over many consumers, choice probabilities become market shares

$$S_i = \frac{e^{V_i}}{\sum_j e^{V_j}}$$

If we assume representative utility is liner, $V_j = \beta' x_j$

$$\ln(S_i) - \ln(S_j) = \beta'(x_i - x_j)$$

Set one alternative to be your reference (usually the outside option) and estimate the linear regression for the other $J - 1$ alternatives

$$\ln(S_i) - \ln(S_0) = \beta'(x_i - x_0) + \omega_i$$

Multinomial Logit Model Example in R

Multinomial Logit Model Example

We are again studying how consumers make choices about expensive and highly energy-consuming systems in their homes. We have data on 900 households in California and the type of heating system in their home. Each household has the following choice set, and we observe the following data

Choice set

- GC: gas central
- GR: gas room
- EC: electric central
- ER: electric room
- HP: heat pump

Alternative-specific data

- IC: installation cost
- OC: annual operating cost

Household demographic data

- income: annual income
- agehed: age of household head
- rooms: number of rooms
- region: home location

Load Dataset

```
## Load tidyverse and mlogit
library(tidyverse)
library(mlogit)
## Load dataset from mlogit package
data('Heating', package = 'mlogit')
```

Dataset

```
## Look at dataset
as_tibble(Heating)
## # A tibble: 900 x 16
##    idcase depvar ic.gc ic.gr ic.ec ic.er ic.hp oc.gc oc.gr oc.ec oc.er
##     <dbl> <fct>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1      1 gc       866   963.  860.  996. 1136.  200.  152.  553.  506.
##  2      2 gc       728.  759.  797.  895.  969.  169.  169.  520.  486.
##  3      3 gc       599.  783.  720.  900. 1048.  166.  138.  439.  405.
##  4      4 er       835.  793.  761.  831. 1049.  181.  147.  483   425.
##  5      5 er       756.  846.  859.  986.  883.  175.  139.  404.  390.
##  6      6 gc       666.  842.  694.  863.  859.  136.  141.  398.  371.
##  7      7 gc       670.  941.  634.  952. 1087.  192.  148.  478.  446.
##  8      8 gc       778. 1022.  813. 1012.  990.  188.  159.  502.  465.
##  9      9 gc       928. 1212.  876. 1025. 1232.  169.  190.  553.  452.
## 10     10 gc       683. 1045.  776.  874.  878.  176.  136.  532.  472.
## # ... with 890 more rows, and 5 more variables: oc.hp <dbl>,
## #   income <dbl>, agehed <dbl>, rooms <dbl>, region <fct>
```

Convert to a Long Dataset

```
## Gather into a long dataset
heating_long <- Heating %>%
  gather(key, value, starts_with('ic.'), starts_with('oc.')) %>%
  separate(key, c('cost', 'alt')) %>%
  spread(cost, value) %>%
  mutate(choice = (depvar == alt)) %>%
  select(-depvar)
```

Long Dataset

```
## Look at long dataset
as_tibble(heating_long)
## # A tibble: 4,500 x 9
##    idcase income agehed rooms region alt      ic    oc choice
##     <dbl>  <dbl>  <dbl> <dbl> <fct>  <chr> <dbl> <dbl> <lgl>
##  1      1      7     25     6 ncostl ec     860.  553. FALSE
##  2      1      7     25     6 ncostl er     996.  506. FALSE
##  3      1      7     25     6 ncostl gc     866   200. TRUE
##  4      1      7     25     6 ncostl gr     963.  152. FALSE
##  5      1      7     25     6 ncostl hp    1136.  238. FALSE
##  6      2      5     60     5 scostl ec     797.  520. FALSE
##  7      2      5     60     5 scostl er     895.  486. FALSE
##  8      2      5     60     5 scostl gc     728.  169. TRUE
##  9      2      5     60     5 scostl gr     759.  169. FALSE
## 10      2      5     60     5 scostl hp     969.  199. FALSE
## # ... with 4,490 more rows
```

Convert Datasets to `mlogit` Format

```
### Convert data to mlogit format
## Convert wide data to mlogit format
heating_mlogit <- mlogit.data(Heating, shape = 'wide',
                              choice = 'depvar', varying = 3:12)
## Convert long data to mlogit format
heating_long_mlogit <- mlogit.data(heating_long, shape = 'long',
                                   choice = 'choice', alt.var = 'alt')
```

Wide Dataset in `mlogit` Format

```
## Look at wide data in mlogit format
as_tibble(heating_mlogit)
## # A tibble: 4,500 x 10
##    idcase depvar income agehed rooms region alt      ic    oc  chid
##     <dbl> <lgl>   <dbl>  <dbl> <dbl> <fct>  <chr> <dbl> <dbl> <int>
##  1      1 FALSE       7     25     6 ncostl ec     860.  553.     1
##  2      1 FALSE       7     25     6 ncostl er     996.  506.     1
##  3      1 TRUE        7     25     6 ncostl gc     866   200.     1
##  4      1 FALSE       7     25     6 ncostl gr     963.  152.     1
##  5      1 FALSE       7     25     6 ncostl hp    1136.  238.     1
##  6      2 FALSE       5     60     5 scostl ec     797.  520.     2
##  7      2 FALSE       5     60     5 scostl er     895.  486.     2
##  8      2 TRUE        5     60     5 scostl gc     728.  169.     2
##  9      2 FALSE       5     60     5 scostl gr     759.  169.     2
## 10      2 FALSE       5     60     5 scostl hp     969.  199.     2
## # ... with 4,490 more rows
```

Long Dataset in `mlogit` Format

```
## Look at long data in mlogit format
as_tibble(heating_long_mlogit)
## # A tibble: 4,500 x 9
##    idcase income agehed rooms region alt      ic    oc choice
##     <dbl>  <dbl>  <dbl> <dbl> <fct>  <fct> <dbl> <dbl> <lgl>
##  1      1      7     25     6 ncostl ec     860.  553. FALSE
##  2      1      7     25     6 ncostl er     996.  506. FALSE
##  3      1      7     25     6 ncostl gc     866   200. TRUE
##  4      1      7     25     6 ncostl gr     963.  152. FALSE
##  5      1      7     25     6 ncostl hp    1136.  238. FALSE
##  6      2      5     60     5 scostl ec     797.  520. FALSE
##  7      2      5     60     5 scostl er     895.  486. FALSE
##  8      2      5     60     5 scostl gc     728.  169. TRUE
##  9      2      5     60     5 scostl gr     759.  169. FALSE
## 10      2      5     60     5 scostl hp     969.  199. FALSE
## # ... with 4,490 more rows
```

Multinomial Logit Model Estimation

```
### Model heating choice as a multinomial logit
## Model choice using cost data and alternative effects
model_mlogit <- heating_mlogit %>%
  mlogit(formula = depvar ~ ic + oc | 1 | 0, data = ., reflevel = 'hp')
```

Model Summary

```
## Summarize model results
model_mlogit %>%
  summary()
##
## Call:
## mlogit(formula = depvar ~ ic + oc | 1 | 0, data = ., reflevel = "hp",
##     method = "nr")
##
## Frequencies of alternatives:
##       hp       ec       er       gc       gr
## 0.055556 0.071111 0.093333 0.636667 0.143333
##
## nr method
## 6 iterations, 0h:0m:0s
## g'(-H)^-1g = 9.58E-06
## successive function values within tolerance limits
##
## Coefficients :
##                    Estimate  Std. Error z-value  Pr(>|z|)
## ec:(intercept)  1.65884594  0.44841936  3.6993 0.0002162 ***
## er:(intercept)  1.85343697  0.36195509  5.1206 3.045e-07 ***
## gc:(intercept)  1.71097930  0.22674214  7.5459 4.485e-14 ***
## gr:(intercept)  0.30826328  0.20659222  1.4921 0.1356640
## ic             -0.00153315  0.00062086 -2.4694 0.0135333 *
## oc             -0.00699637  0.00155408 -4.5019 6.734e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -1008.2
## McFadden R^2:  0.013691
## Likelihood ratio test : chisq = 27.99 (p.value = 8.3572e-07)
```

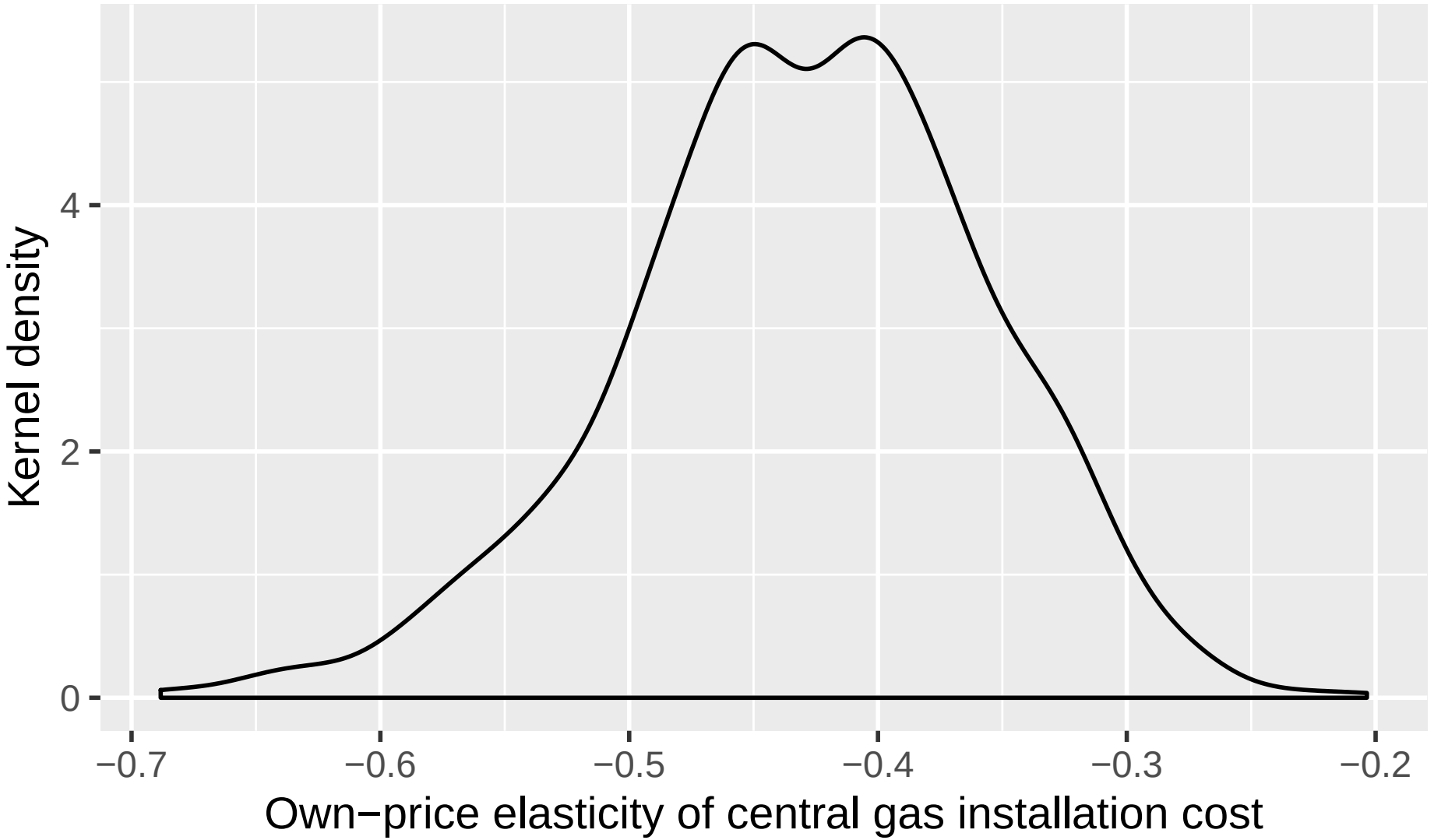
Elasticities

```
### Calculate mean own-price elasticities for central gas
## Calculate probability of central gas
Heating <- Heating %>%
  mutate(prob_gc_mlogit = fitted(model_mlogit,
                                 type = 'probabilities')[, 4])
## Calculate mean own-price elasticities
coef(model_mlogit)[5:6] *
  c(mean(Heating$ic.gc * (1 - Heating$prob_gc_mlogit)),
    mean(Heating$oc.gc * (1 - Heating$prob_gc_mlogit)))
##         ic         oc
## -0.4297750 -0.4347879
```

Distribution of Elasticities

```
### Visualize distibution of own-price elasticity of ic for gc
## Calculate and plot density of own-price elasticity of ic for gc
Heating %>%
  mutate(elasticity = coef(model_mlogit)[5] * ic.gc *
           (1 - prob_gc_mlogit)) %>%
  ggplot(aes(x = elasticity)) +
  geom_density() +
  xlab('Own-price elasticity of central gas installation cost') +
  ylab('Kernel density')
```

Elasticity Distribution Plot



Cross Elasticities

```
### Calculate mean cross-price elasticities of ec with respect to gc
## Calculate mean cross-price elasticity
-coef(model_mlogit)[5:6] *
  c(mean(Heating$ic.gc * Heating$prob_gc_mlogit),
    mean(Heating$oc.gc * Heating$prob_gc_mlogit))
##         ic         oc
## 0.7612192 0.7693978
```

Cost Tradeoffs

How do consumers trade off the installation cost and the annual operating cost?

- What reduction in installation cost offsets a \$1 increase in the annual operating cost?

$$U_{ni} = \alpha_i + \beta_1 IC_{ni} + \beta_2 OC_{ni} + \varepsilon_{ni}$$

$$dU = \beta_1 dIC_{ni} + \beta_2 dOC_{ni}$$

$$dU = 0 \Rightarrow \frac{dIC}{dOC} = -\frac{\beta_2}{\beta_1}$$

```
### Calculate the tradeoff between installation cost and operating cost
## Calculate install cost equivalence of an increse in operating cost
-coef(model_mlogit)[6] / coef(model_mlogit)[5]
##        oc
## -4.563385
```

Implied Discount Rate

What is the implied discount rate of consumers? Assume (for simplicity) that the operating cost accrues in perpetuity.

$$U_{ni} = \alpha_i + \beta_1 IC_{ni} + \beta_2 OC_{ni} + \varepsilon_{ni}$$

Divide by β_1 to express in terms of present day dollars

$$\frac{U_{ni}}{\beta_1} = \frac{\alpha_i}{\beta_1} + IC_{ni} + \frac{\beta_2}{\beta_1} OC_{ni} + \frac{\varepsilon_{ni}}{\beta_1}$$

Re-express using an infinite sum of discounted operating costs

$$\frac{U_{ni}}{\beta_1} = \frac{\alpha_i}{\beta_1} + IC_{ni} + \frac{1}{r} OC_{ni} + \frac{\varepsilon_{ni}}{\beta_1}$$

The last two equations give that

$$r = \frac{\beta_1}{\beta_2}$$

Implied Discount Rate Calculation

```
### Calculte the implied discount rate of consumers
## Calculate the implied discount in perpetuity
coef(model_mlogit)[5] / coef(model_mlogit)[6]
##        ic
## 0.2191356
```

Heterogeneous Coefficients

```
### Model heating choice with heterogeneous cost coefficients
## Model heating choice with costs divided by income
model_mlogit_income <- heating_mlogit %>%
  mlogit(formula = depvar ~ I(ic / income) + I(oc / income) | 1 | 0,
         data = ., reflevel = 'hp')
```

Heterogeneous Coefficients Model Summary

```
## Summarize model results
model_mlogit_income %>%
  summary()
##
## Call:
## mlogit(formula = depvar ~ I(ic/income) + I(oc/income) | 1 | 0,
##     data = ., reflevel = "hp", method = "nr")
##
## Frequencies of alternatives:
##       hp       ec       er       gc       gr
## 0.055556 0.071111 0.093333 0.636667 0.143333
##
## nr method
## 6 iterations, 0h:0m:0s
## g'(-H)^-1g = 1.23E-05
## successive function values within tolerance limits
##
## Coefficients :
##                   Estimate Std. Error z-value  Pr(>|z|)
## ec:(intercept)  0.4570416  0.2867524  1.5939  0.110969
## er:(intercept)  0.7557773  0.2304446  3.2796  0.001039 **
## gc:(intercept)  2.2223040  0.1941569 11.4459 < 2.2e-16 ***
## gr:(intercept)  0.7894157  0.1792350  4.4044 1.061e-05 ***
## I(ic/income)   -0.0022639  0.0019592 -1.1555  0.247874
## I(oc/income)   -0.0053289  0.0027577 -1.9324  0.053315 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -1018.9
## McFadden R^2:  0.0032964
## Likelihood ratio test : chisq = 6.7393 (p.value = 0.034402)
```

Alternative-Specific Coefficients on Demographics

```
### Model heating choice with alternative-specific demographics
## Model heating choice with alternative-specific rooms coefficient
model_mlogit_rooms <- heating_mlogit %>%
  mlogit(formula = depvar ~ ic + oc | rooms | 0, data = .,
         reflevel = 'hp')
```

Alternative-Specific Demographics Model Summary

```
## Summarize model results
model_mlogit_rooms %>%
  summary()
##
## Call:
## mlogit(formula = depvar ~ ic + oc | rooms | 0, data = ., reflevel = "hp",
##     method = "nr")
##
## Frequencies of alternatives:
##       hp       ec       er       gc       gr
## 0.055556 0.071111 0.093333 0.636667 0.143333
##
## nr method
## 6 iterations, 0h:0m:0s
## g'(-H)^-1g = 1.58E-05
## successive function values within tolerance limits
##
## Coefficients :
##                    Estimate  Std. Error z-value  Pr(>|z|)
## ec:(intercept)  1.47832643  0.67123250  2.2024   0.02764 *
## er:(intercept)  1.75977504  0.58813686  2.9921   0.00277 **
## gc:(intercept)  1.77021331  0.44444627  3.9830 6.806e-05 ***
## gr:(intercept)  0.44338366  0.47207112  0.9392   0.34761
## ic             -0.00153161  0.00062169 -2.4636   0.01375 *
## oc             -0.00696375  0.00155563 -4.4765 7.589e-06 ***
## ec:rooms        0.03811449  0.10825890  0.3521   0.72479
## er:rooms        0.01939934  0.10278196  0.1887   0.85029
## gc:rooms       -0.01294329  0.08476424 -0.1527   0.87864
## gr:rooms       -0.03008528  0.09570570 -0.3144   0.75325
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -1007.8
## McFadden R^2:  0.014102
```

Alternative-Specific Coefficients on Costs

```
### Model heating choice with alternative-specific costs
## Model heating choice with alternative-specific cost coefficient
model_mlogit_costs <- heating_mlogit %>%
  mlogit(formula = depvar ~ oc | 1 | ic, data = ., reflevel = 'hp')
```

Alternative-Specific Cost Coefficients Model Summary

```
## Summarize model results
model_mlogit_costs %>%
  summary()
##
## Call:
## mlogit(formula = depvar ~ oc | 1 | ic, data = ., reflevel = "hp",
##     method = "nr")
##
## Frequencies of alternatives:
##       hp       ec       er       gc       gr
## 0.055556 0.071111 0.093333 0.636667 0.143333
##
## nr method
## 6 iterations, 0h:0m:0s
## g'(-H)^-1g = 8.85E-05
## successive function values within tolerance limits
##
## Coefficients :
##                    Estimate  Std. Error z-value Pr(>|z|)
## ec:(intercept)  1.70480715  1.27495400  1.3372 0.181173
## er:(intercept)  2.52929123  1.20296858  2.1025 0.035506 *
## gc:(intercept)  1.46257610  1.01069767  1.4471 0.147870
## gr:(intercept) -0.56099198  1.13846726 -0.4928 0.622182
## oc             -0.00545750  0.00182214 -2.9951 0.002743 **
## hp:ic          -0.00159724  0.00101723 -1.5702 0.116373
## ec:ic          -0.00214993  0.00122995 -1.7480 0.080468 .
## er:ic          -0.00265151  0.00093039 -2.8499 0.004374 **
## gc:ic          -0.00120808  0.00082802 -1.4590 0.144567
## gr:ic          -0.00055589  0.00083672 -0.6644 0.506453
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -1006.2
## McFadden R^2:  0.015635
```

Adamowicz et al. (1994)

Adamowicz et al. (1994)

Research question

- Do stated preference methods and revealed preference methods yield similar results on recreational site choice?

Empirical methods

- Conduct both a stated preference survey and a revealed preference survey on site choice for fishing and other water-based recreation
- Estimate a multinomial logit model for each model separately and a joint model including both surveys

Results

- Independent models appear to be different, but estimating a joint model that includes a scale parameter yields similar underlying preferences

Stated Preference Survey

Method used to value non-market goods

- Traditional application is valuing environmental amenities
- Popular research method in the 1980s–2000s but less popular now
- Any other areas where this could be used to elicit valuation?

Respondents state their choice over a hypothetical set of alternatives

- Survey methods have been refined over the years to remove bias that was present in early surveys
- Do we believe hypothetical decision making reveals true preferences?

Contrast with revealed preference methods

- Researchers document actual choices over a set of real alternatives
- May limit what parameters can be estimated

General Comments and Questions

Both stated choice and revealed choice methods yield data that seem perfect for a multinomial logit model

- But do we believe the assumptions of the logit model in this case?

Estimation of the joint model depends importantly on the scale parameter

- But they never say what scale parameter best fits the joint dataset

Authors apply their results to estimate the change in consumer surplus associated with a change in river flows

- Consumer surplus calculations require a marginal utility of income, which they do not directly estimate
- Welfare effects from the joint model are much closer to the revealed preference model than to the stated preference model
- What other counterfactuals or policy scenarios could we model with these results?

Recap and Looking Ahead

So far

- Discrete choice framework
- Random utility model
- Logit model

But we still do not know how to estimate the logit model!

Coming up

- Maximum likelihood estimation
- Numerical optimization
- Estimating the logit model

Announcements

Reading for next time

- Greene textbook, Chapters 7.1–7.2.1 and 14.1–14.6

Office hours

- Reminder: Tuesdays at 2:00-3:00 in 218 Stockbridge

Upcoming

- Problem Set 1 is posted, due September 24