Lecture 3: Random Utility Model

ResEcon 703: Topics in Advanced Econometrics

Matt Woerman
University of Massachusetts Amherst

Agenda

Last time

- R tutorial

Today

- Discrete Choice
- Random Utility Model
- Linear Probability Model
- Linear Probability Model Example in R

Upcoming

- Reading for next time
  - Train textbook, Chapters 3.1–3.6
  - Gruber and Proterba (1994)
- Problem set
  - Problem Set 1 is posted, due September 24

Discrete Choice

Discrete Choice

Many problems in microeconomics and related fields involve a decision maker choosing between a discrete set of alternatives

- Whether a self-employed person buys health insurance
- Which lake or river an angler visits to fish
- Which city a household chooses to locate in
- Which phone plan a household purchases and when they make calls
- Which appliances a household purchases
- Which health insurance plan an employee chooses
- What pollution control equipment a power plant installs
- Whether and when a city chooses to replace bus engines
- Which automobile a household purchases

Analyzing a Discrete Choice Problem

Three steps to set up and analyze a discrete choice problem

1. Specify the choice set
2. Formulate a model of how the agent chooses among the choice set
3. Estimate the unknown parameters on the model

Choice Set

The choice set defines all of the possible alternatives available to the decision maker

- Example: How to get to campus?
  - Drive alone, carpool, bus, bike, walk, Uber, stay home, etc.

Alternatives must be mutually exclusive and exhaustive

- Mutually exclusive: The agent may choose only one alternative, and choosing that alternative precludes choosing any other alternative
- Exhaustive: That agent must chooses one of the alternatives, so all possible alternatives must be included

The choice set will depend on the context, research question, data availability, etc.

Discrete Choice Model and Estimation

Step 2: Formulate a model of how the agent chooses among the choice set

- Random utility model: coming up next

Step 3: Estimate the unknown parameters on the model

- Estimation methods: the rest of the semester

Random Utility Model

Random Utility Model

Discrete choices are usually modeled under the assumption of utility-maximizing behavior by the decision maker (or profit maximization when the decision maker is a firm)

The random utility model (RUM) provides such a framework

- The agent gets some amount of utility from each of the alternatives
  - The amount of utility can depend on observed characteristics of the alternatives, observed characteristics of the decision maker, and unobserved characteristics
- The agent selects the alternative that provides the greatest utility

Models derived from RUM are consistent with utility (or profit) maximization, even if the decision maker does not maximize utility

- RUMs can be highly flexible and include behavioral and information parameters that diverge from the traditional neoclassical model

Specifying a Random Utility Model

The model from the perspective of the decision maker

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

But we (the econometricians) do not observe U_{nj}!

- We observe
  - The choice
  - Some attributes of each alternative
  - Some attributes of the decision maker
- We will use these data to infer U_{nj}

Utility Decomposition

Decompose the utility of each alternative, U_{nj}, into two components

- Observed factors: V_{nj}
- Unobserved factors: ε_{nj}

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

$V_{nj} = V(x_{nj}, s_n)$ is called representative utility

- x_{nj}: Attributes of the alternatives
- s_n: Attributes of the decision maker
- Also depends on parameters, but we will get to those later

ε_{nj} is everything that affects utility not included in V_{nj}

- Depends importantly on the specification of V_{nj}
- Unobserved, so we treat this term as random
- $f(\varepsilon_n)$ is the joint density of the random vector $\varepsilon_n = \{\varepsilon_{n1}, \ldots, \varepsilon_{nJ}\}$ for decision maker n

Choice Probabilities

U_{nj} contains a random components, so we cannot say for certain which alternative the decision maker will choose

- But we can form probabilities!

The probability the decision maker chooses alternative i is

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

This probability is the cumulative distribution of $\varepsilon_{nj} - \varepsilon_{ni}$

- Multidimensional integral over the density of the unobserved component of utility, $f(\varepsilon_n)$
- Assumptions about $f(\varepsilon_n)$ yield different discrete choice models

Choice Probabilities 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 choice

We specify the representative utility of each alternative as

$$V_c = \alpha T_c + \beta M_c$$
$$V_b = \alpha T_b + \beta M_b$$

Suppose α and β are known (we will talk about estimating them later)

- Then V_c and V_b are known, so we know which has greater representative utility
- But unobserved factors also affect this decision: ε_c and ε_b

The probability of each choice is

$$P_c = \Pr(\varepsilon_b - \varepsilon_c < V_c - V_b)$$
$$P_b = \Pr(\varepsilon_c - \varepsilon_b < V_b - V_c)$$

Properties of the Random Utility Model

The random utility model implies two important properties about discrete choice models

- Only difference in utility matter
  - We ultimately do not care about the level of utility from any alternative, just the comparisons between any two alternatives
  - We can only estimate parameters that capture differences between alternatives
- The scale of utility is arbitrary
  - Scaling all utilities does not change the comparison between alternatives
  - We will usually normalize the variance of the error terms

We will talk about these properties more when we talk about estimation

Linear Probability Model

Binary Choice

The discrete choice problem is greatly simplified with only two alternatives

- With only two alternatives, there is only one comparison to model

The choice probabilities can be fully described with only one modeling equation

$$P_{n1} = \Pr(\varepsilon_{n2} - \varepsilon_{n1} < V_{n1} - V_{n2})$$

- If the choice set is mutually exclusive and exhaustive, then it must be the case that $P_{n2} = 1 - P_{n1}$

We will typically assume representative utility is linear: $V_{ni} = \beta' x_{ni}$

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

Linear Probability Model

Let's abstract from the structural model and consider a non-structural approach to estimate a binary choice model

- From the structural model, the choice is a function of $x_n = \{x_{n1}, x_{n2}\}$

A simple regression to estimate this relationship is

$$Y_n = \alpha' x_n + \omega_n$$

where

- $Y_n = 1$ if and only if n chooses alternative 1

Under standard OLS assumptions

$$\Pr(Y_n = 1 \mid x_n) = E(Y_n \mid x_n) = \alpha' x_n$$

So this is called the linear probability model (LPM)

Pros and Cons of the Linear Probability Model

Pros

- You can estimate the LMP using OLS
  - Regression is fast and easy to run
  - Assumptions are transparent and well-known
- Coefficients can be interpreted as marginal effects

Cons

- Probabilities are not bounded by $[0, 1]$
  - Coefficients can be biased and inconsistent
- Coefficients are not structural parameters
- Error terms are heteroskedastic and not normally distributed

Linear Probability Model Example in R

Linear Probability Model Example

We are studying how consumers make choices about expensive and highly energy-consuming appliances in their homes. We have data on 600 households who rent a studio apartment and whether or not they choose to purchase a window air conditioning unit. For each household, we observe the purchase price of the air conditioner and its annual operating cost. (To simplify things, we assume there is only one "representative" air conditioner for each household and how much the household operates the air conditioner is fixed.) We can use a linear probability model to see how the purchase price and the operating cost affect the decision to purchase.

$$Y_n = \beta_0 + \beta_1 P_n + \beta_2 C_n + \omega_n$$

where

- $Y_n = 1$ if and only if n purchases an air conditioner
- P_n is the purchase price of the air conditioner
- C_n is the annual operating cost of the air conditioner

Load Dataset

```
### Load and look at dataset
## Load tidyverse
library(tidyverse)
```

```
## Load dataset
data <- read_csv('ac_renters.csv')

## Parsed with column specification:
## cols(
##   air_conditioning = col_logical(),
##   cost_system = col_double(),
##   cost_operating = col_double(),
##   income = col_double(),
##   residents = col_double(),
##   city = col_double()
## )
```

Dataset

```
## Look at dataset
data
## # A tibble: 600 x 6
##    air_conditioning cost_system cost_operating income residents  city
##    <lgl>                  <dbl>          <dbl>  <dbl>     <dbl> <dbl>
##  1 FALSE                    620            258     74         1     1
##  2 FALSE                    685            141     74         1     1
##  3 FALSE                    570            152     57         1     1
##  4 TRUE                     497            193     81         1     1
##  5 TRUE                     541            162     59         2     1
##  6 FALSE                    663            160     50         2     1
##  7 FALSE                    579            185     60         1     1
##  8 FALSE                    502            158     61         1     1
##  9 TRUE                     562            132     48         3     1
## 10 FALSE                    495            111     44         1     1
## # ... with 590 more rows
```

Linear Probability Model Regression

```
### Model air conditioning as a linear probability model
## Regress air conditioning on cost variables
reg_lmp <- data %>%
  lm(formula = air_conditioning ~ cost_system + cost_operating)
```

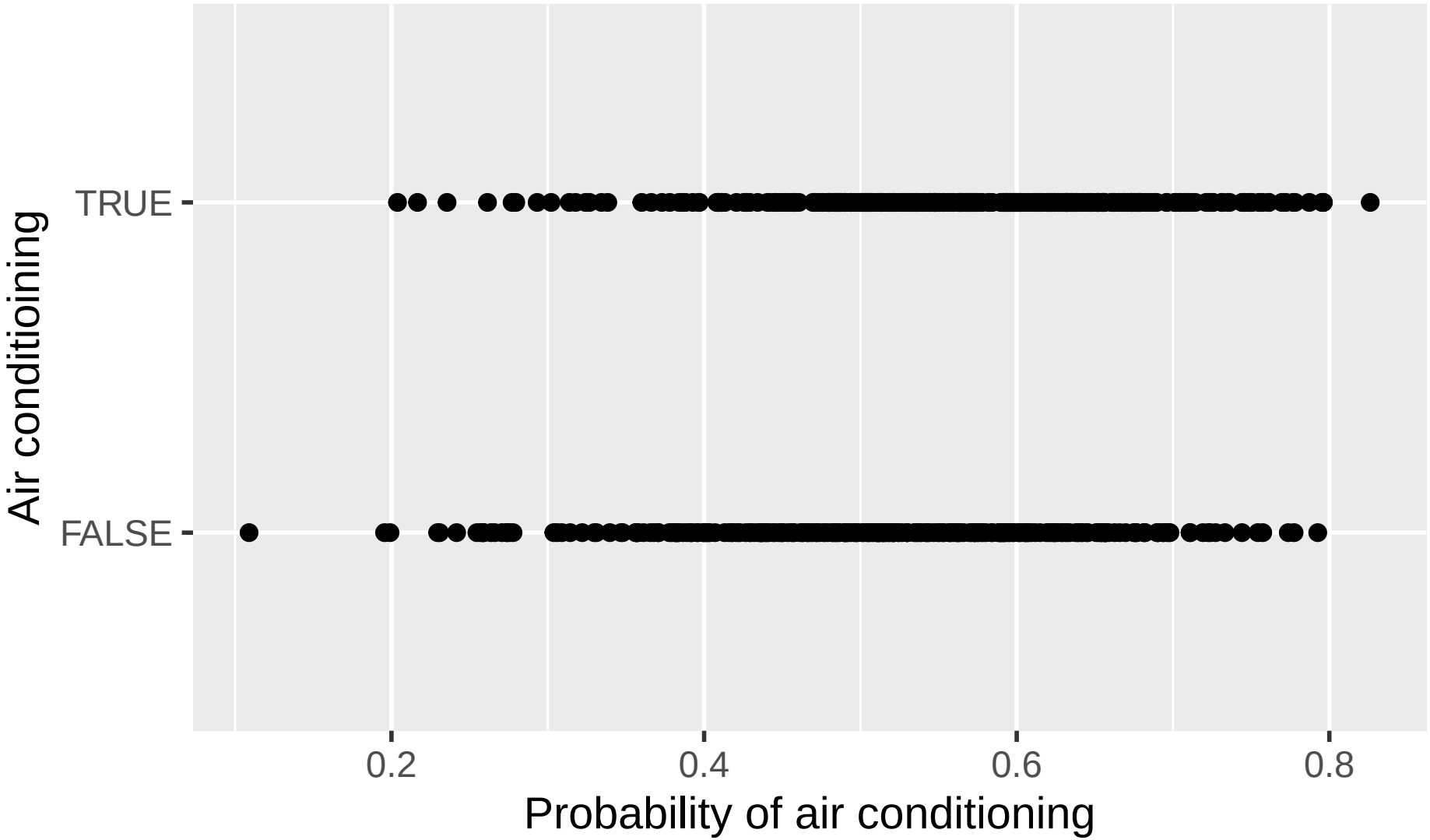
Regression Summary

```
## Summarize regression results
reg_lmp %>%
  summary()
##
## Call:
## lm(formula = air_conditioning ~ cost_system + cost_operating,
##     data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7926 -0.5013  0.2716  0.4314  0.7961
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.6745533  0.2120082   7.899 1.36e-14 ***
## cost_system    -0.0012918  0.0003402  -3.797 0.000161 ***
## cost_operating -0.0023549  0.0004857  -4.849 1.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4832 on 597 degrees of freedom
## Multiple R-squared:  0.06324,Adjusted R-squared:  0.0601
## F-statistic: 20.15 on 2 and 597 DF,  p-value: 3.394e-09
```

Visualize Choice Probability

```
### Visualize probability of air conditioning adoption
## Calculate probability of air conditioining
data <- data %>%
  mutate(probability_ac_lmp = predict(reg_lmp))
## Plot air conditioning vs. probability of air conditioning
data %>%
  ggplot(aes(x = probability_ac_lmp, y = air_conditioning)) +
  geom_point() +
  xlab('Probability of air conditioning') +
  ylab('Air conditioining')
```

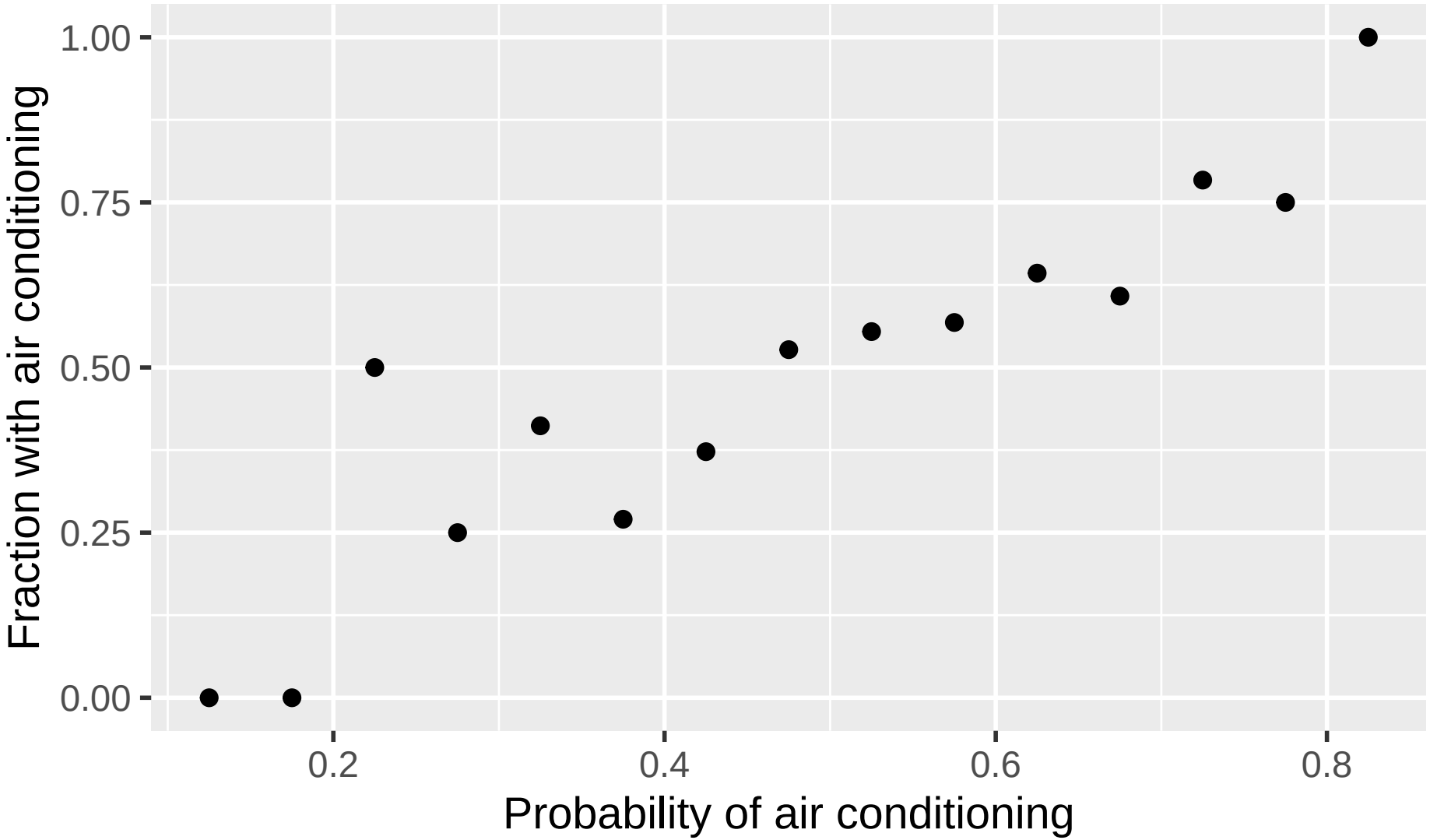
Choice Probability Plot



Visualize Choice Probability with Bins

```
### Visualize probability of air conditioning using bins
## Plot fraction vs. probability of air conditioning using bins
data %>%
  mutate(bin = cut(probability_ac_lmp,
                   breaks = seq(0, 1, 0.05),
                   labels = 1:20)) %>%
  group_by(bin) %>%
  summarize(fraction_ac = mean(air_conditioning)) %>%
  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')
```

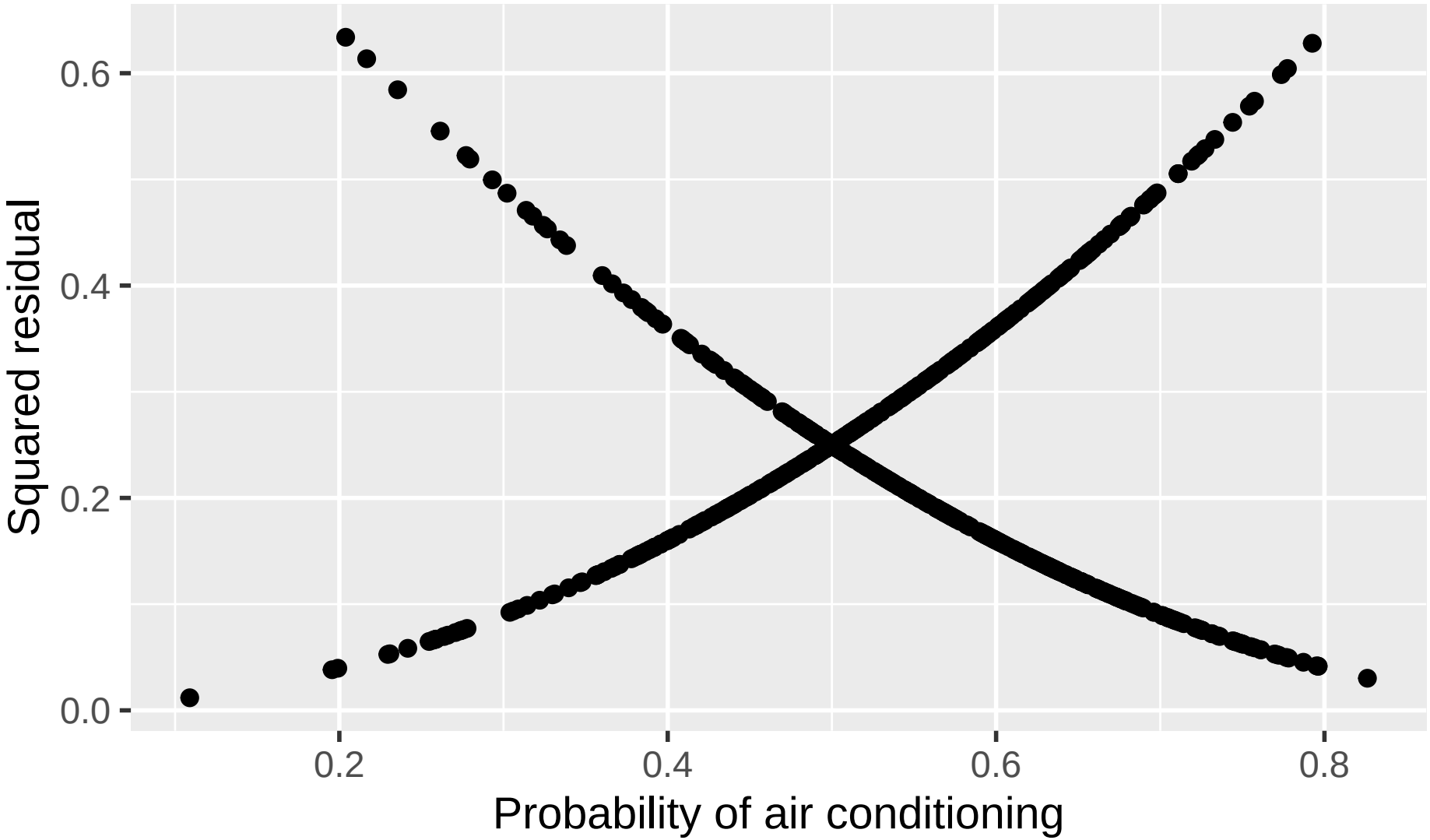
Choice Probability Plot with Bins



Visualize Heteroskedastic Residuals

```
### Visualize heteroskedastic residuals
## Calculate squared residuals
data <- data %>%
  mutate(sq_residual_lmp = (air_conditioning - probability_ac_lmp)^2)
## Plot squared residual vs. probability of air conditioning
data %>%
  ggplot(aes(x = probability_ac_lmp, y = sq_residual_lmp)) +
  geom_point() +
  xlab('Probability of air conditioning') +
  ylab('Squared residual')
```

Heteroskedastic Residuals Plot



Heteroskedastic-Robust Standard Errors

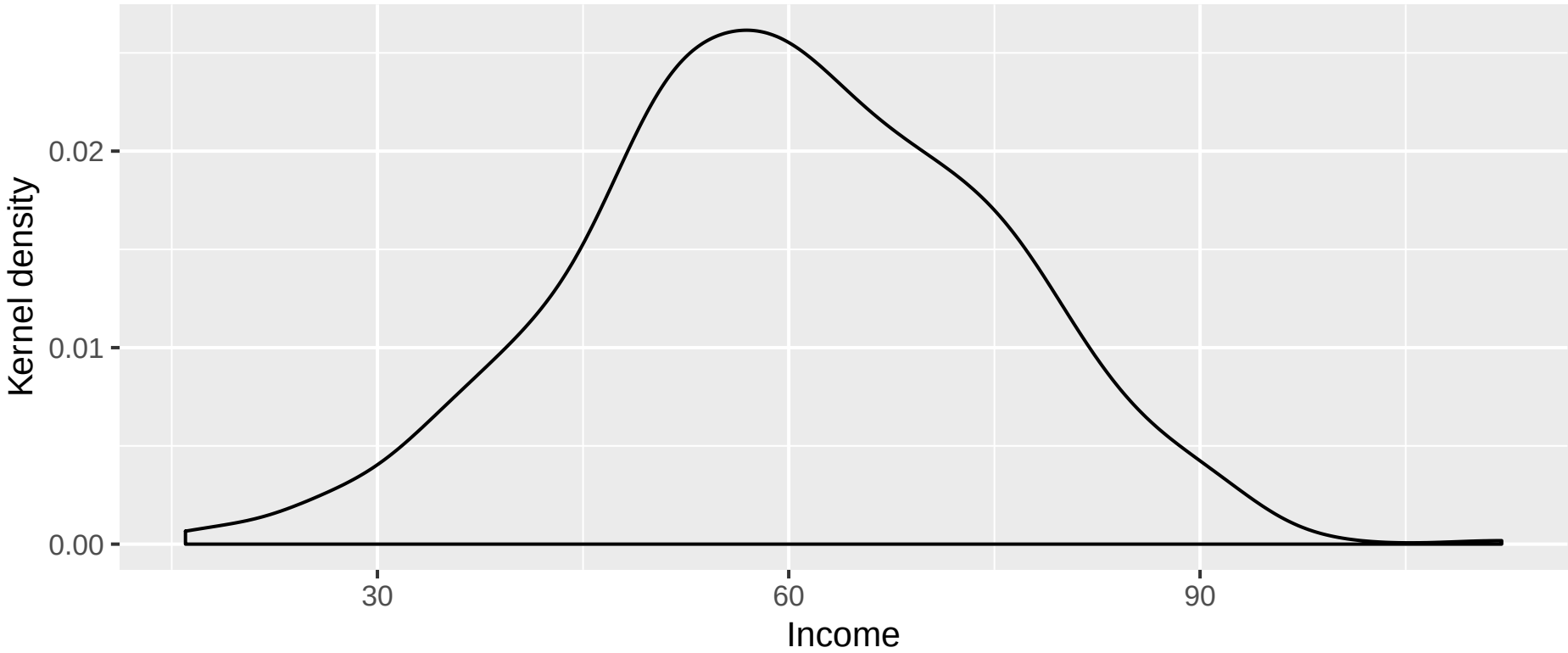
```
### Calculate heteroskedastic-robust standard errors
## Load lmtest and sandwich
library(lmtest)
library(sandwich)
## Summarize regression results with robust standard errors
reg_lmp %>%
  coeftest(vcov = vcovHC(reg_lmp))
##
## t test of coefficients:
##
##                   Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)     1.67455330  0.20499142  8.1689 1.862e-15 ***
## cost_system    -0.00129176  0.00033683 -3.8351 0.0001389 ***
## cost_operating -0.00235491  0.00047989 -4.9071 1.194e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

LMP with Heterogeneous Coefficients

```
### Model air conditioning with heterogeneous cost coefficients
## Regress air conditioning on costs divided by income
reg_lmp_income <- data %>%
  lm(formula = air_conditioning ~ I(cost_system / income) +
       I(cost_operating / income))
## Summarize regression results with robust standard errors
reg_lmp_income %>%
  coeftest(vcov = vcovHC(reg_lmp_income))
##
## t test of coefficients:
##
##                             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                1.3242258  0.0609637 21.7215 < 2.2e-16 ***
## I(cost_system/income)     -0.0340815  0.0076552 -4.4521 1.015e-05 ***
## I(cost_operating/income) -0.1502152  0.0216845 -6.9273 1.113e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Visualize Income Data

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



Marginal Effects Depending on Income

```
### Calculate marginal effects of cost variables
## Calculate marginal effects of costs when income == 60
coef(reg_lmp_income)[2:3] / 60
##    I(cost_system/income) I(cost_operating/income)
##             -0.0005680255             -0.0025035870

## Calculate marginal effects of costs when income == 30
coef(reg_lmp_income)[2:3] / 30
##    I(cost_system/income) I(cost_operating/income)
##              -0.001136051              -0.005007174

## Calculate marginal effects of costs when income == 90
coef(reg_lmp_income)[2:3] / 90
##    I(cost_system/income) I(cost_operating/income)
##             -0.0003786837             -0.0016690580
```

LMP with Age as an Explanatory Variable

```
### Model air conditioning with residents as an explanatory variable
## Regress air conditoning on scaled costs and number of residents
reg_lmp_residents <- data %>%
  lm(formula = air_conditioning ~ I(cost_system / income) +
       I(cost_operating / income) + residents)
## Summarize regression results with robust standard errors
reg_lmp_residents %>%
  coeftest(vcov = vcovHC(reg_lmp_residents))
##
## t test of coefficients:
##
##                             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                 0.7646358  0.0772214  9.9019 < 2.2e-16 ***
## I(cost_system/income)      -0.0346750  0.0082373 -4.2095 2.955e-05 ***
## I(cost_operating/income) -0.1542973  0.0189966 -8.1224 2.640e-15 ***
## residents                   0.3432309  0.0166628 20.5987 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```