Week 4: Logit Model

ResEcon 703: Topics in Advanced Econometrics

Matt Woerman
University of Massachusetts Amherst

Agenda

Last week

- Random utility model

This week's topics

- Logit model
- Logit choice probabilities
- Binary logit model
- Multinomial logit model
- Marginal effects and elasticities
- Logit substitution patterns
- Properties of logit parameters
- Counterfactuals and welfare
- Empirical considerations
- Binary logit model R example
- Multinomial logit model R example

This week's reading

- Train textbook, chapters 3.1–3.6

Logit Model

Random Utility Model Recap

A decision maker chooses the alternative that maximizes utility

- A decision maker, n, faces a choice among J discrete alternatives
- Alternative j provides utility U_{nj} (where $j = 1, \ldots, J$)
- n chooses i if and only if $U_{ni} > U_{nj} \; \forall j \neq i$

We (the econometricians) do not observe utility U_{nj}, so we model it as being composed of

- V_{nj}: Utility from observed attributes
- ε_{nj}: Utility from unobserved attributes, which we treat as random

$$U_{nj} = V_{nj} + \varepsilon_{nj}$$

The probability that decision maker n chooses alternative i is

$$\begin{aligned} P_{ni} &= \Pr(U_{ni} > U_{nj} \; \forall j \neq i) \\ &= \int_{\boldsymbol{\varepsilon}} \mathbb{1}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \; \forall j \neq i) f(\boldsymbol{\varepsilon}_n) d\boldsymbol{\varepsilon}_n \end{aligned}$$

Logit Model

The logit model makes a simple (but sometimes overly strong) assumption about the joint density of unobserved utility, $f(\varepsilon_n)$

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

Why make this assumption about the unobserved component of utilities?

- It yields a simple closed-form expression for choice probabilities

Are there any downsides to making this assumption?

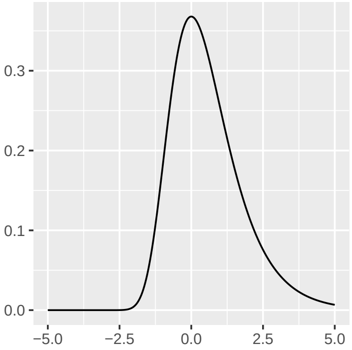
- It implies substitution patterns that may be unrealistic

Type I Extreme Value Density and Distribution

Type I extreme value is similar to a normal distribution but with a fatter tail on one side

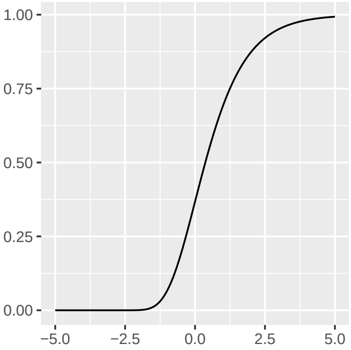
Probability density

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



Cumulative distribution

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

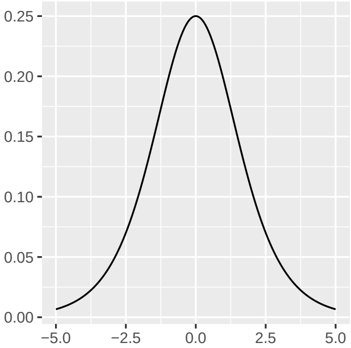


Logistic Density and Distribution

The difference of two type I extreme value draws, $\varepsilon^*_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$, follows the logistic distribution

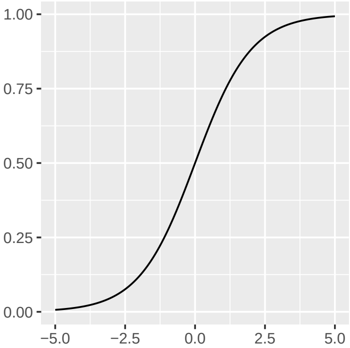
Probability density

$$f(\varepsilon^*_{nji}) = \frac{e^{\varepsilon^*_{nji}}}{\left(1 + e^{\varepsilon^*_{nji}}\right)^2}$$



Cumulative distribution

$$F(\varepsilon^*_{nji}) = \frac{e^{\varepsilon^*_{nji}}}{1 + e^{\varepsilon^*_{nji}}}$$



Logit Choice Probabilities

Logit Choice Probabilities

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

Suppose we know V_{ni}, V_{nj}, and ε_{ni}

- We know the right-hand side of the inequality inside the probability

For a single ε_{nj}, this probability is the cumulative distribution of a type I extreme value random variable

$$
\Pr(\varepsilon_{nj} < \varepsilon_{ni} + V_{ni} - V_{nj} \mid \varepsilon_{ni}) = e^{-e^{-(\varepsilon_{ni} + V_{ni} - V_{nj})}}
$$

We need to know this probability $\forall j \neq i$, not just a single j

- ε_{nj} is i.i.d., so we can take the product of the probability for each ε_{nj}

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

Logit Choice Probabilities

Conditional on knowing ε_{ni}, the choice probability of alternative i is

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

But ε_{ni} is random, so we have to integrate over the density of ε_{ni}

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

- See the textbook for the proof of this equivalence

The probability of n choosing i is a closed-form expression that depends on the representative utility (or observable attributes) of all alternatives

Properties of Logit Choice Probabilities

P_{ni} is always within the range $(0, 1)$

- $P_{ni} \to 1$ as $V_{ni} \to \infty$
- $P_{ni} \to 0$ as $V_{ni} \to -\infty$

Choice probabilities sum to 1

$$\sum_{i=1}^{J} P_{ni} = \sum_{i=1}^{J} \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}} = \frac{\sum_i e^{V_{ni}}}{\sum_j e^{V_{nj}}} = 1$$

Choice probability is a sigmoidal function of representative utility (see the logistic CDF)

- Marginal effects are small when probabilities are close to 0 or 1
- Marginal effects are largest when $P_{ni} = 0.5$

Linear Representative Utility

If we assume representative utility is linear in parameters

$$V_{nj} = \boldsymbol{\beta}' \mathbf{x}_{nj}$$

then the logit choice probability is

$$P_{ni} = \frac{e^{\boldsymbol{\beta}' \mathbf{x}_{ni}}}{\sum_j e^{\boldsymbol{\beta}' \mathbf{x}_{nj}}}$$

Reminders about parameters and estimation

- With linear representative utility, the structural parameters will typically give us the marginal utility of attributes, characteristics, etc.
  - We can use different models of representative utility to get different structural parameters
- We want to find the parameter values that make choice probabilities consistent with observed choices

Binary Logit Model

Binary Logit Choice Probabilities

Under the logit assumption, choice probabilities for a binary choice are

$$P_{n1} = \frac{e^{V_{n1}}}{e^{V_{n1}} + e^{V_{n2}}}$$
$$P_{n2} = \frac{e^{V_{n2}}}{e^{V_{n1}} + e^{V_{n2}}}$$

- There are several alternate ways to express these choice probabilities

A more succinct expression for the choice probability of alternative 1 is

$$P_{n1} = \frac{1}{1 + e^{-(V_{n1} - V_{n2})}}$$

If we assume representative utility is linear, $V_{nj} = \beta' \boldsymbol{x}_{nj}$

$$P_{n1} = \frac{1}{1 + e^{-\beta'(\boldsymbol{x}_{n1} - \boldsymbol{x}_{n2})}}$$

But this choice probability is nonlinear, so we cannot use OLS

Binary Logit Odds Ratio

We can instead calculate the odds ratio of alternative 1

$$\frac{P_{n1}}{1 - P_{n1}} = \frac{P_{n1}}{P_{n2}} = e^{V_{n1} - V_{n2}}$$

Then the log odds ratio of alternative 1 is

$$\ln\left(\frac{P_{n1}}{1 - P_{n1}}\right) = V_{n1} - V_{n2}$$

If we assume representative utility is linear, $V_{nj} = \beta' \boldsymbol{x}_{nj}$

$$\ln\left(\frac{P_{n1}}{1 - P_{n1}}\right) = \beta'(\boldsymbol{x}_{n1} - \boldsymbol{x}_{n2})$$

We might express this log odds ratio more generally as

$$\ln\left(\frac{P_{n1}}{1 - P_{n1}}\right) = \beta' \boldsymbol{x}_n$$

where $\boldsymbol{x}_n = \{\boldsymbol{x}_{n1}, \boldsymbol{x}_{n2}\}$ contains data about both alternatives

Binary Logit Estimation in R

The log odds ratio for the binary logit model is

$$\ln\left(\frac{P_{n1}}{1 - P_{n1}}\right) = \beta' \boldsymbol{x}_n$$

Now we have an expression with a linear right-hand side, but the left-hand side is nonlinear!

- This model is part of the family known as "generalized linear models"
- We can estimate this model in R using the `glm()` function with the argument `family = "binomial"`

This only works for a binary logit model because it is implicitly a single comparison that can be fully represented with one equation

- Estimation gets more complicated with more than two alternatives

Binary Logit Example

A person chooses whether to take a car (c) or a bus (b) to work

- We observe the time, T, and cost, M, of each alternative

We specify the representative utility of each alternative as

$$V_{nj} = \beta_{0j} + \beta_1 T_{nj} + \beta_2 M_{nj}$$

Under the logit assumption, the choice probability of driving is

$$\begin{aligned} P_{nc} &= \frac{e^{\beta_{0c} + \beta_1 T_{nc} + \beta_2 M_{nc}}}{e^{\beta_{0c} + \beta_1 T_{nc} + \beta_2 M_{nc}} + e^{\beta_{0b} + \beta_1 T_{nb} + \beta_2 M_{nb}}} \\ &= \frac{1}{1 + e^{-(\beta_{0c} - \beta_{ob}) - \beta_1 (T_{nc} - T_{ncb}) - \beta_2 (M_{nc} - M_{nb})}} \end{aligned}$$

The log odds ratio of driving is

$$\ln\left(\frac{P_{nc}}{1 - P_{nc}}\right) = (\beta_{0c} - \beta_{0b}) + \beta_1 (T_{nc} - T_{nb}) + \beta_2 (M_{nc} - M_{nb})$$

Multinomial Logit Model

Multinomial Logit Model

With a multinomial discrete choice (more than two alternatives), the problem becomes more complicated

- We cannot reduce the choice down to one choice probability

Under the logit assumption, the choice probabilities of the alternatives are

$$
\begin{aligned}
P_{n1} &= \frac{e^{V_{n1}}}{e^{V_{n1}} + e^{V_{n2}} + \cdots + e^{V_{nJ}}} \\
P_{n2} &= \frac{e^{V_{n2}}}{e^{V_{n1}} + e^{V_{n2}} + \cdots + e^{V_{nJ}}} \\
&\vdots \\
P_{nJ} &= \frac{e^{V_{nJ}}}{e^{V_{n1}} + e^{V_{n2}} + \cdots + e^{V_{nJ}}}
\end{aligned}
$$

Multinomial Logit Estimation in R

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

If we assume representative utility is linear, $V_{nj} = \beta' \boldsymbol{x}_{nj}$

$$P_{ni} = \frac{e^{\beta' \boldsymbol{x}_{ni}}}{\sum_j e^{\beta' \boldsymbol{x}_{nj}}}$$

We will use the `mlogit` package in R to estimate multinomial logit models

- Find the values of the structural parameters that make choice probabilities consistent with observed choices

The `mlogit` package requires us to

1. Organize the data to identify decision makers and alternatives
2. Specify the formula for representative utility

Marginal Effects and Elasticities

Marginal Effects

Unlike a linear probability model, the structural parameters of a logit model cannot be interpreted as marginal effects on probability

- But we can use the choice probabilities and parameters to derive the marginal effects!

The marginal effect of z_{ni}, an observed attribute of alternative i, on P_{ni} is

$$\frac{\partial P_{ni}}{\partial z_{ni}} = \frac{\partial \left(e^{V_{ni}} / \sum_j e^{V_{nj}} \right)}{\partial z_{ni}}$$
$$= \frac{\partial V_{ni}}{\partial z_{ni}} P_{ni}(1 - P_{ni})$$

If V_{ni} is linear in z_{ni} with coefficient β_z, then the marginal effect is

$$\frac{\partial P_{ni}}{\partial z_{ni}} = \beta_z P_{ni}(1 - P_{ni})$$

Cross Marginal Effects

The marginal effect of z_{nj}, an observed attribute of alternative j, on P_{ni} is

$$\frac{\partial P_{ni}}{\partial z_{nj}} = \frac{\partial \left(e^{V_{ni}} / \sum_k e^{V_{nk}} \right)}{\partial z_{nj}}$$
$$= -\frac{\partial V_{nj}}{\partial z_{nj}} P_{ni} P_{nj}$$

If V_{nj} is linear in z_{nj} with coefficient β_z, then the marginal effect is

$$\frac{\partial P_{ni}}{\partial z_{nj}} = -\beta_z P_{ni} P_{nj}$$

In a binary logit model, these marginal effect expressions are negatives of each another

Elasticities

Sometimes elasticities are more informative than marginal effects

- Percent changes rather than level changes

The elasticity of P_{ni} with respect to z_{ni}, an observed attribute of alternative i, is

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

If V_{ni} is linear in z_{ni} with coefficient β_z, then the elasticity is

$$E_{iz_{ni}} = \beta_z z_{ni}(1 - P_{ni})$$

Cross Elasticities

The elasticity of P_{ni} with respect to z_{nj}, an observed attribute of alternative j, is

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

If V_{nj} is linear in z_{nj} with coefficient β_z, then the elasticity is

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

The cross-elasticity of P_{ni} with respect to z_{nj} depends only on attributes of alternative j and not on alternative i

Logit Substitution Patterns

Independence of Irrelevant Alternatives

The ratio of any two logit choice probabilities is

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

This ratio only depends on attributes of alternatives i and k, so the relative probability of choosing i over k 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

But there is one major downside to the IIA property

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 attributes 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 the IIA property, 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

The cross-elasticity of P_{ni} with respect to z_{nj} is given by (assuming linearity of representative utility)

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

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

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

Properties of Logit Parameters

Variation in Preferences

A decision maker's preferences can vary for many reasons, some of which are observable, but others are not

- The logit model can only explicitly capture variation due to observable attributes
- Future models will allow for unobservable variation

Consider some sources of preference variation in the car-or-bus commute choice

- Some people hate driving and some people love it, but we do not directly observe this preference
  - We cannot include this variation in the logit model
- People with higher incomes care less about the cost of each alternative

$$\beta_n = \frac{\beta}{I_n} \quad \Rightarrow \quad U_{nc} = \alpha T_{nc} + \beta \frac{M_{nc}}{I_n} + \varepsilon_{nc}$$

The logit model allows for parameters to be a function of observable data

Scale Parameter

In the logit model, we assume the unobserved and random component of utility has variance $\pi^2/6$

- This assumption may seem restrictive, but we can use a scale parameter, σ, to allow for a different variance

Suppose the random utility, ε_{nj}^*, actually has variance $\sigma^2 \times (\pi^2/6)$

$$U_{nj}^* = V_{nj} + \varepsilon_{nj}^*$$

Dividing by σ gives a scaled model

$$U_{nj} = \frac{V_{nj}}{\sigma} + \varepsilon_{nj} \text{ where } \varepsilon_{nj} = \frac{\varepsilon_{nj}^*}{\sigma}$$

The variance of the scaled random utility is

$$Var(\varepsilon_{nj}) = \frac{1}{\sigma^2} Var(\varepsilon_{nj}^*) = \frac{\pi^2}{6}$$

Logit Choice Probabilities with a Scale Parameter

In the scaled model, choice probabilities are

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

If V_{nj} is linear in parameters with coefficients $\boldsymbol{\beta}^*$

$$P_{ni} = \frac{e^{(\boldsymbol{\beta}^*/\sigma)'\mathbf{x}_{ni}}}{\sum_j e^{(\boldsymbol{\beta}^*/\sigma)'\mathbf{x}_{nj}}}$$

But $\boldsymbol{\beta}^*$ and σ are not separately identified, so we can only estimate their ratio, $\boldsymbol{\beta} = \boldsymbol{\beta}^*/\sigma$, which gives the standard logit expression

$$P_{ni} = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_{ni}}}{\sum_j e^{\boldsymbol{\beta}'\mathbf{x}_{nj}}}$$

Parameters are estimated relative to the variance of unobserved utility

Heteroskedasticity and the Scale Parameter

Different subsets of decision makers may each have a different variance of random utility

- We can use scale parameters to account for this groupwise heteroskedasticity
- We can estimate the relative scale parameters of each group compared to one reference group

Suppose we have commute data for both Amherst (A) and Boston (B)

- The scale parameters for each city are σ^A and σ^B with $k = (\sigma^B/\sigma^A)^2$

$$\text{Amherst:} \quad P_{ni} = \frac{e^{\beta' \mathbf{x}_{ni}}}{\sum_j e^{\beta' \mathbf{x}_{nj}}}$$

$$\text{Boston:} \quad P_{ni} = \frac{e^{(\beta/\sqrt{k})' \mathbf{x}_{ni}}}{\sum_j e^{(\beta/\sqrt{k})' \mathbf{x}_{nj}}}$$

Counterfactuals and Welfare

Counterfactual Simulations

An advantage of a structural econometric model is the ability to conduct counterfactual simulations and calculate their welfare consequences

You can compare outcomes in the observed empirical setting to outcomes in an alternate setting with some aspect manipulated

- Different attributes or data
- Different choice set
- Different structural parameters

Examples of counterfactual simulations

- Estimate the demand for education in order to simulate the effects of a school voucher program
- Estimate how farmers choose which crop to plant in order to simulate the effects of a groundwater sustainability policy
- Estimate the supply of labor in order to simulate the effects of an income tax change

Simulating Individual Choices in the Logit Model

We cannot simulate discrete choices with certainty, but we can use choice probabilities to simulate choices in expectation

$$E(Y_{ni}) = P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}$$

where $Y_{ni} = 1$ if and only if n chooses i

The change in the expectation of a decision maker's choice due to a change in the choice setting is

$$\Delta E(Y_{ni}) = P^1_{ni} - P^0_{ni} = \frac{e^{V^1_{ni}}}{\sum_{j=1}^{J^1} e^{V^1_{nj}}} - \frac{e^{V^0_{ni}}}{\sum_{j=1}^{J^0} e^{V^0_{nj}}}$$

where the 1 superscript denotes the counterfactual and the 0 superscript denotes the observed empirical setting

- Note: You must "simulate" choices in the original setting!

Simulating Aggregate Choices in the Logit Model

We can simulate the aggregate number of decision makers expected to choose alternative i, which we denote A_i, by summing the individual expectation over all agents

$$E(A_i) = \sum_{n=1}^{N} E(Y_{ni}) = \sum_{n=1}^{N} P_{ni}$$

The expected change in the aggregate number of decision makers choosing alternative i due to a change in the choice setting is

$$\Delta E(A_i) = \sum_{n=1}^{N} P_{ni}^1 - \sum_{n=1}^{N} P_{ni}^0 = \sum_{n=1}^{N} \frac{e^{V_{ni}^1}}{\sum_{j=1}^{J^1} e^{V_{nj}^1}} - \sum_{n=1}^{N} \frac{e^{V_{ni}^0}}{\sum_{j=1}^{J^0} e^{V_{nj}^0}}$$

- Reminder: You must "simulate" choices in the original setting!

Consumer Surplus

The logit model gives a closed-form expression for consumer surplus

- Monetary gain to a consumer from "purchasing" a good for less than the value the consumer places on the good

If we know the marginal utility of income for decision maker n, which we denote α_n, then the consumer surplus from a choice setting is

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

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

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

Consumer Surplus in Counterfactuals

The expected consumer surplus that decision maker n obtains when faced with a choice setting is

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

The expected change in consumer surplus due to a change in the choice setting is

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

- Reminder: You must "simulate" consumer surplus in the original setting!

Empirical Considerations

Market-Level Data

The logit model can be (and often is) estimated from market-level data

- You observe the price, market share, and attributes 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, $V_j = \beta' \boldsymbol{x}_j$

$$\ln(S_i) - \ln(S_j) = \beta'(\boldsymbol{x}_i - \boldsymbol{x}_j)$$

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

$$\ln(S_i) - \ln(S_0) = \beta'(\boldsymbol{x}_i - \boldsymbol{x}_0) + \omega_i$$

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 include lagged or future variables to capture "dynamics"
- We can 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 preferences of a decision maker that affect their choices are unlikely to be independent over time

Exogeneity

This entire discussion of the logit model relies on the exogeneity of the data (attributes of alternatives, etc.)

$$E(\varepsilon_n \mid \boldsymbol{x}_n) = \mathbf{0}$$

- If the data are endogenous, then our structural parameter estimates may be biased

Example of endoegenity in the car-or-bus commute choice

- If a commuter likes to drive, they will not care about living close to a bus stop
- If a commuter likes to take the bus, they are more likely to live close to a bus stop

We will talk about how to deal with endogeneity later in the course

Binary Logit Model R Example

Binary Choice Example

We are studying how consumers make choices about expensive and highly energy-consuming appliances in their homes.

- We have (simulated) data on 600 households that rent apartments without air conditioning. These households must choose whether or not to purchase a window air conditioning unit. (To simplify things, we assume there is only one "representative" air conditioner for each household and its price and operating cost are exogenous.)
- We observe the following data about each household and its "representative" air conditioner
  - An indicator if they purchase the air conditioner (TRUE/FALSE)
  - The purchase price of the air conditioner ($)
  - The annual operating cost of the air conditioner ($ per year)
  - The household's electricity price (cents per kWh)
  - The size of the household's apartment (square feet)
  - The household's annual income ($1000s)
  - The number of residents in the household (people)
  - An indicator for the household's city (1, 2, or 3)

Random Utility Model for Air Conditioner Choice

We model the utility to household n of not purchasing an air conditioned ($j = 0$) or purchasing an air conditioner ($j = 1$) as

$$U_{n0} = V_{n0} + \varepsilon_{n0}$$
$$U_{n1} = V_{n1} + \varepsilon_{n1}$$

where V_{nj} depends on the data about alternative j and household n

The probability that household n purchases an air conditioner is

$$P_{n1} = \Pr(\varepsilon_{n0} - \varepsilon_{n1} < V_{n1} - V_{n0})$$

- Only differences in utility—not the actual values of utility—affect this probability
- What is the difference in utility to household n from purchasing an air conditioner vs. not purchasing an air conditioner?

Representative Utility for Air Conditioner Choice

$$P_{n1} = \Pr(\varepsilon_{n0} - \varepsilon_{n1} < V_{n1} - V_{n0})$$

What is the difference in utility to household n from purchasing an air conditioner vs. not purchasing an air conditioner?

- They gain utility from having air conditioning
- They lose utility from paying the purchase price of the air conditioner
- They lose utility from paying the annual operating cost of the air conditioner

We can model the difference in utility as

$$V_{n1} - V_{n0} = \beta_0 + \beta_1 P_n + \beta_2 C_n$$

where

- P_n is the purchase price of the air conditioner
- C_n is the annual operating cost of the air conditioner
- β_0, β_1, and β_2 are utility parameters to be estimated

Binary Logit Model of Air Conditioner Choice

Under the logit assumption, the choice probability of purchasing air conditioning becomes

$$
\begin{aligned}
P_{n1} &= \frac{1}{1 + e^{-(V_{n1} - V_{n0})}} \\
&= \frac{1}{1 + e^{-(\beta_0 + \beta_1 P_n + \beta_2 C_n)}}
\end{aligned}
$$

Alternatively, the log odds ratio of purchasing air conditioning is

$$
\begin{aligned}
\ln\left(\frac{P_{n1}}{1 - P_{n1}}\right) &= V_{n1} - V_{n0} \\
&= \beta_0 + \beta_1 P_n + \beta_2 C_n
\end{aligned}
$$

We can estimate this "generalized linear model" in R using the `glm()` function with the argument `family = "binomial"`

Load Dataset

`read_csv()` is a `tidyverse` function to read a .csv file into a tibble

```
## Load tidyverse
library(tidyverse)
## Load dataset
ac_data <- read_csv('ac_renters.csv')
```

Dataset

```
## Look at dataset
ac_data
## # A tibble: 600 x 8
##    air_conditio~1 cost_~2 cost_~3 elec_~4 squar~5 income resid~6  city
##    <lgl>            <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl> <dbl>
##  1 FALSE              513     247    12.8     541     47       2     1
##  2 FALSE              578     138     9.6     384     64       1     1
##  3 TRUE               658     171    10.7     619     86       2     1
##  4 FALSE              615     198    11.5     624     49       2     1
##  5 FALSE              515     165    10.5     365     56       1     1
##  6 FALSE              588     143     9.7     411     39       2     1
##  7 TRUE               643     153    10.1     529     58       2     1
##  8 FALSE              676     182    11       694     46       2     1
##  9 TRUE               516     137     9.6     305     75       1     1
## 10 TRUE               544     185    11.1     454     68       3     1
## # ... with 590 more rows, and abbreviated variable names
## #   1: air_conditioning, 2: cost_system, 3: cost_operating,
## #   4: elec_price, 5: square_feet, 6: residents
```

Generalized Linear Model

We want to estimate the generalized linear model

$$\ln\left(\frac{P_{n1}}{1-P_{n1}}\right) = \beta_0 + \beta_1 P_n + \beta_2 C_n$$

`glm()` is the R function to fit a generalized linear model

- The argument `family = "binomial"` indicates the "link" between the nonlinear left-hand side and the linear right-hand side

```
## Model air conditioning as a function of cost variables
binary_logit <- glm(formula =
                      air_conditioning ~ cost_system + cost_operating,
                    family = 'binomial',
                    data = ac_data)
```

Model Summary

summary() summarizes the results of the model

```
## Summarize model results
summary(binary_logit)
##
## Call:
## glm(formula = air_conditioning ~ cost_system + cost_operating,
##     family = "binomial", data = ac_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9130  -1.1849   0.7523   0.9929   1.8579
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     4.436647   0.953225   4.654 3.25e-06 ***
## cost_system    -0.002974   0.001504  -1.978    0.048 *
## cost_operating -0.015428   0.002355  -6.552 5.67e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 820.53  on 599  degrees of freedom
## Residual deviance: 764.97  on 597  degrees of freedom
## AIC: 770.97
##
## Number of Fisher Scoring iterations: 4
```

Interpreting Coefficients

`coef()` is the R function to display only the model coefficients

```
## Display model coefficients
coef(binary_logit)
##    (Intercept)    cost_system cost_operating
##    4.436646631   -0.002974325   -0.015427739
```

How do we interpret these coefficients?

- Air conditioning generates 4.44 “utils” of utility
- An additional $100 of purchase price reduces utility by 0.30
- An additional $100 of annual operating cost reduces utility by 1.54

Fitted Utilities

predict() calculates the fitted values of the model

```
## Calculate utility of air conditioning
ac_data <- ac_data %>%
  mutate(utility_ac_logit = predict(binary_logit))
## Look at utilities and other data
ac_data %>%
  select(air_conditioning, starts_with('cost'), utility_ac_logit)
## # A tibble: 600 x 4
##    air_conditioning cost_system cost_operating utility_ac_logit
##    <lgl>                  <dbl>          <dbl>            <dbl>
##  1 FALSE                    513            247          -0.900
##  2 FALSE                    578            138           0.588
##  3 TRUE                     658            171          -0.159
##  4 FALSE                    615            198          -0.447
##  5 FALSE                    515            165           0.359
##  6 FALSE                    588            143           0.482
##  7 TRUE                     643            153           0.164
##  8 FALSE                    676            182          -0.382
##  9 TRUE                     516            137           0.788
## 10 TRUE                     544            185          -0.0355
## # ... with 590 more rows
```

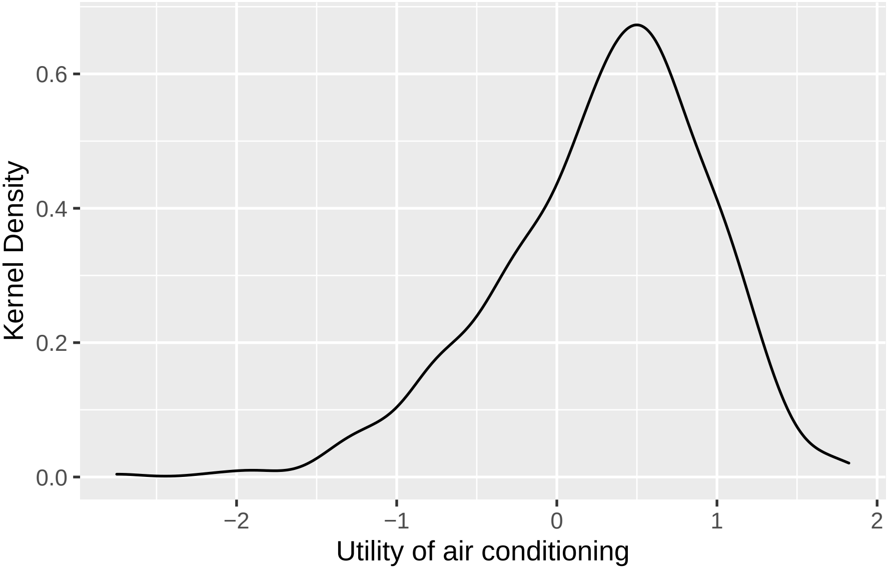
Kernel Density of Fitted Utilities

ggplot is a highly flexible and powerful system for creating visualizations in R

- Data visualization is beyond the scope of this course, and many good ggplot tutorials and references exist

```
## Plot density of utilities
ac_data %>%
  ggplot(aes(x = utility_ac_logit)) +
  geom_density() +
  xlab('Utility of air conditioning') +
  ylab('Kernel Density')
```

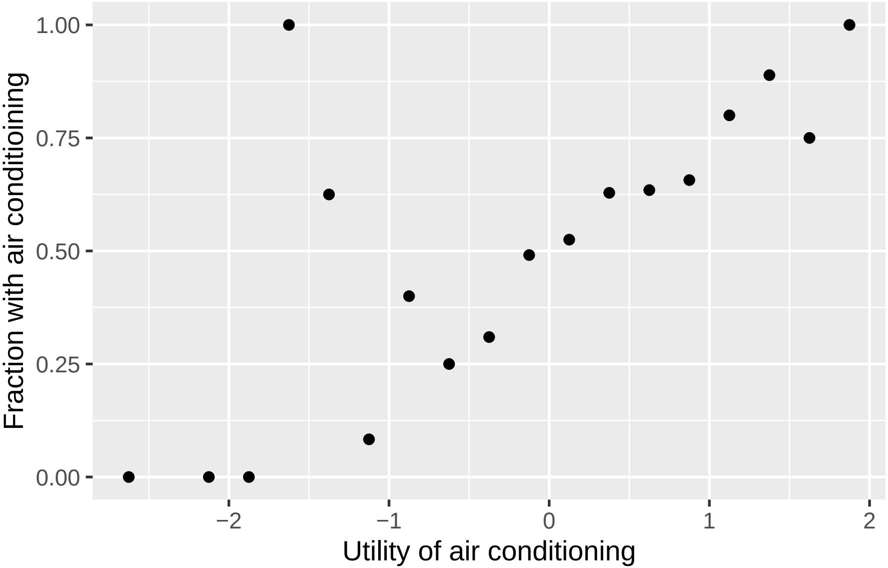
Kernel Density of Fitted Utilities



Plot of Utility vs. Adoption

```
## Plot fraction vs. utility of air conditioning using bins
ac_data %>%
  mutate(bin = cut(utility_ac_logit,
                   breaks = seq(-3, 2, 0.25),
                   labels = 1:20)) %>%
  group_by(bin) %>%
  summarize(fraction_ac = mean(air_conditioning), .groups = 'drop') %>%
  mutate(bin = as.numeric(bin),
         bin_mid = 0.25 * (bin - 1) - 2.875) %>%
  ggplot(aes(x = bin_mid, y = fraction_ac)) +
  geom_point() +
  xlab('Utility of air conditioning') +
  ylab('Fraction with air conditioining')
```

Plot of Utility vs. Adoption



Choice Probabilities

We can use the fitted utility values to calculate each household's choice probability of adopting air conditioning

$$P_{n1} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 P_n + \beta_2 C_n)}}$$

```
## Calculate choice probability of air conditioning
ac_data <- ac_data %>%
  mutate(probability_ac_logit = 1 / (1 + exp(-utility_ac_logit)))
```

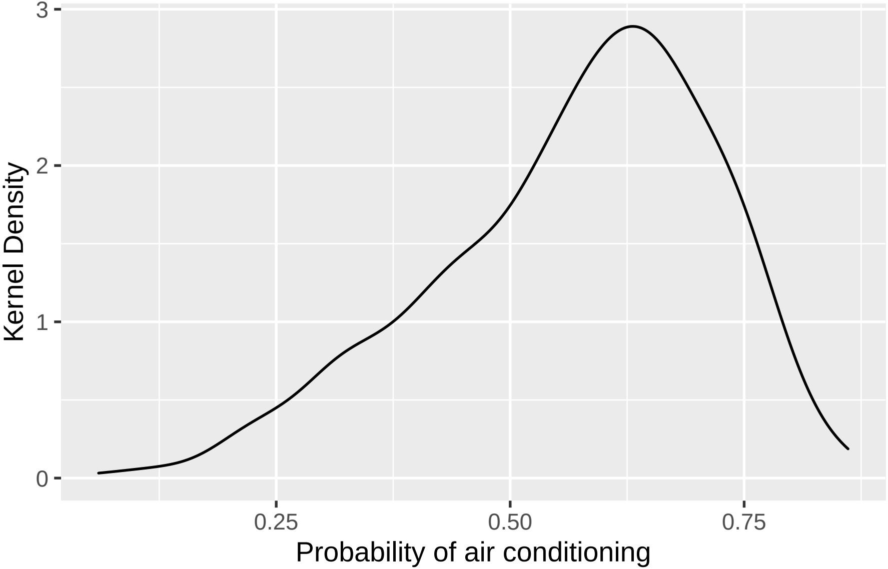
Choice Probabilities

```
## Look at utilities and probabilities
ac_data %>%
  select(air_conditioning, utility_ac_logit, probability_ac_logit)
## # A tibble: 600 x 3
##    air_conditioning utility_ac_logit probability_ac_logit
##    <lgl>                       <dbl>                <dbl>
##  1 FALSE                     -0.900                 0.289
##  2 FALSE                      0.588                 0.643
##  3 TRUE                      -0.159                 0.460
##  4 FALSE                     -0.447                 0.390
##  5 FALSE                      0.359                 0.589
##  6 FALSE                      0.482                 0.618
##  7 TRUE                       0.164                 0.541
##  8 FALSE                     -0.382                 0.406
##  9 TRUE                       0.788                 0.687
## 10 TRUE                      -0.0355                0.491
## # ... with 590 more rows
```

Kernel Density of Choice Probabilities

```r
## Plot density of probabilities
ac_data %>%
  ggplot(aes(x = probability_ac_logit)) +
  geom_density() +
  xlab('Probability of air conditioning') +
  ylab('Kernel Density')
```

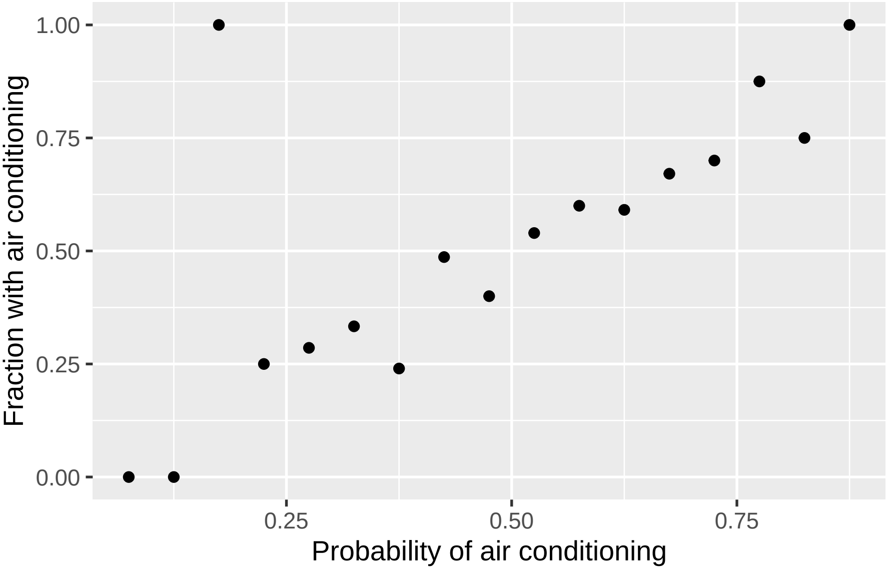
Kernel Density of Choice Probabilities



Plot of Probability vs. Adoption

```
## Plot fraction vs. probability of air conditioning using bins
ac_data %>%
  mutate(bin = cut(probability_ac_logit,
                   breaks = seq(0, 1, 0.05),
                   labels = 1:20)) %>%
  group_by(bin) %>%
  summarize(fraction_ac = mean(air_conditioning), .groups = 'drop') %>%
  mutate(bin = as.numeric(bin),
         bin_mid = 0.05 * (bin - 1) + 0.025) %>%
  ggplot(aes(x = bin_mid, y = fraction_ac)) +
  geom_point() +
  xlab('Probability of air conditioning') +
  ylab('Fraction with air conditioning')
```

Plot of Probability vs. Adoption



Marginal Effects

We can calculate the marginal effects of each cost variable

$$\frac{\partial P_{ni}}{\partial z_{ni}} = \beta_z P_{ni}(1 - P_{ni})$$

```
## Calculate the marginal effect of each cost variable
ac_data <- ac_data %>%
  mutate(marg_eff_system = coef(binary_logit)[2] *
           probability_ac_logit * (1 - probability_ac_logit),
         marg_eff_operating = coef(binary_logit)[3] *
           probability_ac_logit * (1 - probability_ac_logit))
```

Elasticities

We can calculate the elasticities of each cost variable

$$E_{iz_{ni}} = \beta_z z_{ni}(1 - P_{ni})$$

```
## Calculate the elasticity of each cost variable
ac_data <- ac_data %>%
  mutate(elasticity_system = coef(binary_logit)[2] *
           cost_system * (1 - probability_ac_logit),
         elasticity_operating = coef(binary_logit)[3] *
           cost_operating * (1 - probability_ac_logit))
```

Heterogeneous Marginal Effects and Elasticities

These marginal effects and elasticities are heterogeneous because households have different costs and choice probabilities

```
## Look at marginal effects and elasticities
ac_data %>%
  select(starts_with('marg_eff'), starts_with('elasticity'))
## # A tibble: 600 x 4
##    marg_eff_system marg_eff_operating elasticity_system elasticity_o~1
##              <dbl>              <dbl>             <dbl>          <dbl>
##  1       -0.000611           -0.00317            -1.08          -2.71
##  2       -0.000683           -0.00354            -0.614         -0.760
##  3       -0.000739           -0.00383            -1.06          -1.42
##  4       -0.000708           -0.00367            -1.12          -1.86
##  5       -0.000720           -0.00374            -0.630         -1.05
##  6       -0.000702           -0.00364            -0.668         -0.842
##  7       -0.000739           -0.00383            -0.878         -1.08
##  8       -0.000717           -0.00372            -1.19          -1.67
##  9       -0.000639           -0.00331            -0.480         -0.661
## 10       -0.000743           -0.00386            -0.823         -1.45
## # ... with 590 more rows, and abbreviated variable name
## #   1: elasticity_operating
```

Summary of Marginal Effects and Elasticities

`summary()` also summarizes the variables of a data frame or tibble

```
## Summarize marginal effects and elasticities
ac_data %>%
  select(starts_with('marg_eff'), starts_with('elasticity')) %>%
  summary()
##  marg_eff_system      marg_eff_operating   elasticity_system
##  Min.   :-0.0007436   Min.   :-0.0038569   Min.   :-1.7444
##  1st Qu.:-0.0007282   1st Qu.:-0.0037774   1st Qu.:-0.9452
##  Median :-0.0006923   Median :-0.0035911   Median :-0.7295
##  Mean   :-0.0006642   Mean   :-0.0034454   Mean   :-0.7719
##  3rd Qu.:-0.0006237   3rd Qu.:-0.0032352   3rd Qu.:-0.5510
##  Max.   :-0.0001682   Max.   :-0.0008722   Max.   :-0.2180
##  elasticity_operating
##  Min.   :-5.1039
##  1st Qu.:-1.4375
##  Median :-0.9176
##  Mean   :-1.1154
##  3rd Qu.:-0.6317
##  Max.   :-0.1415
```

Binary Logit with Heterogeneous Parameters

We have estimated a single "average" parameter for each variable

- But in reality, marginal utility is likely to vary by income

$$V_{n1} - V_{n0} = \beta_0 + \beta_{1n} P_n + \beta_{2n} C_n + \varepsilon_n$$

$$\beta_{1n} = \frac{\beta_1}{I_n} \quad \text{and} \quad \beta_{2n} = \frac{\beta_2}{I_n}$$

Estimate a model using price or cost as a share of income

$$\ln\left(\frac{P_{n1}}{1 - P_{n1}}\right) = \beta_0 + \beta_1 \frac{P_n}{I_n} + \beta_2 \frac{C_n}{I_n}$$

Use `I()` around math inside your R formula

```r
## Model air conditioning as a function of costs divided by income
binary_logit_inc <- glm(formula =
                          air_conditioning ~ I(cost_system / income) +
                          I(cost_operating / income),
                        family = 'binomial',
                        data = ac_data)
```

Binary Logit with Heterogeneous Parameters

```
## Summarize model results
summary(binary_logit_inc)
##
## Call:
## glm(formula = air_conditioning ~ I(cost_system/income) + I(cost_operating/income),
##     family = "binomial", data = ac_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9101  -0.7204   0.3585   0.6704   2.2796
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                7.2768     0.6318  11.518  < 2e-16 ***
## I(cost_system/income)     -0.3936     0.0645  -6.103 1.04e-09 ***
## I(cost_operating/income)  -1.1010     0.1653  -6.661 2.71e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 820.53  on 599  degrees of freedom
## Residual deviance: 539.46  on 597  degrees of freedom
## AIC: 545.46
##
## Number of Fisher Scoring iterations: 5
```

Interpreting Heterogeneous Parameters

```r
## Display model coefficients
coef(binary_logit_inc)
##             (Intercept)    I(cost_system/income)
##               7.2768120               -0.3935995
## I(cost_operating/income)
##              -1.1010374
```

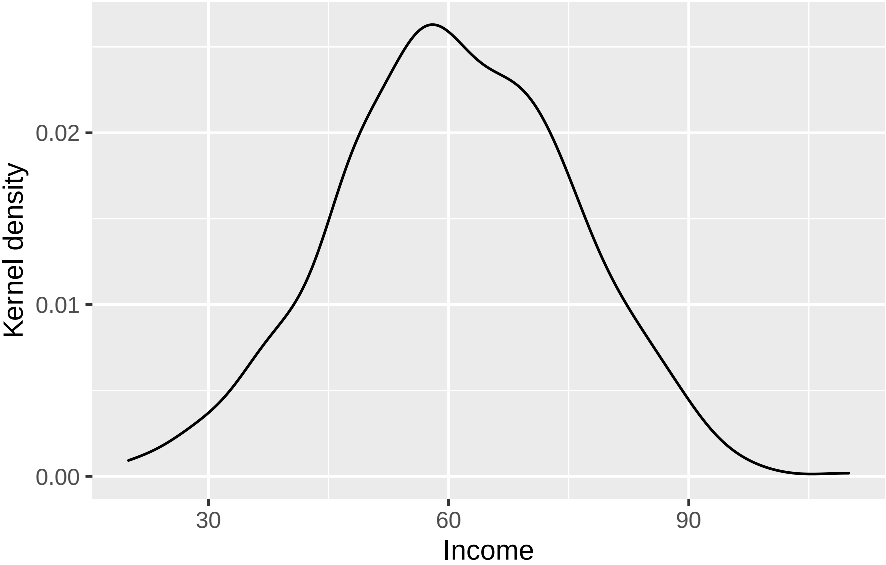
How do we interpret these parameters?

- Air conditioning generates 7.28 “utils” of utility
- An additional 0.1 percentage point of purchase price as a share of income reduces utility by 0.39
- An additional 0.1 percentage point of annual operating cost as a share of income reduces utility by 1.10

Kernel Density of Income

```
## Plot kernel density of income
ac_data %>%
  ggplot(aes(x = income)) +
  geom_density() +
  xlab('Income') +
  ylab('Kernel density')
```

Kernel Density of Income



Marginal Utility Depending on Income

What are the marginal utilities at $30,000 income? $60,000? $90,000?

$$\beta_{1n} = \frac{\beta_1}{I_n} \quad \text{and} \quad \beta_{2n} = \frac{\beta_2}{I_n}$$

```
## Calculate marginal utility of costs when income == 30
coef(binary_logit_inc)[2:3] / 30
##    I(cost_system/income) I(cost_operating/income)
##              -0.01311998              -0.03670125

## Calculate marginal utility of costs when income == 60
coef(binary_logit_inc)[2:3] / 60
##    I(cost_system/income) I(cost_operating/income)
##             -0.006559992             -0.018350623

## Calculate marginal utility of costs when income == 90
coef(binary_logit_inc)[2:3] / 90
##    I(cost_system/income) I(cost_operating/income)
##             -0.004373328             -0.012233749
```

Cost Trade-Offs

We can use the structural parameters of this model to determine how consumers trade off the purchase price and the annual operating cost

- If the annual operating cost were to increase by $1, what reduction in the purchase price would leave consumers no worse off?

$$
\begin{aligned}
U_{n1} &= \beta_0 + \beta_1 P_n + \beta_2 C_n + \varepsilon_{n1} \\
dU_{n1} &= \beta_1 dP_n + \beta_2 dC_n \\
dU_{n1} = 0 &\quad \Rightarrow \quad \frac{dP_n}{dC_n} = -\frac{\beta_2}{\beta_1}
\end{aligned}
$$

```
## Calculate system cost equivalence of an increase in operating cost
-coef(binary_logit)[3] / coef(binary_logit)[2]
## cost_operating
##      -5.186972
```

Implied Discount Rate

We can also use these structural parameters to determine what discount rate is implied by air conditioner purchase decisions

- How the future is valued or "discounted" compared to today

If we assume an infinite time horizon for the annual operating cost, a general formula for a household's expected utility after purchasing an air conditioner is

$$U_{n1} = \alpha_0 + \alpha_1 \left(P_n + \frac{1}{\gamma} C_n \right) + \omega_n$$

where α_1 is the marginal utility of income and γ is the discount rate

From our model, the utility from purchasing an air conditioner is

$$U_{n1} = \beta_0 + \beta_1 P_n + \beta_2 C_n + \varepsilon_{n1}$$

We have not estimated γ in the binary logit model, but we can use our structural parameters to calculate it

Implied Discount Rate Calculation

$$U_{n1} = \alpha_0 + \alpha_1 \left(P_n + \frac{1}{\gamma} C_n \right) + \omega_n$$
$$U_{n1} = \beta_0 + \beta_1 P_n + \beta_2 C_n + \varepsilon_{n1}$$

These two expressions for the equivalent utility imply that

$$\alpha_1 = \beta_1$$
$$\frac{\alpha_1}{\gamma} = \beta_2$$

which we can combine and rewrite as

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

```
## Calculate the implied discount rate
coef(binary_logit)[2] / coef(binary_logit)[3]
## cost_system
##   0.1927907
```

Multinomial Logit Model R Example

Multinomial Choice Example

We are studying how consumers make choices about expensive and highly energy-consuming appliances in their homes, but now with different data

- We have (real) 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

Random Utility Model of Heating System Choice

We model the utility to household n of installing heating system j as

$$U_{nj} = V_{nj} + \varepsilon_{nj}$$

where V_{nj} depends on the data about alternative j and household n

The probability that household n installs heating system i is

$$P_{ni} = \int_{\boldsymbol{\varepsilon}} \mathbb{1}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \; \forall j \neq i) f(\boldsymbol{\varepsilon}_n) d\boldsymbol{\varepsilon}_n$$

Under the logit assumption, these choice probabilities simplify to

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

Representative Utility of Heating System Choice

What might affect the utility of the different heating systems?

- Installation cost
- Annual operating cost
- Heating system technology
  - Gas systems might be preferred to electric systems
  - Central systems might be preferred to room systems
- Anything else?

We model the representative utility of heating system j to household n as

$$V_{nj} = \alpha_j + \beta_1 IC_{nj} + \beta_2 OC_{nj}$$

Multinomial Logit Model of Heating System Choice

Under the logit assumption, the choice probability that household n installs heating system i is

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

We model the representative utility of heating system j to household n as

$$V_{nj} = \alpha_j + \beta_1 IC_{nj} + \beta_2 OC_{nj}$$

Substituting this representative utility into the choice probabilities gives

$$P_{ni} = \frac{e^{\alpha_i + \beta_1 IC_{ni} + \beta_2 OC_{ni}}}{\sum_j e^{\alpha_j + \beta_1 IC_{nj} + \beta_2 OC_{nj}}}$$

We can use the `mlogit` package in R to estimate the six parameters that make these choice probabilities consistent with observed choices

Load Dataset

The `Heating` dataset is part of the `mlogit` package, so we can load it using the `data()` function

```
## Load tidyverse and mlogit
library(tidyverse)
library(mlogit)
## Load dataset from mlogit package
data('Heating', package = 'mlogit')
## Rename dataset to lowercase
heating <- Heating
```

Dataset

```
## Look at dataset
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>
```

This dataset is in a wide format

Convert to a Long Dataset

We can instead represent the exact same data in a long format

```
## Pivot into a long dataset
heating_long <- heating %>%
  pivot_longer(contains('.')) %>%
  separate(name, c('name', 'alt')) %>%
  pivot_wider() %>%
  mutate(choice = (depvar == alt)) %>%
  select(-depvar)
```

Long Dataset

```
## Look at long dataset
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 gc      866  200. TRUE
##  2      1      7     25     6 ncostl gr      963.  152. FALSE
##  3      1      7     25     6 ncostl ec      860.  553. FALSE
##  4      1      7     25     6 ncostl er      996.  506. FALSE
##  5      1      7     25     6 ncostl hp     1136.  238. FALSE
##  6      2      5     60     5 scostl gc      728.  169. TRUE
##  7      2      5     60     5 scostl gr      759.  169. FALSE
##  8      2      5     60     5 scostl ec      797.  520. FALSE
##  9      2      5     60     5 scostl er      895.  486. FALSE
## 10      2      5     60     5 scostl hp      969.  199. FALSE
## # ... with 4,490 more rows
```

This dataset is in a long format

Format Datasets for `mlogit` Package

The first step to use the `mlogit` package in R is to convert the data frame to an indexed data frame

- The indexing adds additional "structure" to the data frame to define the choice setting (decision maker, household, etc.) and the alternatives

The `dfidx()` function—from the `dfidx` package, which is loaded automatically when you load the `mlogit` package—converts a data frame to an indexed data frame

- Type `?dfidx` for the help file
- See the vignettes on the `dfidx` and `mlogit` CRAN pages for more information
  - `cran.r-project.org/web/packages/dfidx/index.html`
  - `cran.r-project.org/web/packages/mlogit/index.html`

Using the dfidx() Function to Convert Datasets

There are many different ways to specify the "structure" of a data frame in the dfidx() function, but these arguments work in many cases:

- data: data frame you wanted to be converted
- shape: 'wide' or 'long' for the format of the data frame
- choice: variable that contains the choice indicator
- The fourth argument depends on the format of the data frame
  - For a wide data frame, varying: numeric vector defining which variables contain alternative-specific data
  - For a long data frame, idx: two-element character vector defining which which variables contain identifiers for the choice situation and alternative

```
## Convert wide data to dfidx format
heating_dfidx <- dfidx(heating, shape = 'wide',
                       choice = 'depvar', varying = 3:12)
## Convert long data to dfidx format
heating_long_dfidx <- dfidx(heating_long, shape = 'long',
                            choice = 'choice', idx = c('idcase', 'alt'))
```

Wide Dataset in dfidx Format

```r
## Look at wide data in dfidx format
tibble(heating_dfidx)
## # A tibble: 4,500 x 9
##    idcase depvar income agehed rooms region    ic    oc idx$id1 $id2
##     <dbl> <lgl>   <dbl>  <dbl> <dbl> <fct>  <dbl> <dbl>   <int> <fct>
##  1      1 FALSE       7     25     6 ncostl  860.  553.       1 ec
##  2      1 FALSE       7     25     6 ncostl  996.  506.       1 er
##  3      1 TRUE        7     25     6 ncostl  866   200.       1 gc
##  4      1 FALSE       7     25     6 ncostl  963.  152.       1 gr
##  5      1 FALSE       7     25     6 ncostl 1136.  238.       1 hp
##  6      2 FALSE       5     60     5 scostl  797.  520.       2 ec
##  7      2 FALSE       5     60     5 scostl  895.  486.       2 er
##  8      2 TRUE        5     60     5 scostl  728.  169.       2 gc
##  9      2 FALSE       5     60     5 scostl  759.  169.       2 gr
## 10      2 FALSE       5     60     5 scostl  969.  199.       2 hp
## # ... with 4,490 more rows
```

Long Dataset in dfidx Format

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

Estimate Multinomial Logit Model in `mlogit` Package

The second step to use the `mlogit` package in R is to specify a `formula` for representative utility

- This `formula` is more flexible—and more complicated—than we have used with `lm()` or `glm()`

The `mlogit()` function takes a `formula` with four separate sets of covariates to allow for four different kinds of parameters

$$\texttt{mlogit}(\texttt{formula} = \texttt{y} \sim \texttt{a} \mid \texttt{b} \mid \texttt{c} \mid \texttt{d})$$

- `a`: Variables with common parameters
- `b`: Individual-specific variables with alternative-specific parameters
- `c`: Alternative-specific variables with alternative-specific parameters
- `d`: Individual-specific variables that affect the scale parameter

See the vignettes on the `mlogit` CRAN page for more information

- `cran.r-project.org/web/packages/mlogit/index.html`

Multinomial Logit Model Estimation

We model the representative utility of heating system j to household n as

$$V_{nj} = \alpha_j + \beta_1 IC_{nj} + \beta_2 OC_{nj}$$

```
## Model choice using cost data and alternative effects
model_mlogit <- mlogit(formula = depvar ~ ic + oc | 1 | 0,
                       data = heating_dfidx,
                       reflevel = 'hp')
```

Model Summary

```
## Summarize model results
summary(model_mlogit)
##
## Call:
## mlogit(formula = depvar ~ ic + oc | 1 | 0, data = heating_dfidx,
##     reflevel = "hp", method = "nr")
##
## Frequencies of alternatives:choice
##       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|)
## (Intercept):ec  1.65884594  0.44841936  3.6993 0.0002162 ***
## (Intercept):er  1.85343697  0.36195509  5.1206 3.045e-07 ***
## (Intercept):gc  1.71097930  0.22674214  7.5459 4.485e-14 ***
## (Intercept):gr  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)
```

Interpreting Parameters

`coef()` is the R function to display only the model coefficients

```
## Display model coefficients
coef(model_mlogit)
## (Intercept):ec (Intercept):er (Intercept):gc (Intercept):gr
##    1.658845944    1.853436967    1.710979303    0.308263280
##             ic             oc
##   -0.001533153   -0.006996368
```

How do we interpret these coefficients?

- Electric central, electric room, and gas central provide more utility than heat pump
- Gas room provides the same utility as heat pump
  - Parameter is positive but not statistically significant
- An additional $100 of installation cost reduces utility by 0.15
- An additional $100 of annual operating cost reduces utility by 0.70

Alternative-Specific Parameters on Demographics

We might think that the utility from central systems vs. room systems depends on the number of rooms in the home

- We can estimate alternative-specific parameters on the number of rooms

$$V_{nj} = \alpha_j + \beta_1 IC_{nj} + \beta_2 OC_{nj} + \gamma_j R_n$$

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

Alternative-Specific Parameters on Demographics

```
## Summarize model results
summary(model_mlogit_rooms)
##
## Call:
## mlogit(formula = depvar ~ ic + oc | rooms | 0, data = heating_dfidx,
##     reflevel = "hp", method = "nr")
##
## Frequencies of alternatives:choice
##       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|)
## (Intercept):ec  1.47832643  0.67123250  2.2024   0.02764 *
## (Intercept):er  1.75977504  0.58813686  2.9921   0.00277 **
## (Intercept):gc  1.77021331  0.44444627  3.9830 6.806e-05 ***
## (Intercept):gr  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 ***
## rooms:ec        0.03811449  0.10825890  0.3521   0.72479
## rooms:er        0.01939934  0.10278196  0.1887   0.85029
## rooms:gc       -0.01294329  0.08476424 -0.1527   0.87864
## rooms:gr       -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
## Likelihood ratio test : chisq = 28.831 (p.value = 6.5482e-05)
```

Alternative-Specific Parameters on Costs

We might think that the marginal utility of installation cost depends on the type of heating system

- We can estimate alternative-specific parameters on installation cost

$$V_{nj} = \alpha_j + \beta_{1j} IC_{nj} + \beta_2 OC_{nj}$$

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

Alternative-Specific Parameters on Costs

```
## Summarize model results
summary(model_mlogit_costs)
##
## Call:
## mlogit(formula = depvar ~ oc | 1 | ic, data = heating_dfidx,
##     reflevel = "hp", method = "nr")
##
## Frequencies of alternatives:choice
##       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|)
## (Intercept):ec  1.70480715  1.27495400  1.3372 0.181173
## (Intercept):er  2.52929123  1.20296858  2.1025 0.035506 *
## (Intercept):gc  1.46257610  1.01069767  1.4471 0.147870
## (Intercept):gr -0.56099198  1.13846726 -0.4928 0.622182
## oc             -0.00545750  0.00182214 -2.9951 0.002743 **
## ic:hp          -0.00159724  0.00101723 -1.5702 0.116373
## ic:ec          -0.00214993  0.00122995 -1.7480 0.080468 .
## ic:er          -0.00265151  0.00093039 -2.8499 0.004374 **
## ic:gc          -0.00120808  0.00082802 -1.4590 0.144567
## ic:gr          -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
## Likelihood ratio test : chisq = 31.965 (p.value = 1.6569e-05)
```

Fitted Utilities

`fitted()` with `type = 'linpred'` calculates the fitted utilities of the model

$$V_{nj} = \alpha_j + \beta_1 IC_{nj} + \beta_2 OC_{nj}$$

```
## Look at fitted utilities
fitted(model_mlogit, type = 'linpred') %>%
  head()
##          hp        ec        er         gc        gr
## 1 -3.405191 -3.530883 -3.210579 -1.0138360 -2.229100
## 2 -2.879079 -3.202592 -2.921923 -0.5850563 -2.035239
## 3 -2.806872 -2.516635 -2.358279 -0.3665739 -1.856372
## 4 -3.167658 -2.887513 -2.395670 -0.8349672 -1.937065
## 5 -2.602633 -2.487319 -2.382925 -0.6711906 -1.961024
## 6 -2.781384 -2.190857 -2.064932 -0.2594666 -1.968485
```

Fitted Choice Probabilities

`fitted()` with `type = 'probabilities'` calculates the fitted choice probabilities of the model

$$P_{ni} = \frac{e^{\alpha_i + \beta_1 IC_{ni} + \beta_2 OC_{ni}}}{\sum_j e^{\alpha_j + \beta_1 IC_{nj} + \beta_2 OC_{nj}}}$$

```
## Look at fitted choice probabilities
fitted(model_mlogit, type = 'probabilities') %>%
  head()
##           hp         ec         er        gc        gr
## 1 0.05791494 0.05107444 0.07035738 0.6329116 0.1877416
## 2 0.06701658 0.04849337 0.06420595 0.6644519 0.1558322
## 3 0.05565974 0.07440281 0.08716904 0.6387765 0.1439919
## 4 0.05489595 0.07264503 0.11879834 0.5657376 0.1879231
## 5 0.08219005 0.09223575 0.10238514 0.5670663 0.1561227
## 6 0.05112739 0.09228185 0.10466584 0.6366616 0.1152634
```

Marginal Effects

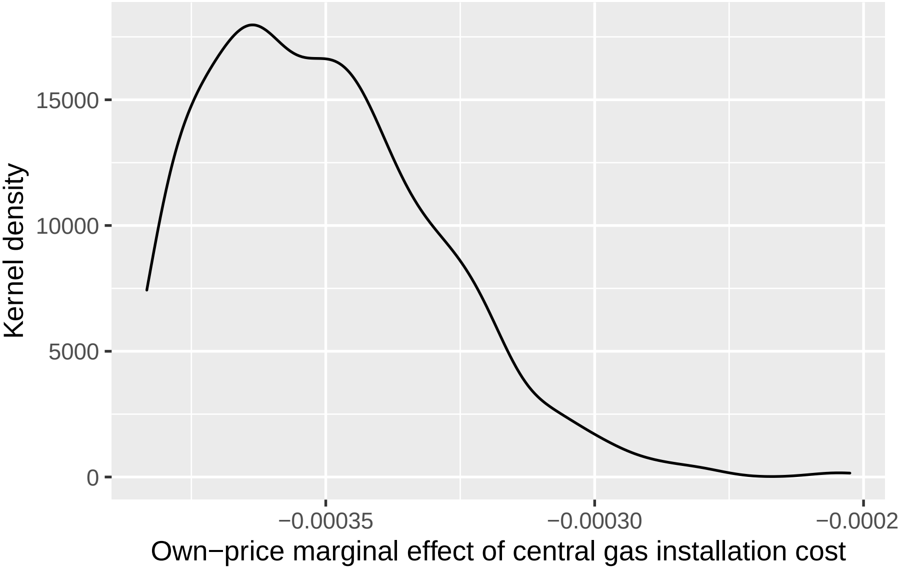
$$\frac{\partial P_{ni}}{\partial z_{ni}} = \beta_z P_{ni}(1 - P_{ni})$$

```r
## Calculate probability of central gas
heating <- heating %>%
  mutate(prob_gc_mlogit =
           fitted(model_mlogit, type = 'probabilities')[, 4])
## Calculate own-price marginal effects for central gas
heating <- heating %>%
  mutate(mfx_gc_ic_mlogit =
           coef(model_mlogit)[5] * prob_gc_mlogit * (1 - prob_gc_mlogit),
         mfx_gc_oc_mlogit =
           coef(model_mlogit)[6] * prob_gc_mlogit * (1 - prob_gc_mlogit))
```

Distribution of Installation Cost Marginal Effect

```r
## Plot kernel density of own-price elasticity of ic for gc
heating %>%
  ggplot(aes(x = mfx_gc_ic_mlogit)) +
  geom_density() +
  xlab('Own-price marginal effect of central gas installation cost') +
  ylab('Kernel density')
```

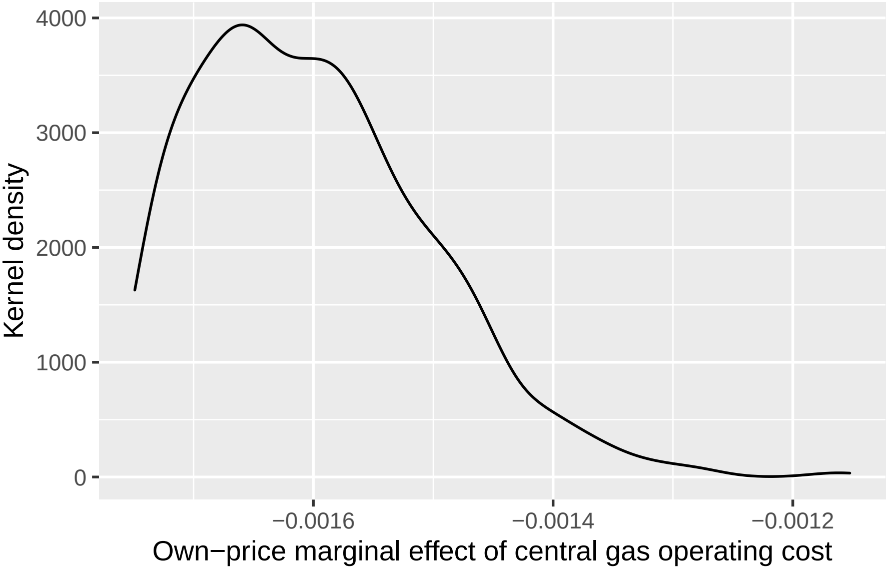
Distribution of Installation Cost Marginal Effect



Distribution of Operating Cost Marginal Effect

```
## Plot kernel density of own-price elasticity of oc for gc
heating %>%
  ggplot(aes(x = mfx_gc_oc_mlogit)) +
  geom_density() +
  xlab('Own-price marginal effect of central gas operating cost') +
  ylab('Kernel density')
```

Distribution of Operating Cost Marginal Effect



Elasticities

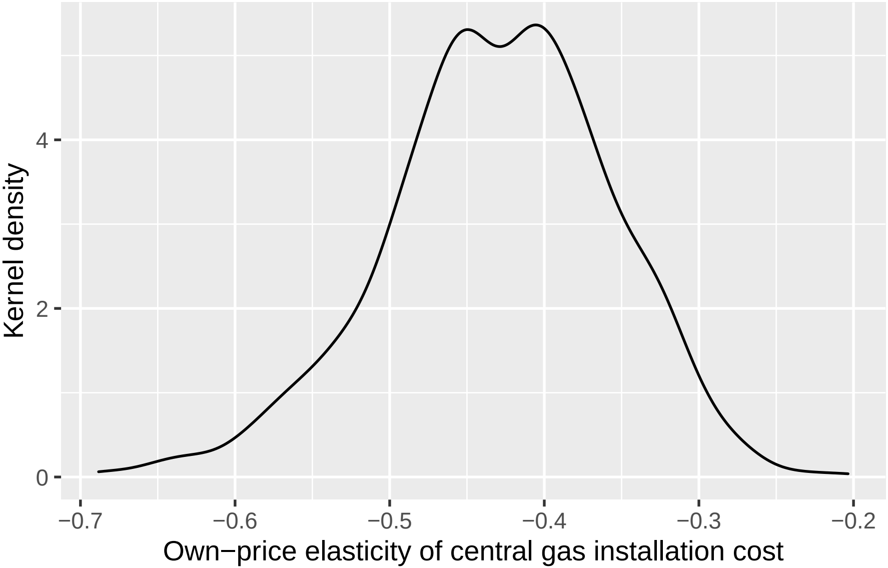
$$E_{iz_{ni}} = \beta_z z_{ni}(1 - P_{ni})$$

```
## Calculate own-price elasticities for central gas
heating <- heating %>%
  mutate(elas_gc_ic_mlogit =
           coef(model_mlogit)[5] * ic.gc * (1 - prob_gc_mlogit),
         elas_gc_oc_mlogit =
           coef(model_mlogit)[6] * oc.gc * (1 - prob_gc_mlogit))
```

Distribution of Installation Cost Elasticity

```
## Plot kernel density of own-price elasticity of ic for gc
heating %>%
  ggplot(aes(x = elas_gc_ic_mlogit)) +
  geom_density() +
  xlab('Own-price elasticity of central gas installation cost') +
  ylab('Kernel density')
```

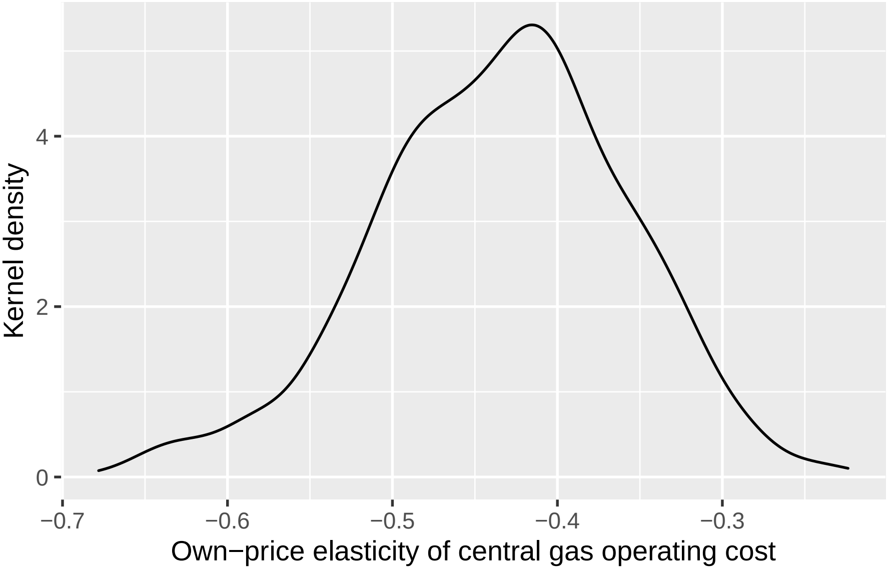
Distribution of Installation Cost Elasticity



Distribution of Operating Cost Elasticity

```
## Plot kernel density of own-price elasticity of oc for gc
heating %>%
  ggplot(aes(x = elas_gc_oc_mlogit)) +
  geom_density() +
  xlab('Own-price elasticity of central gas operating cost') +
  ylab('Kernel density')
```

Distribution of Operating Cost Elasticity



Cross Elasticities

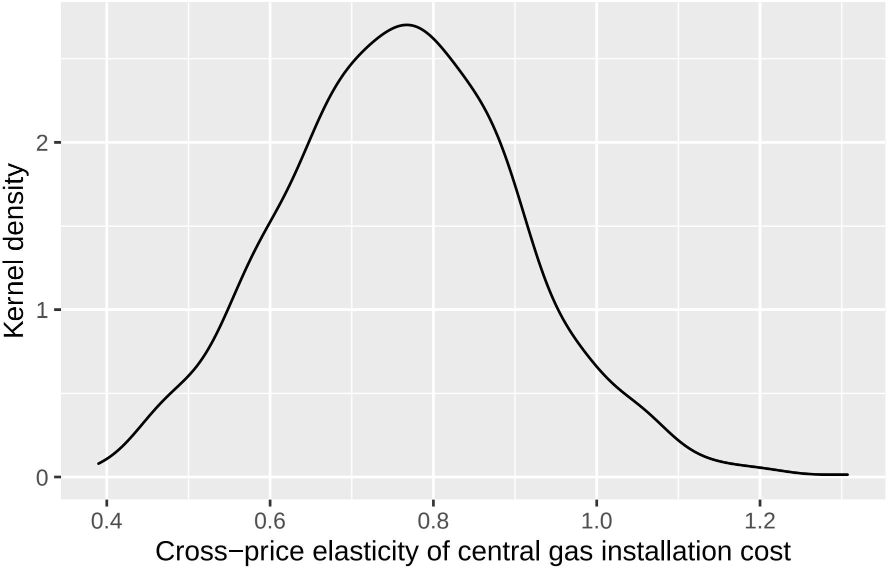
$$E_{iz_{nj}} = -\beta_z z_{nj} P_{nj}$$

```
## Calculate own-price elasticities for central gas
heating <- heating %>%
  mutate(crosselas_gc_ic_mlogit =
           -coef(model_mlogit)[5] * ic.gc * prob_gc_mlogit,
         crosselas_gc_oc_mlogit =
           -coef(model_mlogit)[6] * oc.gc * prob_gc_mlogit)
```

Distribution of Installation Cost Cross Elasticity

```r
## Plot kernel density of cross-price elasticity of ic for gc
heating %>%
  ggplot(aes(x = crosselas_gc_ic_mlogit)) +
  geom_density() +
  xlab('Cross-price elasticity of central gas installation cost') +
  ylab('Kernel density')
```

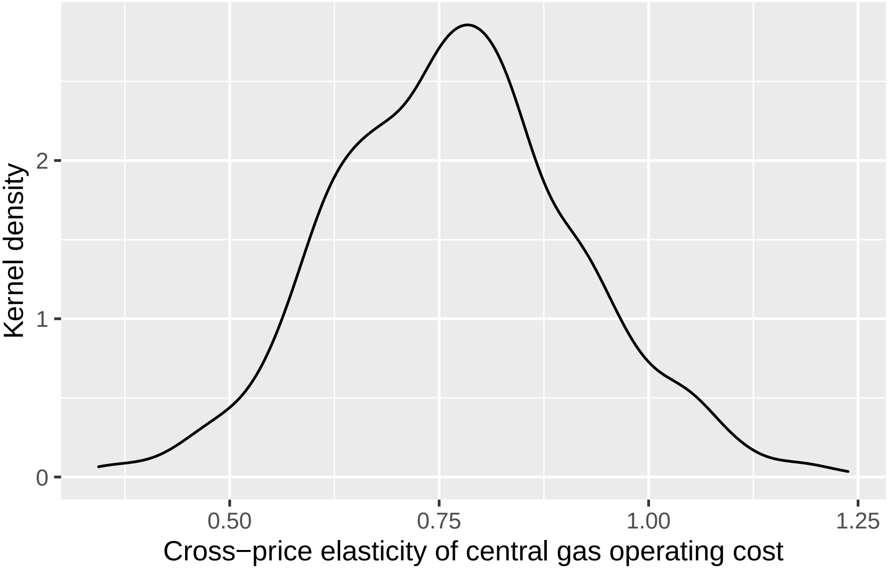
Distribution of Installation Cost Cross Elasticity



Distribution of Operating Cost Cross Elasticity

```r
## Plot kernel density of cross-price elasticity of oc for gc
heating %>%
  ggplot(aes(x = crosselas_gc_oc_mlogit)) +
  geom_density() +
  xlab('Cross-price elasticity of central gas operating cost') +
  ylab('Kernel density')
```

Distribution of Operating Cost Cross Elasticity



Mean Marginal Effects

`effects()` with `type = 'aa'` calculates the full set of $J \times J$ marginal effects for one `covariate` at the data means

- Columns correspond to outcomes
- Rows correspond to covariates

```
## Calculate marginal effects of ic at data means
effects(model_mlogit, covariate = 'ic', type = 'aa')
##               hp            ec            er            gc
## hp -8.017461e-05  5.724478e-06  7.570980e-06  5.469354e-05
## ec  5.724478e-06 -9.643283e-05  9.224310e-06  6.663736e-05
## er  7.570980e-06  9.224310e-06 -1.245630e-04  8.813208e-05
## gc  5.469354e-05  6.663737e-05  8.813209e-05 -3.513129e-04
## gr  1.218561e-05  1.484667e-05  1.963565e-05  1.418499e-04
##               gr
## hp  1.218561e-05
## ec  1.484667e-05
## er  1.963565e-05
## gc  1.418499e-04
## gr -1.885179e-04
```

Mean Elasticities

`effects()` with `type = 'rr'` calculates the full set of $J \times J$ elasticities for one `covariate` at the data means

- Columns correspond to outcomes
- Rows correspond to covariates

```
## Calculate elasticities of ic at data means
effects(model_mlogit, covariate = 'ic', type = 'rr')
##             hp          ec          er          gc          gr
## hp -1.51559794  0.08881805  0.08881805  0.08881806  0.08881806
## ec  0.08526389 -1.17888750  0.08526389  0.08526389  0.08526390
## er  0.13456469  0.13456470 -1.37394760  0.13456469  0.13456469
## gc  0.76749550  0.76749551  0.76749551 -0.42349863  0.76749549
## gr  0.20290165  0.20290166  0.20290166  0.20290164 -1.21031326
```

Multinomial Logit with Heterogeneous Parameters

We have estimated a single “average” parameter for each cost variable

- But in reality, marginal utility is likely to vary by income

$$V_{nj} = \alpha_j + \beta_{1n} IC_{nj} + \beta_{2n} OC_{nj}$$

$$\beta_{1n} = \frac{\beta_1}{I_n} \quad \text{and} \quad \beta_{2n} = \frac{\beta_2}{I_n}$$

Estimate a model using each cost as a share of income

$$V_{nj} = \alpha_j + \beta_1 \frac{IC_{nj}}{I_n} + \beta_2 \frac{OC_{nj}}{I_n}$$

```r
## Model heating choice with costs divided by income
model_mlogit_inc <- mlogit(formula = depvar ~
                             I(ic / income) + I(oc / income) | 1 | 0,
                           data = heating_dfidx,
                           reflevel = 'hp')
```

Multinomial Logit with Heterogeneous Parameters

```
## Summarize model results
summary(model_mlogit_inc)
##
## Call:
## mlogit(formula = depvar ~ I(ic/income) + I(oc/income) | 1 | 0,
##     data = heating_dfidx, reflevel = "hp", method = "nr")
##
## Frequencies of alternatives:choice
##       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|)
## (Intercept):ec  0.4570416  0.2867524  1.5939  0.110969
## (Intercept):er  0.7557773  0.2304446  3.2796  0.001039 **
## (Intercept):gc  2.2223040  0.1941569 11.4459 < 2.2e-16 ***
## (Intercept):gr  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)
```

Interpreting Heterogeneous Parameters

```
## Display model coefficients
coef(model_mlogit_inc)
## (Intercept):ec (Intercept):er (Intercept):gc (Intercept):gr
##    0.457041639    0.755777343    2.222304017    0.789415732
##   I(ic/income)   I(oc/income)
##   -0.002263881   -0.005328921
```

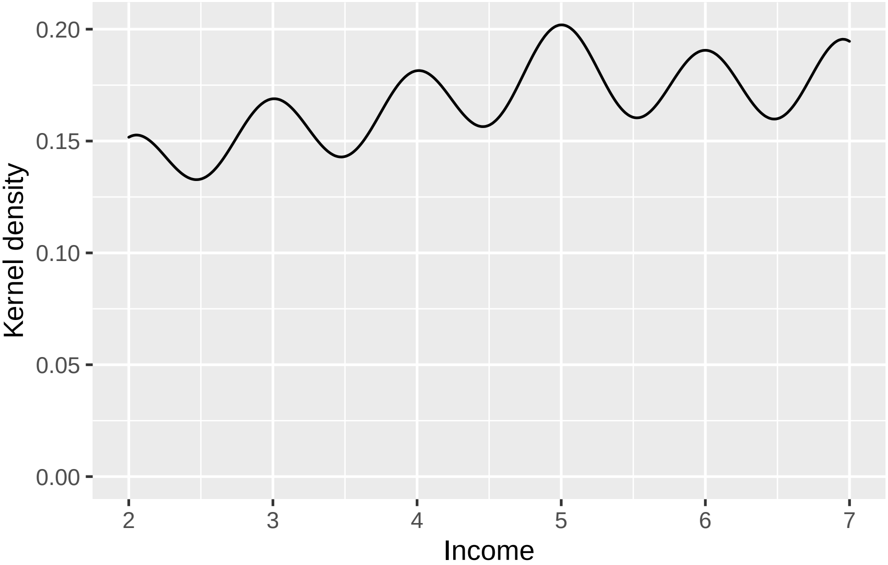
How do we interpret these parameters? (Ignoring statistical significance momentarily)

- Electric central, electric room, gas central, and gas room provide more utility than heat pump
- An additional 0.1 percentage point of installation cost as a share of income reduces utility by 0.02
- An additional 0.1 percentage point of annual operating cost as a share of income reduces utility by 1.05

Kernel Density of Income

```r
## Plot kernel density of income
heating %>%
  ggplot(aes(x = income)) +
  geom_density() +
  xlab('Income') +
  ylab('Kernel density')
```

Kernel Density of Income



Marginal Utility Depending on Income

What are the marginal utilities at \$30,000 income? \$50,000? \$70,000?

$$\beta_{1n} = \frac{\beta_1}{I_n} \quad \text{and} \quad \beta_{2n} = \frac{\beta_2}{I_n}$$

```
## Calculate marginal utility of costs when income == 3
coef(model_mlogit_inc)[5:6] / 3
##  I(ic/income)  I(oc/income)
## -0.0007546271 -0.0017763071

## Calculate marginal utility of costs when income == 5
coef(model_mlogit_inc)[5:6] / 5
##  I(ic/income)  I(oc/income)
## -0.0004527763 -0.0010657843

## Calculate marginal utility of costs when income == 7
coef(model_mlogit_inc)[5:6] / 7
##  I(ic/income)  I(oc/income)
## -0.0003234116 -0.0007612745
```

Multinomial Logit with Scale Parameters

We might think that the variance of the random utility term differs by region

- Weather plays an important role in the choice of a heating system, but it is only represented in our model through random utility

Estimate a model with a scale parameter for each region

$$V_{nj} = \frac{\alpha_j}{\sigma_r} + \frac{\beta_1}{\sigma_r} IC_{nj} + \frac{\beta_2}{\sigma_r} OC_{nj}$$

- But the `mlogit` package currently has a bug that causes the scale parameter estimation to fail if the data are too “clean”
- First, we have to strategically introduce `NA` values into our data that will not affect estimation

Scale Parameter Bug Work Around

```
## Create new dataset to work around bug
heating_nas <- heating_long %>%
  mutate(available = 1)
heating_nas <- heating_nas %>%
  bind_rows(heating_nas %>%
            filter(idcase == 1) %>%
            mutate(idcase = 9999) %>%
            mutate(available = 1 * choice))
heating_nas_dfidx <- dfidx(heating_nas, shape = 'long',
                          idx = c('idcase', 'alt'), choice = 'choice',
                          subset = available == 1)
```

Multinomial Logit with Scale Parameters

$$V_{nj} = \frac{\alpha_j}{\sigma_r} + \frac{\beta_1}{\sigma_r} IC_{nj} + \frac{\beta_2}{\sigma_r} OC_{nj}$$

The `mlogit()` function takes a `formula` with four separate sets of covariates to allow for four different kinds of parameters

$$\texttt{mlogit}(\texttt{formula} = \texttt{y} \sim \texttt{a} \mid \texttt{b} \mid \texttt{c} \mid \texttt{d})$$

- `a`: Variables with common parameters
- `b`: Individual-specific variables with alternative-specific parameters
- `c`: Alternative-specific variables with alternative-specific parameters
- `d`: Individual-specific variables that affect the scale parameter

```
## Model heating choice with regional scale parameters
model_mlogit_region <- mlogit(formula =
                                choice ~ ic + oc | 1 | 0 | region,
                              data = heating_nas_dfidx,
                              reflevel = 'hp')
```

Multinomial Logit with Scale Parameters

```
## Summarize model results
summary(model_mlogit_region)
##
## Call:
## mlogit(formula = choice ~ ic + oc | 1 | 0 | region, data = heating_nas_dfidx,
##     reflevel = "hp")
##
## Frequencies of alternatives:choice
##       hp       ec       er       gc       gr
## 0.055494 0.071032 0.093230 0.637070 0.143174
##
## bfgs method
## 7 iterations, 0h:0m:0s
## g'(-H)^-1g = 3.68E-07
## gradient close to zero
##
## Coefficients :
##                     Estimate  Std. Error z-value  Pr(>|z|)
## (Intercept):ec    1.55071130  0.44163351  3.5113 0.0004459 ***
## (Intercept):er    1.71844069  0.37035863  4.6399 3.485e-06 ***
## (Intercept):gc    1.57538121  0.24996788  6.3023 2.932e-10 ***
## (Intercept):gr    0.27826211  0.19426156  1.4324 0.1520266
## ic               -0.00143420  0.00061017 -2.3505 0.0187479 *
## oc               -0.00649788  0.00161426 -4.0253 5.690e-05 ***
## sig.regionscostl -0.03372534  0.10097328 -0.3340 0.7383776
## sig.regionmountn  0.04999917  0.15412748  0.3244 0.7456342
## sig.regionncostl -0.21788091  0.08281907 -2.6308 0.0085183 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -1003.5
## McFadden R^2:  0.018708
## Likelihood ratio test : chisq = 38.264 (p.value = 3.3387e-07)
```

Interpreting Scale Parameters

```
## Display model coefficients
coef(model_mlogit_region)
##   (Intercept):ec    (Intercept):er    (Intercept):gc    (Intercept):gr
##      1.550711303       1.718440694       1.575381209       0.278262115
##               ic                oc sig.regionscostl sig.regionmountn
##     -0.001434204      -0.006497878      -0.033725336       0.049999166
## sig.regionncostl
##     -0.217880913
```

How do we interpret these scale parameters?

$$\sigma_r = 1 + \sum_r \delta_r \mathbb{1}(\texttt{region}_n = r)$$

- Valley region is excluded and its scale parameter is normalized to one
- South coastal and mountain regions have the same random utility variance as the valley region and, hence, the same marginal utility parameters
- North coastal region has a smaller random utility variance than the valley region and, hence, larger (in absolute value) marginal utility parameters

Cost Trade-Offs

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?

$$
\begin{aligned}
U_{ni} &= \alpha_i + \beta_1 IC_{ni} + \beta_2 OC_{ni} + \varepsilon_{ni} \\
dU_{ni} &= \beta_1 dIC_{ni} + \beta_2 dOC_{ni} \\
dU_{ni} &= 0 \quad \Rightarrow \quad \frac{dIC_{ni}}{dOC_{ni}} = -\frac{\beta_2}{\beta_1}
\end{aligned}
$$

```
## Calculate ic equivalence of an increase in oc
-coef(model_mlogit)[6] / coef(model_mlogit)[5]
##        oc
## -4.563385
```

Implied Discount Rate

We can also use these structural parameters to determine what discount rate is implied by heating system decisions

- How the future is valued or "discounted" compared to today

If we assume an infinite time horizon for the annual operating cost, a general formula for a household's expected utility after installing heating system i is

$$U_{ni} = \phi_i + \phi_1 \left(IC_{ni} + \frac{1}{\gamma} OC_{ni} \right) + \omega_{ni}$$

where ϕ_1 is the marginal utility of income and γ is the discount rate

From our model, the utility from installing heating system i is

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

We have not estimated γ in the multinomial logit model, but we can use our structural parameters to calculate it

Implied Discount Rate Calculation

$$U_{ni} = \phi_i + \phi_1 \left(IC_{ni} + \frac{1}{\gamma} OC_{ni} \right) + \omega_{ni}$$
$$U_{ni} = \alpha_i + \beta_1 IC_{ni} + \beta_2 OC_{ni} + \varepsilon_{ni}$$

These two expressions for the equivalent utility imply that

$$\phi_1 = \beta_1$$
$$\frac{\phi_1}{\gamma} = \beta_2$$

which we can combine and rewrite as

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

```
## Calculate the implied discount rate
coef(model_mlogit)[5] / coef(model_mlogit)[6]
##        ic
## 0.2191356
```

Counterfactual Heat Pump Subsidy

To incentivize the adoption of heat pumps—the most energy efficient heating system—the Public Utilities Commission is considering a 50% subsidy on the installation cost of a heat pump

They want to know:

1. How would this subsidy have changed the number of heat pumps adopted by these 900 households?
2. How much consumer surplus would this subsidy have generated for these 900 households?

Counterfactual Dataset

The first step to conduct this counterfactual simulation is to construct the counterfactual dataset of costs with a 50% subsidy on heat pump installations

```
## Create data with 50% heat pump subsidy
heating_subsidy <- Heating %>%
  mutate(ic.hp = 0.5 * ic.hp)
## Convert subsidy data to dfidx format
heating_subsidy_dfidx <- dfidx(heating_subsidy, shape = 'wide',
                               choice = 'depvar', varying = 3:12)
```

Counterfactual Dataset

```
## Look at subsidy dataset
tibble(heating_subsidy_dfidx)
## # A tibble: 4,500 x 9
##    idcase depvar income agehed rooms region    ic    oc idx$id1 $id2
##     <dbl> <lgl>   <dbl>  <dbl> <dbl> <fct>  <dbl> <dbl>   <int> <fct>
##  1      1 FALSE       7     25     6 ncostl  860.  553.       1 ec
##  2      1 FALSE       7     25     6 ncostl  996.  506.       1 er
##  3      1 TRUE        7     25     6 ncostl  866   200.       1 gc
##  4      1 FALSE       7     25     6 ncostl  963.  152.       1 gr
##  5      1 FALSE       7     25     6 ncostl  568.  238.       1 hp
##  6      2 FALSE       5     60     5 scostl  797.  520.       2 ec
##  7      2 FALSE       5     60     5 scostl  895.  486.       2 er
##  8      2 TRUE        5     60     5 scostl  728.  169.       2 gc
##  9      2 FALSE       5     60     5 scostl  759.  169.       2 gr
## 10      2 FALSE       5     60     5 scostl  484.  199.       2 hp
## # ... with 4,490 more rows
```

Counterfactual Heating System Adoption

`predict()` with argument `newdata` defined as an indexed data frame calculates choice probabilities for every decision maker and alternative

$$\Delta E(A_i) = \sum_{n=1}^{N} P_{ni}^1 - \sum_{n=1}^{N} P_{ni}^0 = \sum_{n=1}^{N} \frac{e^{V_{ni}^1}}{\sum_{j=1}^{J^1} e^{V_{nj}^1}} - \sum_{n=1}^{N} \frac{e^{V_{ni}^0}}{\sum_{j=1}^{J^0} e^{V_{nj}^0}}$$

```r
## Calculate aggregate choices using observed data
agg_choices_obs <- predict(model_mlogit,
                           newdata = heating_dfidx) %>%
  colSums()
## Calculate aggregate choices using subsidy data
agg_choices_subsidy <- predict(model_mlogit,
                               newdata = heating_subsidy_dfidx) %>%
  colSums()
```

Counterfactual Heating System Adoption

```
## Calculate difference between aggregate choices in levels
agg_choices_subsidy - agg_choices_obs
##        hp         ec         er         gc         gr
##  53.793777  -3.929684  -5.171338 -36.437873  -8.254882

## Calculate difference between aggregate choices in percentages
(agg_choices_subsidy - agg_choices_obs) / agg_choices_obs
##         hp          ec          er          gc          gr
##  1.07587554 -0.06140132 -0.06156355 -0.06359140 -0.06399133
```

Counterfactual Consumer Surplus

`logsum()` with argument `data` defined as an indexed data frame calculates the log-sum term for every decision maker

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

```
## Calculate total log-sum value using observed data
logsum_obs <- logsum(model_mlogit, data = heating_dfidx) %>%
  sum()
## Calculate total log-sum value using subsidy data
logsum_subsidy <- logsum(model_mlogit, data = heating_subsidy_dfidx) %>%
  sum()
## Calculate change in consumer surplus from subsidy
(logsum_subsidy - logsum_obs) / (-coef(model_mlogit)[5])
##       ic
## 38466.63
```