Week 3: Random Utility Model

ResEcon 703: Topics in Advanced Econometrics

Matt Woerman
University of Massachusetts Amherst

Agenda

Last two weeks

- Structural estimation
- R tutorial

This week's topics

- Discrete choice
- Random utility model
- Choice probabilities
- Linear probability model
- Linear probability model R example

This week's reading

- Train textbook, chapters 1–2

Discrete Choice

How to Construct a Structural Econometric Model

1. Start with economic theory
2. Transform economic model into econometric model
3. Estimate the econometric model

We will use a consistent framework for the first 1.5 steps

- Step 1: Discrete choice to maximize utility
- Step 2a: Random utility model
- (Step 2b: We will make different assumptions about the unobservables in the random utility model, yielding different econometric models)

Discrete Choice

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

- Which travel mode a commuter uses to get to work
- Which health insurance plan an employee chooses
- Which recreation site to visit
- What consumer goods a household purchases
- Which job a worker chooses
- Which city a household chooses to locate in
- What pollution control equipment a power plant installs
- Whether a city replaces a bus engine
- Which automobile a household purchases
- Which crop a farmer plants

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 of the model
   - These are the structural parameters that describe the decision maker's behavior, preferences, etc.

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
  - The general model is coming up next
  - We will talk about a few specific models throughout the semester

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 chosen alternative
  - Some attributes of each alternative
  - Some attributes of the decision maker
- We will use these data to infer U_{nj} and how each attribute affects U_{nj}

Model of Utility

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

- Utility of observed factors: V_{nj}
- Utility of unobserved factors: ε_{nj}

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

$V_{nj} = V(\boldsymbol{x}_{nj}, \boldsymbol{s}_n)$ is called representative utility

- $\boldsymbol{x}_{nj}$: Vector of attributes of the alternative
- $\boldsymbol{s}_n$: Vector of attributes of the decision maker

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

- Depends importantly on the specification of V_{nj}
- We treat this term as a random variable from our perspective
- $f(\boldsymbol{\varepsilon}_n)$ is the joint density of the random vector $\boldsymbol{\varepsilon}_n = \{\varepsilon_{n1}, \ldots, \varepsilon_{nJ}\}$ for decision maker n

Representative Utility

We model representative utility, V_{nj}, as a function of

- $\boldsymbol{x}_{nj}$: Vector of attributes of the alternative
- $\boldsymbol{s}_n$: Vector of attributes of the decision maker
- β: Vector of structural parameters

We usually specify representative utility as a linear function

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

- A linear function is highly flexible and can include interactions, squared terms, etc.
- Most utility functions can be closely approximated by a function that is linear in parameters
- Non-linear utility can greatly complicate estimation

Structural Parameters

With linear representative utility, the total utility that alternative j gives decision maker n is

$$U_{nj} = \beta' \boldsymbol{x}_{nj} + \varepsilon_{nj}$$

The structural parameters, β, tell us how the observable attributes relate to the unobserved utility

- With linear representative utility, these parameters are interpreted as marginal utilities
- We can recover other structural parameters with different models of utility, profit, etc.

We want to find the structural parameters that make these utilities consistent with the observed choices

- More on how to do this for the rest of the semester

Choice Probabilities

Choice Probabilities

If you knew the representative utility, V_{nj}, for every alternative, could you say for certain what the decision maker would choose?

- No!

We assume the decision maker chooses the alternative that maximizes total utility, not representative utility

- We model utility as containing an unobserved random component
- Knowing representative utility is not sufficient to make a definitive statement about which alternative maximizes utility

If we cannot model discrete choices with certainty, what can we do?

- We can make probabilistic statements!

Choice probabilities play an important role in discrete choice models

- Probability of the decision maker choosing each of the alternatives

General Formula for Choice Probabilities

The probability that decision maker n chooses alternative i is

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

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

- Multidimensional integral over the density of the unobserved component of utility, $f(\boldsymbol{\varepsilon}_n)$
- Assumptions about $f(\boldsymbol{\varepsilon}_n)$ 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 alternative

We specify the representative utility of each alternative as

$$V_{nc} = \beta_{0c} + \beta_1 T_{nc} + \beta_2 M_{nc}$$
$$V_{nb} = \beta_{0b} + \beta_1 T_{nb} + \beta_2 M_{nb}$$

Suppose the β coefficients are known

- Then V_{nc} and V_{nb} are known, so we know which travel mode has greater representative utility
- But unobserved factors also affect this decision: ε_{nc} and ε_{nb}

The choice probability of driving is

$$\begin{aligned} P_{nc} &= \Pr(\varepsilon_{nb} - \varepsilon_{nc} < V_{nc} - V_{nb}) \\ &= \Pr(\varepsilon_{nb} - \varepsilon_{nc} < (\beta_{0c} + \beta_1 T_{nc} + \beta_2 M_{nc}) - (\beta_{0b} + \beta_1 T_{nb} + \beta_2 M_{nb})) \\ &= \Pr(\varepsilon_{nb} - \varepsilon_{nc} < (\beta_{0c} - \beta_{ob}) + \beta_1 (T_{nc} - T_{nb}) + \beta_2 (M_{nc} - M_{nb})) \end{aligned}$$

Using Choice Probabilities To Estimate Parameters

We want to estimate the structural parameters of the model that describe the decision maker's preferences, behavior, etc.

- How do these choice probabilities help us with that?

We want to fit the model to the data, but what makes for a good fit?

- When decision maker n chooses alternative i, we want that choice probability, P_{ni}, to be close to 1
- And we want all other choice probabilities to be close to 0
- So "fitting the model" means finding the structural parameters that fit the choice probabilities to the observed choices

How we do that will depend on the assumptions we make about $f(\varepsilon_n)$

- More on that starting next week

Properties of the Random Utility Model

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

The general formula for choice probabilities reveals two important properties about the random utility model

- 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 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: $V_{ni} = \beta' \boldsymbol{x}_{ni}$

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

Linear Probability Model

Abstract from our structural model for the rest of this lecture and consider a nonstructural approach to estimate a binary choice model

- We will return to structural estimation next week and for the remainder of the course

From the structural model, the choice probability P_{n1} is a nonlinear function of data about each alternative, $\boldsymbol{x}_n = \{\boldsymbol{x}_{n1}, \boldsymbol{x}_{n2}\}$

- A simple linear analog of this choice probability is

$$Y_n = \boldsymbol{\alpha}' \boldsymbol{x}_n + \omega_n$$

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

Under standard OLS assumptions

$$\Pr(Y_n = 1 \mid \boldsymbol{x}_n) = E(Y_n \mid \boldsymbol{x}_n) = \boldsymbol{\alpha}' \boldsymbol{x}_n$$

So this OLS regression model is called the linear probability model (LPM)

Linear Probability Model 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

The choice probability of driving is

$$\begin{aligned} P_{nc} &= \Pr(\varepsilon_{nb} - \varepsilon_{nc} < V_{nc} - V_{nb}) \\ &= \Pr(\varepsilon_{nb} - \varepsilon_{nc} < (\beta_{oc} + \beta_1 T_{nc} + \beta_2 M_{nc}) - (\beta_{0b} + \beta_1 T_{nb} + \beta_2 M_{nb})) \\ &= \Pr(\varepsilon_{nb} - \varepsilon_{nc} < (\beta_{0c} - \beta_{ob}) + \beta_1 (T_{nc} - T_{nb}) + \beta_2 (M_{nc} - M_{nb})) \end{aligned}$$

A nonstructural approach to estimate this choice is to use OLS to estimate the linear probability model

$$Y_n = \alpha_0 + \alpha_1 T_{nc} + \alpha_2 T_{nb} + \alpha_3 M_{nc} + \alpha_4 M_{nb} + \omega_n$$

where $Y_n = 1$ if and only if n chooses to drive

Pros and Cons of the Linear Probability Model

Pros

- You can estimate the LPM 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
  - Marginal effects are generally not the same thing as marginal utility or any other parameter that defines preferences, behavior, etc.
- Error terms are heteroskedastic and not normally distributed

Whether the pros outweigh the cons depends on your context, research question, data, etc.

Nonstructural Approach for Multinomial Choice

We can use OLS to nonstructurally estimate a linear probability model of a binary choice

- It may or may not be the best method, but it is feasible and has some advantages

What if you have a multinomial choice (more than two alternatives)?

- Can you think of an OLS (or other nonstructural) approach to estimate all of the comparisons implicit in a multinomial choice?

As we move to more complicated choice settings, a structural approach becomes the most feasible way to credibly estimate a model of discrete choice

Linear Probability 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

Nonstructural Approach to Air Conditioner Choice

$$P_{n1} = \Pr(\varepsilon_{n0} - \varepsilon_{n1} < \beta_0 - \beta_1 P_n - \beta_2 C_n)$$

Can we estimate the parameters of this model using an OLS regression?

- No, it is nonlinear
- But we can use this model to inform a nonstructural approach to estimation

The probability that household n purchases an air conditioner depends on:

- The price of the air conditioner
- The annual operating cost of the air conditioner
- Some parameters
- Random errors

We can construct a linear probability model with all of these characteristics

Linear Probability Model for Air Conditioner Choice

We can use a linear probability model to see how the purchase price and the operating cost affect the decision to purchase.

$$Y_n = \alpha_0 + \alpha_1 P_n + \alpha_2 C_n + \omega_n$$

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

Note that this regression model is not the same as our structural model!

$$P_{n1} = \Pr(\varepsilon_{n0} - \varepsilon_{n1} < \beta_0 - \beta_1 P_n - \beta_2 C_n)$$

- But the choice probability from the structural model has informed what should go in our nonstructural OLS regression

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')

## Rows:  600 Columns:  8
## - Column specification ----------------------
## Delimiter:  ","
## dbl (7):  cost_system, cost_operating, elec_price, square_feet,
inc...
## lgl (1):  air_conditioning
##
## i Use ‘spec()‘ to retrieve the full column specification for this
data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet
this message.
```

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
```

Linear Probability Model Regression

We want to estimate the linear probability model

$$Y_n = \alpha_0 + \alpha_1 P_n + \alpha_2 C_n + \omega_n$$

`lm()` is the R function to fit a linear model (i.e., run an OLS regression)

```
## Regress air conditioning on cost variables
reg_lpm <- lm(formula = air_conditioning ~ cost_system + cost_operating,
              data = ac_data)
```

Regression Summary

`summary()` summarizes the results of the regression

```
## Summarize regression results
summary(reg_lpm)
##
## Call:
## lm(formula = air_conditioning ~ cost_system + cost_operating,
##     data = ac_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8755 -0.5068  0.2454  0.3955  0.8406
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.5045413  0.2073398   7.256 1.24e-12 ***
## cost_system    -0.0006750  0.0003362  -2.008   0.0451 *
## cost_operating -0.0034690  0.0004856  -7.144 2.66e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4739 on 597 degrees of freedom
## Multiple R-squared:  0.0891,Adjusted R-squared:  0.08605
## F-statistic:  29.2 on 2 and 597 DF,  p-value: 7.971e-13
```

Interpreting Coefficients

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

```
## Display regression coefficients
coef(reg_lpm)
##    (Intercept)    cost_system cost_operating
##   1.5045412592  -0.0006750464  -0.0034689824
```

How do we interpret these coefficients?

- An additional \$100 of purchase price decreases the probability of purchase by 6.75 percentage points
- An additional \$100 of annual operating cost decreases the probability of purchase by 34.69 percentage points

Fitted Probabilities

`predict()` calculates the fitted values of the regression

```
## Calculate probability of air conditioning
ac_data <- ac_data %>%
  mutate(probability_ac_lpm = predict(reg_lpm))
## Look at probabilities and other data
ac_data %>%
  select(air_conditioning, starts_with('cost'), probability_ac_lpm)
## # A tibble: 600 x 4
##    air_conditioning cost_system cost_operating probability_ac_lpm
##    <lgl>                  <dbl>          <dbl>                <dbl>
##  1 FALSE                    513            247                0.301
##  2 FALSE                    578            138                0.636
##  3 TRUE                     658            171                0.467
##  4 FALSE                    615            198                0.403
##  5 FALSE                    515            165                0.585
##  6 FALSE                    588            143                0.612
##  7 TRUE                     643            153                0.540
##  8 FALSE                    676            182                0.417
##  9 TRUE                     516            137                0.681
## 10 TRUE                     544            185                0.496
## # ... with 590 more rows
```

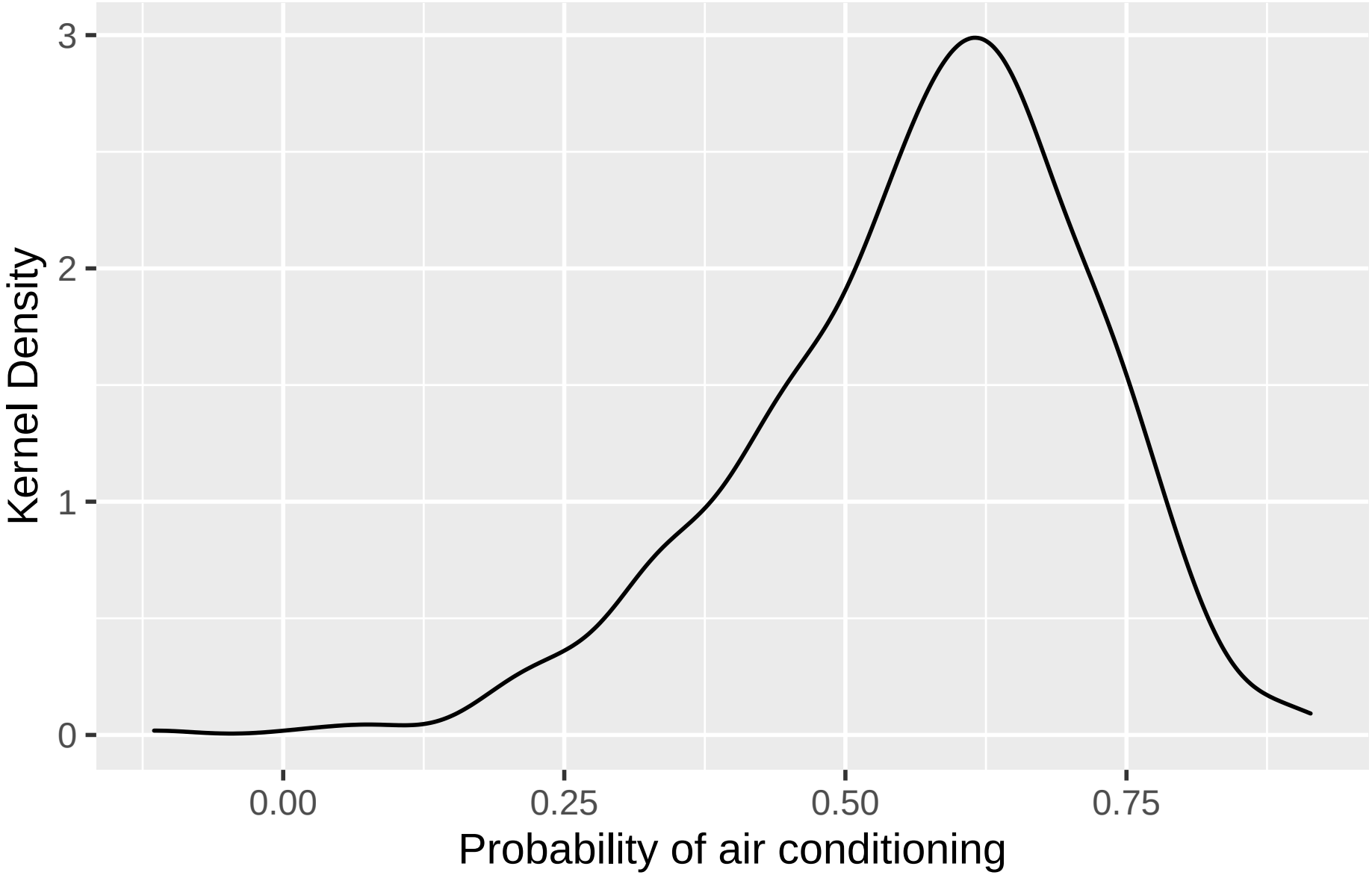
Kernel Density of Fitted Probabilities

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 probabilities
ac_data %>%
  ggplot(aes(x = probability_ac_lpm)) +
  geom_density() +
  xlab('Probability of air conditioning') +
  ylab('Kernel Density')
```

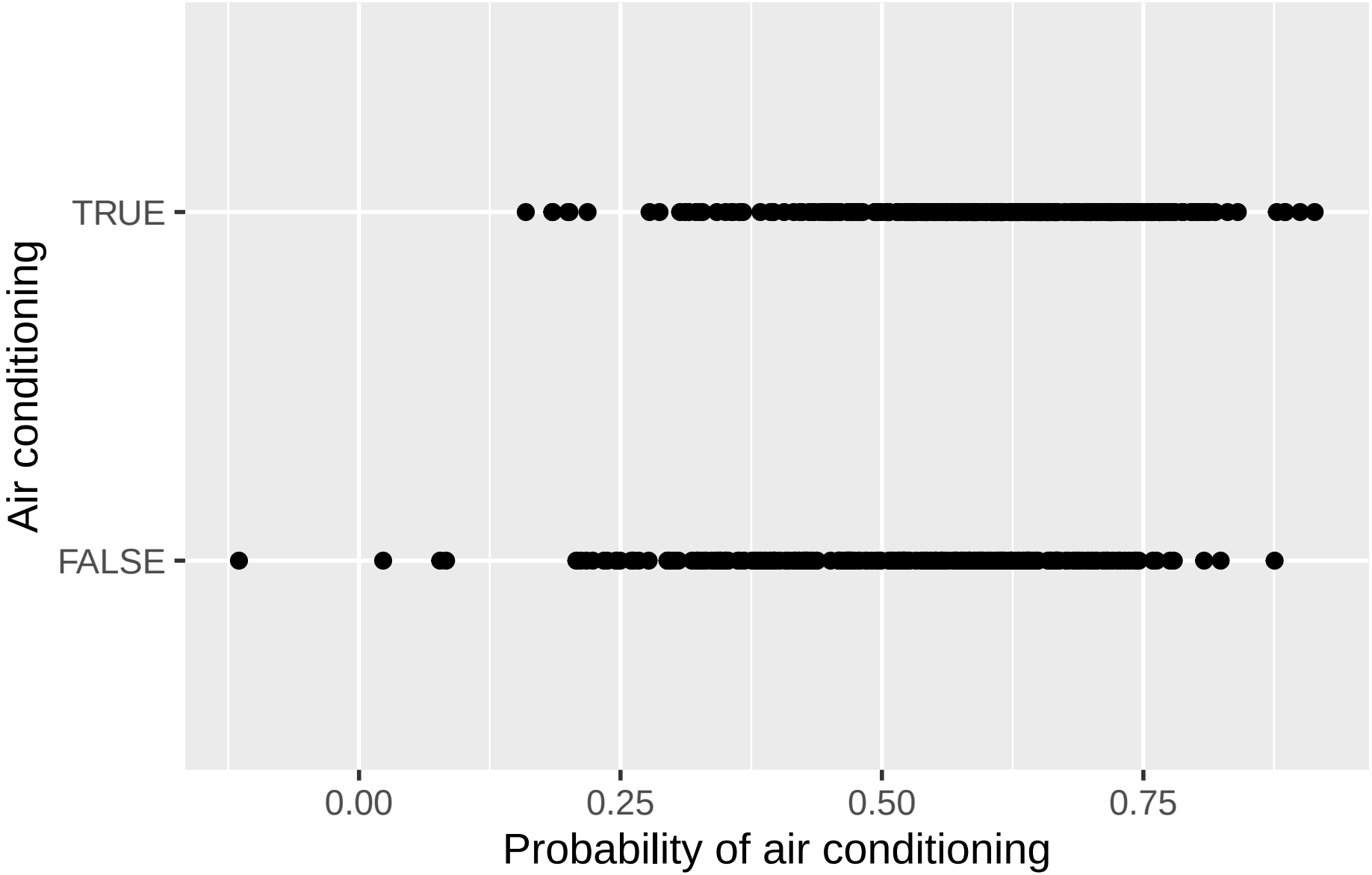
Kernel Density of Fitted Probabilities



Plot of Probability vs. Adoption

```
## Plot air conditioning vs. probability of air conditioning
ac_data %>%
  ggplot(aes(x = probability_ac_lpm, y = air_conditioning)) +
  geom_point() +
  xlab('Probability of air conditioning') +
  ylab('Air conditioning')
```

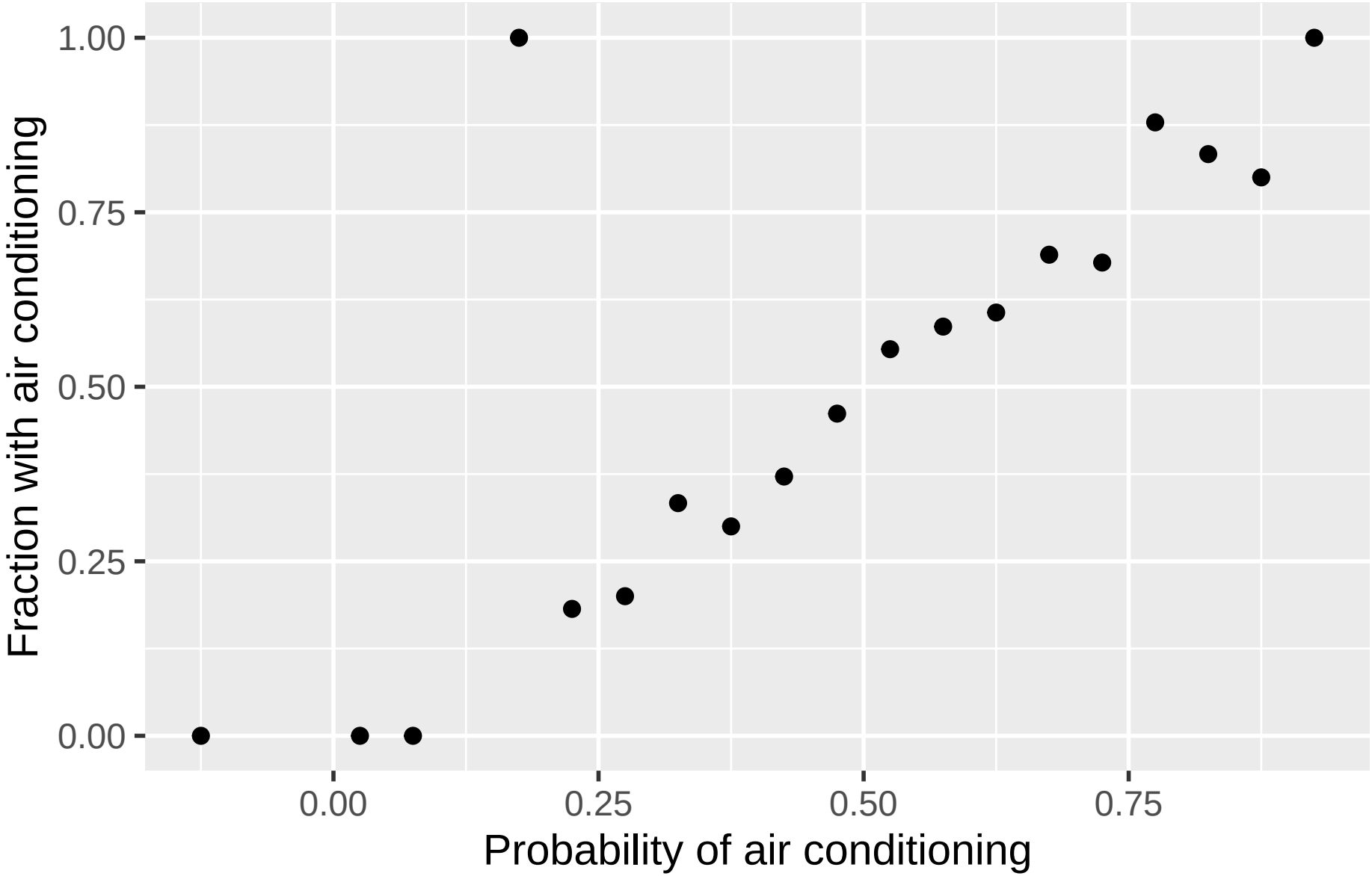
Plot of Probability vs. Adoption



Plot of Probability vs. Adoption with Bins

```
## Plot fraction vs. probability of air conditioning using bins
ac_data %>%
  mutate(bin = cut(probability_ac_lpm,
                   breaks = seq(-0.2, 1, 0.05),
                   labels = 1:24)) %>%
  group_by(bin) %>%
  summarize(fraction_ac = mean(air_conditioning), .groups = 'drop') %>%
  mutate(bin = as.numeric(bin),
         bin_mid = 0.05 * (bin - 1) + 0.025 - 0.2) %>%
  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 with Bins



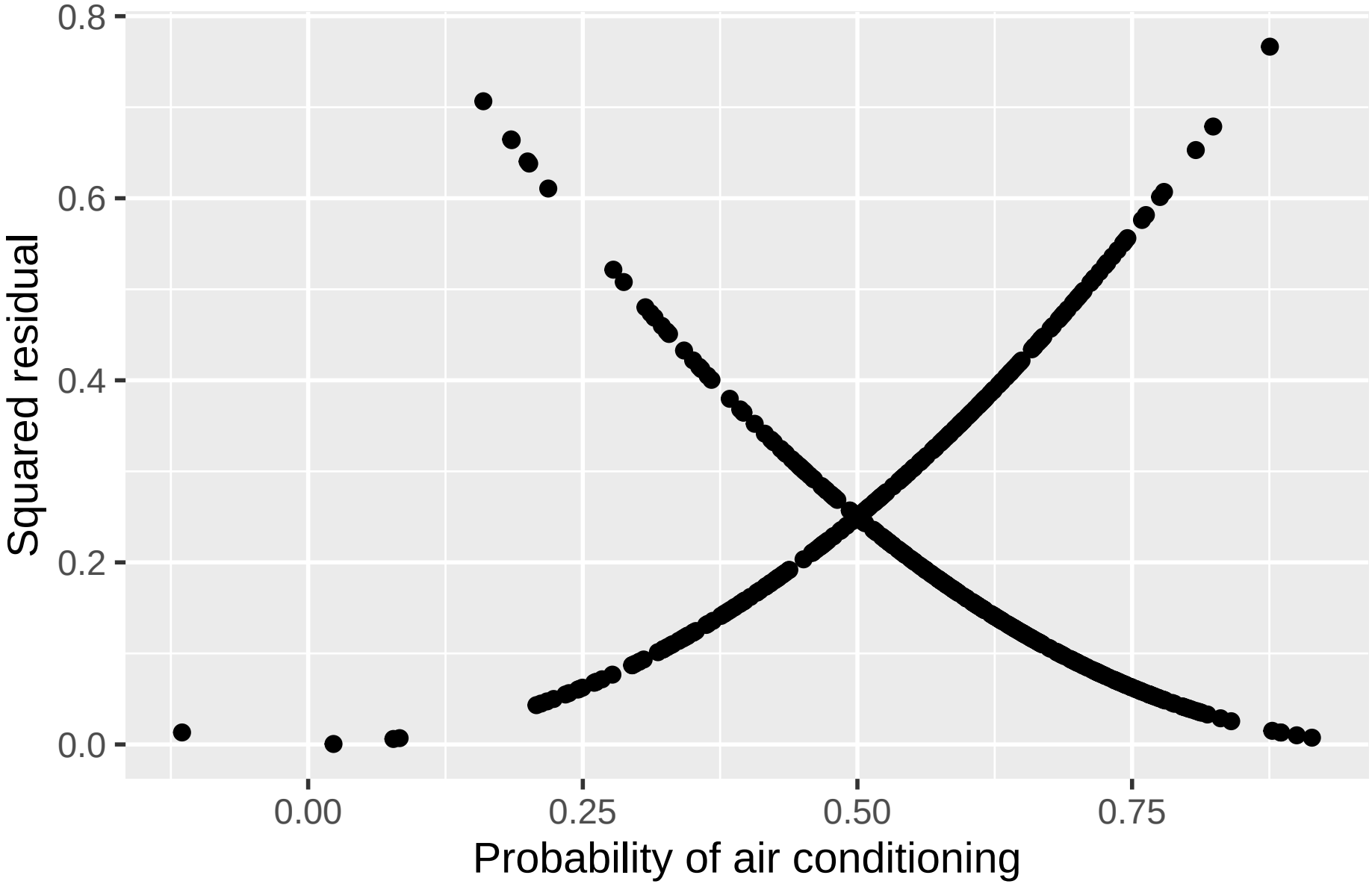
Heteroskedastic and Non-Normal Residuals

One criticism of a linear probability model is that the error terms are heteroskedastic and not normally distributed

- We can see this issue by calculating the residual for each observation and plotting residuals as a function of fitted probability

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

Heteroskedastic and Non-Normal Residuals



Heteroskedastic-Robust Standard Errors

`coeftest()` is a function from the `lmtest` package to test coefficients and summarize results using an alternate variance-covariance matrix

- `vcovHC()` is a function from the `sandwich` package to estimate a heteroskedastic-robust variance-covariance matrix

```
## Load lmtest and sandwich
library(lmtest)
library(sandwich)
## Summarize regression results with robust standard errors
reg_lpm %>%
  coeftest(vcov = vcovHC(reg_lpm))
##
## t test of coefficients:
##
##                    Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)     1.50454126  0.20074465  7.4948 2.408e-13 ***
## cost_system    -0.00067505  0.00033335 -2.0251   0.04331 *
## cost_operating -0.00346898  0.00046080 -7.5282 1.908e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

LPM with Heterogeneous Coefficients

We have estimated a single "average" effect for each price or cost variable

- But in reality, these effects are likely to vary by income

$$Y_n = \alpha_0 + \alpha_{1n} P_n + \alpha_{2n} C_n + \omega_n$$

$$\alpha_{1n} = \frac{\alpha_1}{I_n} \quad \text{and} \quad \alpha_{2n} = \frac{\alpha_2}{I_n}$$

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

$$Y_n = \alpha_0 + \alpha_1 \frac{P_n}{I_n} + \alpha_2 \frac{C_n}{I_n} + \omega_n$$

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

```
## Regress air conditioning on costs divided by income
reg_lpm_inc <- lm(formula = air_conditioning ~ I(cost_system / income) +
                    I(cost_operating / income),
                  data = ac_data)
```

LPM with Heterogeneous Coefficients

```
## Summarize regression results with robust standard errors
reg_lpm_inc %>%
  coeftest(vcov = vcovHC(reg_lpm_inc))
##
## t test of coefficients:
##
##                             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                 1.431453   0.060426 23.6893 < 2.2e-16 ***
## I(cost_system/income)      -0.036390   0.007809 -4.6599 3.903e-06 ***
## I(cost_operating/income) -0.176198   0.020526 -8.5842 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting Heterogeneous Coefficients

```
## Display regression coefficients
coef(reg_lpm_inc)
##                (Intercept)    I(cost_system/income)
##                 1.43145289               -0.03638958
## I(cost_operating/income)
##                -0.17619786
```

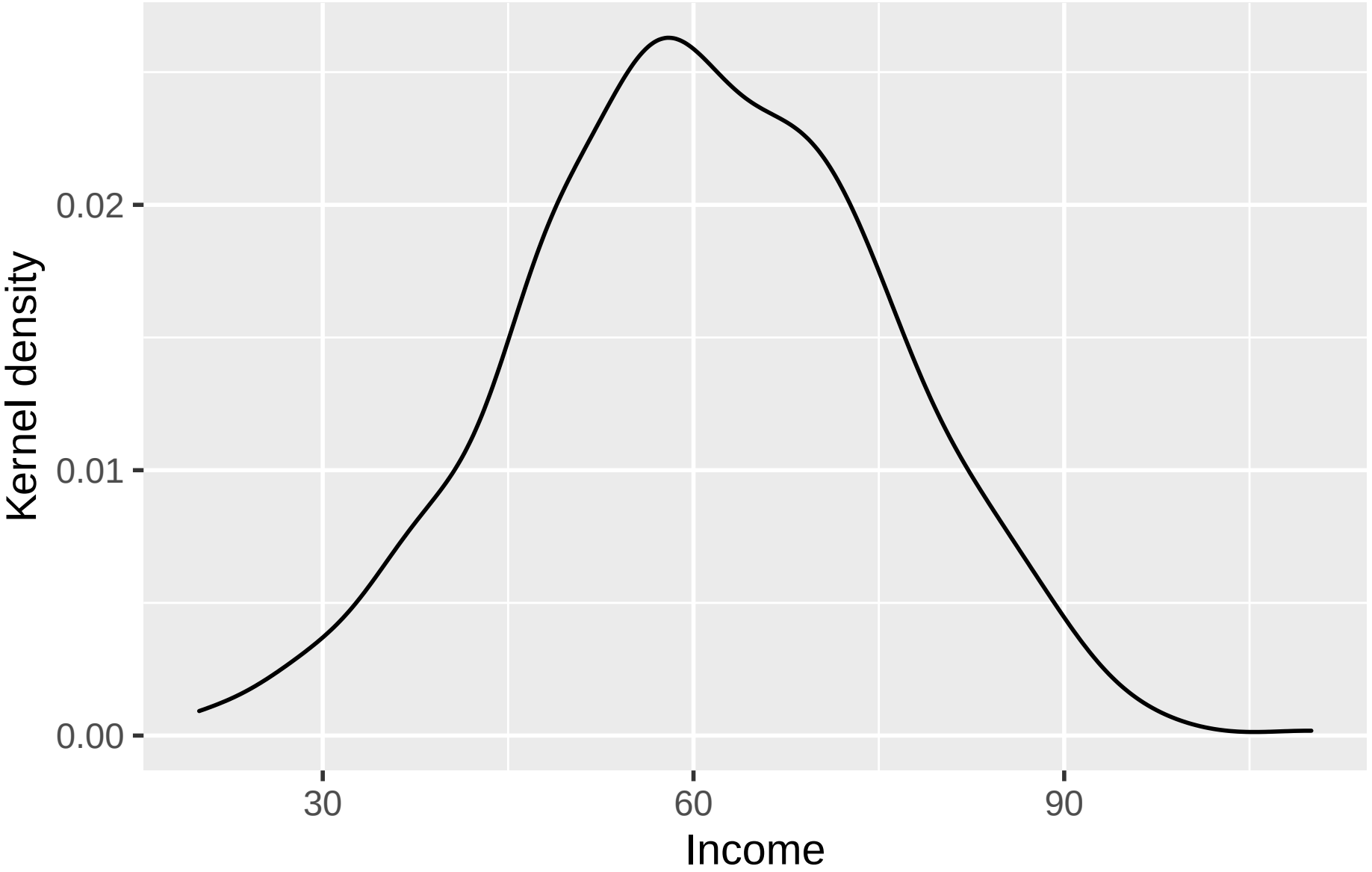
How do we interpret these coefficients?

- An additional 0.1 percentage point of purchase price as a share of income decreases the probability of purchase by 3.64 percentage points
- An additional 0.1 percentage point of annual operating cost as a share of decreases the probability of purchase by 17.62 percentage points

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 Effects Depending on Income

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

$$\alpha_{1n} = \frac{\alpha_1}{I_n} \quad \text{and} \quad \alpha_{2n} = \frac{\alpha_2}{I_n}$$

```
## Calculate marginal effects of costs when income == 30
coef(reg_lpm_inc)[2:3] / 30
##    I(cost_system/income) I(cost_operating/income)
##             -0.001212986             -0.005873262

## Calculate marginal effects of costs when income == 60
coef(reg_lpm_inc)[2:3] / 60
##    I(cost_system/income) I(cost_operating/income)
##             -0.0006064931             -0.0029366309

## Calculate marginal effects of costs when income == 90
coef(reg_lpm_inc)[2:3] / 90
##    I(cost_system/income) I(cost_operating/income)
##             -0.0004043287             -0.0019577540
```

LPM with Residents as an Explanatory Variable

We can include other attributes of households in the linear probability model

- Number of residents, size of apartment, etc.

```
## Regress air conditioning on scaled costs and number of residents
reg_lpm_res <- lm(formula = air_conditioning ~ I(cost_system / income) +
                    I(cost_operating / income) + residents,
                  data = ac_data)
## Summarize regression results with robust standard errors
reg_lpm_res %>%
  coeftest(vcov = vcovHC(reg_lpm_res))
##
## t test of coefficients:
##
##                              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)                 0.8592807  0.0719445  11.9437 < 2.2e-16 ***
## I(cost_system/income)      -0.0357436  0.0074886  -4.7731 2.285e-06 ***
## I(cost_operating/income) -0.1841899  0.0176557 -10.4323 < 2.2e-16 ***
## residents                   0.3430781  0.0159653  21.4890 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```