Lecture 16: Mixed Logit Model II

ResEcon 703: Topics in Advanced Econometrics

Matt Woerman
University of Massachusetts Amherst

Agenda

Last time

- Mixed Logit Model

Today

- Mixed Logit Model Example in R

Upcoming

- Reading for next time
  - Train textbook, Chapter 10
- Problem sets
  - Problem Set 3 is due
  - Problem Set 4 will be posted soon, due November 21

Mixed Logit

The utility that decision maker n obtains from alternative j is

$$U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj}$$

For a given β_n, the conditional choice probability is a standard logit

$$L_{ni}(\beta_n) = \frac{e^{\beta_n' x_{ni}}}{\sum_{j=1}^{J} e^{\beta_n' x_{nj}}}$$

We model β as a random variable with density $f(\beta \mid \theta)$ and integrate over this density to get the choice probability

$$P_{ni} = \int \frac{e^{\beta' x_{ni}}}{\sum_{j=1}^{J} e^{\beta' x_{nj}}} f(\beta \mid \theta) d\beta$$

We estimate θ, the parameters that define the distributions of coefficients

Mixed Logit Model Example in R

Mixed Logit Model Example

We are again studying how consumers make choices about expensive and highly energy-consuming systems in their homes. We have data on 250 households in California and the type of HVAC (heating, ventilation, and air conditioning) system in their home. Each household has the following choice set, and we observe the following data

Choice set

- GCC: gas central with AC
- ECC: electric central with AC
- ERC: electric room with AC
- HPC: heat pump with AC
- GC: gas central
- EC: electric central
- ER: electric room

Alternative-specific data

- ICH: installation cost for heat
- ICCA: installation cost for AC
- OCH: operating cost for heat
- OCCA: operating cost for AC

Household demographic data

- income: annual income

Load Dataset

```r
### Load and look at dataset
## Load tidyverse and mlogit
library(tidyverse)
library(mlogit)
## Load dataset from mlogit package
data('HC', package = 'mlogit')
```

Dataset

```
## Look at dataset
as_tibble(HC)
## # A tibble: 250 x 18
##    depvar ich.gcc ich.ecc ich.erc ich.hpc ich.gc ich.ec ich.er  icca
##    <fct>    <dbl>   <dbl>   <dbl>   <dbl>  <dbl>  <dbl>  <dbl> <dbl>
##  1 erc       9.7     7.86    8.79   11.4    24.1   24.5   7.37  27.3
##  2 hpc       8.77    8.69    7.09    9.37   28     32.7   9.33  26.5
##  3 gcc       7.43    8.86    6.94   11.7    25.7   31.7   8.14  22.6
##  4 gcc       9.18    8.93    7.22   12.1    29.7   26.7   8.04  25.3
##  5 gcc       8.05    7.02    8.44   10.5    23.9   28.4   7.15  25.4
##  6 gcc       9.32    8.03    6.22   12.6    27.0   21.4   8.6   19.9
##  7 gc        7.11    8.78    7.36   12.4    22.9   28.6   6.41  27.0
##  8 hpc       9.38    7.48    6.72    8.93   26.2   27.9   7.3   18.1
##  9 gcc       8.08    7.39    8.79   11.2    23.0   22.6   7.85  22.6
## 10 gcc       6.24    4.88    7.46    8.28   19.8   27.5   6.88  25.8
## # ... with 240 more rows, and 9 more variables: och.gcc <dbl>,
## #   och.ecc <dbl>, och.erc <dbl>, och.hpc <dbl>, och.gc <dbl>,
## #   och.ec <dbl>, och.er <dbl>, occa <dbl>, income <dbl>
```

Format Dataset in a Long Format

```r
### Format dataset
## Gather into a long dataset
hvac_long <- HC %>%
  mutate(id = 1:n()) %>%
  gather(key, value, starts_with('ich.'), starts_with('och.')) %>%
  separate(key, c('cost', 'alt')) %>%
  spread(cost, value) %>%
  mutate(choice = (depvar == alt)) %>%
  select(-depvar)
```

Dataset in a Long Format

```
## Look at long dataset
as_tibble(hvac_long)
## # A tibble: 1,750 x 8
##     icca  occa income    id alt     ich   och choice
##    <dbl> <dbl>  <dbl> <int> <chr> <dbl> <dbl> <lgl>
##  1  17    2.79     60   133 ec    20.3   4.52 FALSE
##  2  17    2.79     60   133 ecc    8.46  4.52 FALSE
##  3  17    2.79     60   133 er     7.7   4.32 FALSE
##  4  17    2.79     60   133 erc    8.16  4.32 FALSE
##  5  17    2.79     60   133 gc    25.3   2.26 FALSE
##  6  17    2.79     60   133 gcc    6.33  2.26 TRUE
##  7  17    2.79     60   133 hpc   11.1   1.63 FALSE
##  8  18.1  2.55     50    14 ec    25.6   5.21 FALSE
##  9  18.1  2.55     50    14 ecc   11.2   5.21 FALSE
## 10  18.1  2.55     50    14 er     9.3   3.8  FALSE
## # ... with 1,740 more rows
```

Clean Dataset

```r
## Combine heating and cooling costs into one variable
hvac_clean <- hvac_long %>%
  mutate(cooling = (nchar(alt) == 3),
         ic = if_else(cooling, ich + icca, ich),
         oc = if_else(cooling, och + occa, och)) %>%
  select(id, alt, choice, ic, oc, income)
```

Cleaned Dataset

```
## Look at cleaned dataset
as_tibble(hvac_clean)
## # A tibble: 1,750 x 6
##       id alt   choice    ic    oc income
##    <int> <chr> <lgl>  <dbl> <dbl>  <dbl>
##  1   133 ec    FALSE   20.3  4.52     60
##  2   133 ecc   FALSE   25.5  7.31     60
##  3   133 er    FALSE    7.7  4.32     60
##  4   133 erc   FALSE   25.2  7.11     60
##  5   133 gc    FALSE   25.3  2.26     60
##  6   133 gcc   TRUE    23.3  5.05     60
##  7   133 hpc   FALSE   28.1  4.42     60
##  8    14 ec    FALSE   25.6  5.21     50
##  9    14 ecc   FALSE   29.2  7.76     50
## 10    14 er    FALSE    9.3  3.8      50
## # ... with 1,740 more rows
```

Convert Dataset to mlogit Format

```
## Convert cleaned dataset to mlogit format
hvac_mlogit <- mlogit.data(hvac_clean, shape = 'long',
                           choice = 'choice', alt.var = 'alt')
```

Dataset in mlogit Format

```
## Look at data in mlogit format
as_tibble(hvac_mlogit)
## # A tibble: 1,750 x 6
##       id alt   choice    ic    oc income
##    <int> <fct> <lgl>  <dbl> <dbl>  <dbl>
##  1   133 ec    FALSE   20.3  4.52     60
##  2   133 ecc   FALSE   25.5  7.31     60
##  3   133 er    FALSE    7.7  4.32     60
##  4   133 erc   FALSE   25.2  7.11     60
##  5   133 gc    FALSE   25.3  2.26     60
##  6   133 gcc   TRUE    23.3  5.05     60
##  7   133 hpc   FALSE   28.1  4.42     60
##  8    14 ec    FALSE   25.6  5.21     50
##  9    14 ecc   FALSE   29.2  7.76     50
## 10    14 er    FALSE    9.3  3.8      50
## # ... with 1,740 more rows
```

Mixed Logit Models to Estimate with `mlogit()`

The representative utility of each alternative is

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

Random coefficients to consider

- β_1 and β_2 distributed normal
- β_1 and β_2 distributed log-normal
- β_1 and β_2 distributed log-normal and correlated

Mixed Logit Models in R

```r
### Mixed logit model using the mlogit package
## Help file for the mlogit function
?mlogit
## Arguments for mlogit mixed logit functionality
mlogit(formula, data, reflevel, rpar, correlation, R, seed, ...)
```

`mlogit()` arguments for mixed logit

1. `formula, data, reflevel`: same as a multinomial logit model
2. `rpar`: named vector of random coefficients and their distributions
3. `correlation`: `TRUE` models random coefficient correlations
4. `R`: number of simulations in MSLE
5. `seed`: seed for random draws in simulation

Mixed Logit with Normal Distributions

```r
### Model HVAC choice as a mixed logit
## Model choice using alternative intercepts and cost data with normal
## coefficients
model_1 <- hvac_mlogit %>%
  mlogit(formula = choice ~ ic + oc | 1 | 0, data = ., reflevel = 'hpc',
         rpar = c(ic = 'n', oc = 'n'), R = 1000, seed = 321)
```

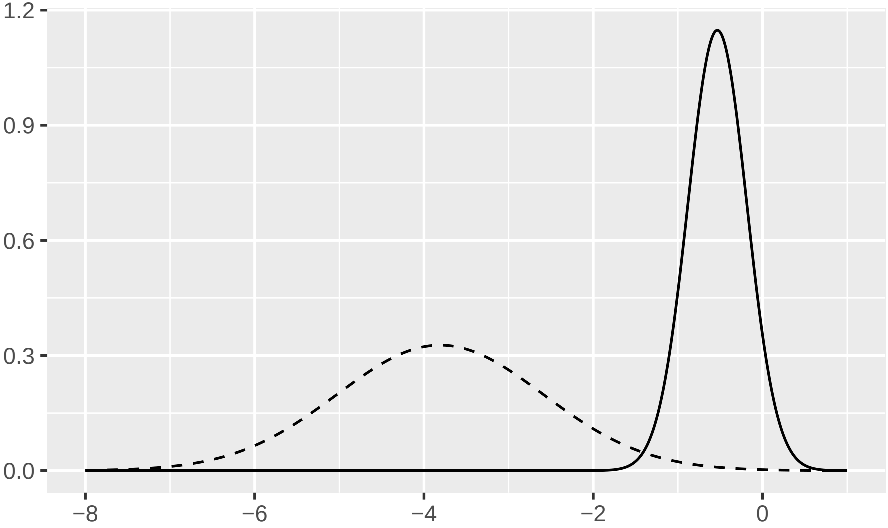
Model Results with Normal Distributions

```
## Summarize model results
model_1 %>%
  summary()
##
## Call:
## mlogit(formula = choice ~ ic + oc | 1 | 0, data = ., reflevel = "hpc",
##     rpar = c(ic = "n", oc = "n"), R = 1000, seed = 321)
##
## Frequencies of alternatives:
##   hpc    ec   ecc    er   erc    gc   gcc
## 0.104 0.004 0.016 0.032 0.004 0.096 0.744
##
## bfgs method
## 23 iterations, 0h:0m:19s
## g'(-H)^-1g = 2.33E-07
## gradient close to zero
##
## Coefficients :
##                  Estimate Std. Error z-value Pr(>|z|)
## ec:(intercept)  -13.59196    6.59838 -2.0599 0.039409 *
## ecc:(intercept)   4.55227    1.91196  2.3809 0.017268 *
## er:(intercept)  -26.84257   15.08477 -1.7794 0.075166 .
## erc:(intercept)   3.34782    2.28886  1.4627 0.143560
## gc:(intercept)  -15.47410    7.12512 -2.1718 0.029873 *
## gcc:(intercept)   4.60586    1.42533  3.2314 0.001232 **
## ic               -0.53396    0.24426 -2.1860 0.028816 *
## oc               -3.80889    1.41603 -2.6898 0.007149 **
## sd.ic             0.34764    0.24761  1.4040 0.160331
## sd.oc            -1.22028    0.82478 -1.4795 0.139003
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -192.75
## McFadden R^2:  0.14416
```

Normally Distributed Coefficients

```r
## Plot distributions of random coefficients
ggplot(data = data.frame(x = c(-8, 1)), aes(x)) +
  stat_function(fun = dnorm, n = 1001,
                args = list(mean = model_1$coefficients[7],
                            sd = abs(model_1$coefficients[9]))) +
  stat_function(fun = dnorm, n = 1001,
                args = list(mean = model_1$coefficients[8],
                            sd = abs(model_1$coefficients[10])),
                linetype = 'dashed') +
  xlab(NULL) +
  ylab(NULL)
```

Normally Distributed Coefficients Plot



Mixed Logit with Log-Normal Distributions

```r
### Model HVAC choice as a mixed logit
## Model choice using alternative intercepts and cost data with
## log-normal coefficients
model_2 <- hvac_mlogit %>%
  mlogit(formula = choice ~ ic + oc | 1 | 0, data = ., reflevel = 'hpc',
         rpar = c(ic = 'ln', oc = 'ln'), R = 1000, seed = 321)

## Warning in log(start[ln]):  NaNs produced
## Error in if (abs(x - oldx) < ftol) {:  missing value where TRUE/FALSE
needed
```

Format Dataset for Log-Normal Distributions

```r
### Reformat dataset with negative costs
## Convert cleaned dataset to mlogit format with negative costs
hvac_mlogit_neg <- mlogit.data(hvac_clean, shape = 'long',
                               choice = 'choice', alt.var = 'alt',
                               opposite = c('ic', 'oc'))
```

Dataset Formatted for Log-Normal Distributions

```
## Look at data in mlogit format
as_tibble(hvac_mlogit_neg)
## # A tibble: 1,750 x 6
##       id alt   choice    ic    oc income
##    <int> <fct> <lgl>  <dbl> <dbl>  <dbl>
##  1   133 ec    FALSE  -20.3 -4.52     60
##  2   133 ecc   FALSE  -25.5 -7.31     60
##  3   133 er    FALSE   -7.7 -4.32     60
##  4   133 erc   FALSE  -25.2 -7.11     60
##  5   133 gc    FALSE  -25.3 -2.26     60
##  6   133 gcc   TRUE   -23.3 -5.05     60
##  7   133 hpc   FALSE  -28.1 -4.42     60
##  8    14 ec    FALSE  -25.6 -5.21     50
##  9    14 ecc   FALSE  -29.2 -7.76     50
## 10    14 er    FALSE   -9.3 -3.8      50
## # ... with 1,740 more rows
```

Mixed Logit with Log-Normal Distributions

```r
### Model HVAC choice as a mixed logit
## Model choice using alternative intercepts and cost data with
## log-normal coefficients
model_2 <- hvac_mlogit_neg %>%
  mlogit(formula = choice ~ ic + oc | 1 | 0, data = ., reflevel = 'hpc',
         rpar = c(ic = 'ln', oc = 'ln'), R = 1000, seed = 321)
```

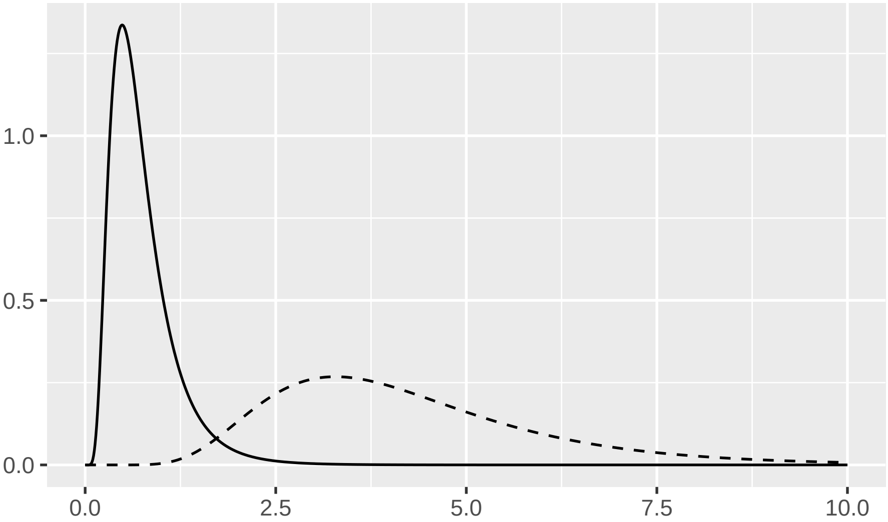
Model Results with Log-Normal Distributions

```
## Summarize model results
model_2 %>%
  summary()
##
## Call:
## mlogit(formula = choice ~ ic + oc | 1 | 0, data = ., reflevel = "hpc",
##     rpar = c(ic = "ln", oc = "ln"), R = 1000, seed = 321)
##
## Frequencies of alternatives:
##   hpc    ec   ecc    er   erc    gc   gcc
## 0.104 0.004 0.016 0.032 0.004 0.096 0.744
##
## bfgs method
## 40 iterations, 0h:0m:35s
## g'(-H)^-1g = 6.71E-08
## gradient close to zero
##
## Coefficients :
##                  Estimate Std. Error z-value  Pr(>|z|)
## ec:(intercept)  -17.93432    9.93335 -1.8055 0.0710020 .
## ecc:(intercept)   4.72345    2.07762  2.2735 0.0229964 *
## er:(intercept)  -39.61551   27.17697 -1.4577 0.1449269
## erc:(intercept)   3.43598    2.51989  1.3635 0.1727113
## gc:(intercept)  -19.91239    9.68510 -2.0560 0.0397843 *
## gcc:(intercept)   4.60527    1.39448  3.3025 0.0009583 ***
## ic               -0.43815    0.45590 -0.9611 0.3365171
## oc                1.36185    0.36511  3.7300 0.0001915 ***
## sd.ic             0.53322    0.25379  2.1010 0.0356401 *
## sd.oc             0.41586    0.13312  3.1240 0.0017839 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -189.72
## McFadden R^2:  0.15761
```

Log-Normally Distributed Coefficients

```r
## Plot distributions of random coefficients
ggplot(data = data.frame(x = c(0, 10)), aes(x)) +
  stat_function(fun = dlnorm, n = 1001,
                args = list(mean = model_2$coefficients[7],
                            sd = abs(model_2$coefficients[9]))) +
  stat_function(fun = dlnorm, n = 1001,
                args = list(mean = model_2$coefficients[8],
                            sd = abs(model_2$coefficients[10])),
                linetype = 'dashed') +
  xlab(NULL) +
  ylab(NULL)
```

Log-Normally Distributed Coefficients Plot



Mixed Logit with Correlated Log-Normal Distributions

```r
### Model HVAC choice as a mixed logit
## Model choice using alternative intercepts and cost data with
## correlated log-normal coefficients
model_3 <- hvac_mlogit_neg %>%
  mlogit(formula = choice ~ ic + oc | 1 | 0, data = ., reflevel = 'hpc',
         rpar = c(ic = 'ln', oc = 'ln'), correlation = TRUE, R = 1000,
         seed = 321)
```

Model Results with Correlated Log-Normal Distributions

```
## Summarize model results
model_3 %>%
  summary()
##
## Call:
## mlogit(formula = choice ~ ic + oc | 1 | 0, data = ., reflevel = "hpc",
##     rpar = c(ic = "ln", oc = "ln"), R = 1000, correlation = TRUE,
##     seed = 321)
##
## Frequencies of alternatives:
##   hpc    ec   ecc    er   erc    gc   gcc
## 0.104 0.004 0.016 0.032 0.004 0.096 0.744
##
## bfgs method
## 34 iterations, 0h:0m:34s
## g'(-H)^-1g = 6.52E-07
## gradient close to zero
##
## Coefficients :
##                  Estimate Std. Error z-value  Pr(>|z|)
## ec:(intercept)  -18.13141    6.29638 -2.8797  0.003981 **
## ecc:(intercept)   3.64970    1.45116  2.5150  0.011902 *
## er:(intercept)  -38.22791   16.94682 -2.2558  0.024086 *
## erc:(intercept)   2.43339    1.96325  1.2395  0.215172
## gc:(intercept)  -21.53875    7.32033 -2.9423  0.003258 **
## gcc:(intercept)   3.79386    0.88938  4.2657 1.992e-05 ***
## ic               -0.45353    0.30527 -1.4857  0.137368
## oc                1.26077    0.26512  4.7555 1.979e-06 ***
## chol.ic:ic        0.56759    0.24467  2.3198  0.020350 *
## chol.ic:oc        0.39855    0.19520  2.0418  0.041174 *
## chol.oc:oc        0.22016    0.23527  0.9358  0.349388
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Choleski Transformation for Multivariate Normals

One way to draw multivariate (log-)normal random variables is to draw independent normal random variables and apply a Choleski transformation

We want β_1 and β_2 to be multivariate log-normal

$$\ln \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{\beta_1} \\ \mu_{\beta_2} \end{pmatrix}, \Omega \right)$$

We draw two standard normal random variables, ω_1 and ω_2, and apply

$$\ln \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \mu_{\beta_1} \\ \mu_{\beta_2} \end{pmatrix} + \begin{pmatrix} s_{11} & 0 \\ s_{21} & s_{22} \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}$$

The elements of the variance-covariance matrix, Ω, are

$$\begin{aligned} Var(\ln \beta_1) &= s_{11}^2 \\ Var(\ln \beta_2) &= s_{21}^2 + s_{22}^2 \\ Cov(\ln \beta_1, \ln \beta_2) &= s_{11} s_{21} \end{aligned}$$

Variance and Covariance of Log-Normal Coefficients

```
## Calculate coefficient variances and covariance
model_3_vcov <- c(model_3$coefficients[9]^2,
                  model_3$coefficients[10]^2 +
                    model_3$coefficients[11]^2,
                  model_3$coefficients[9] *
                    model_3$coefficients[10]) %>%
  setNames(c('var.ic', 'var.oc', 'cov.ic:oc'))
model_3_vcov
##    var.ic    var.oc cov.ic:oc
## 0.3221626 0.2073116 0.2262136

## Calculate coefficient variances and covariance using vcov
vcov(model_3, what = 'rpar') %>%
  summary()
##           Estimate Std. Error z-value Pr(>|z|)
## var.ic     0.32216    0.27775  1.1599   0.2461
## var.oc     0.20731    0.16139  1.2845   0.1990
## cov.ic:oc  0.22621    0.18578  1.2177   0.2234
```

Implied Discount Rate

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

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

Divide by β_1 to express in terms of present day dollars

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

The net present value of a stream of future payments in perpetuity is

$$NPV_{ni} = \frac{1}{r} OC_{ni}$$

The last two equations give that

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

Distribution of Implied Discount Rate

If β_1 and β_2 are distributed log-normal with

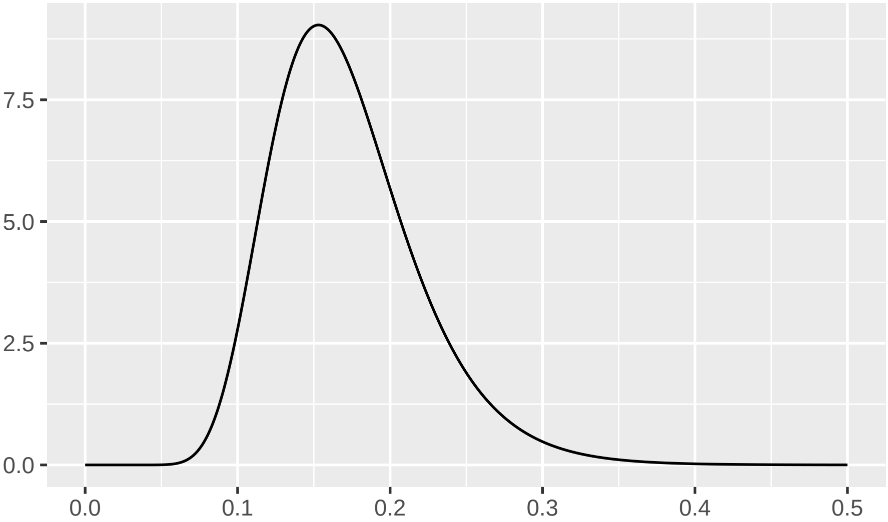
$$\ln \beta_1 \sim N(\mu_{\beta_1}, \sigma^2_{\beta_1})$$
$$\ln \beta_2 \sim N(\mu_{\beta_2}, \sigma^2_{\beta_2})$$

then the ratio is distributed log-normal with

$$\ln \frac{\beta_1}{\beta_2} \sim N(\mu_{\beta_1} - \mu_{\beta_2}, \sigma^2_{\beta_1} + \sigma^2_{\beta_2} - 2\sigma_{\beta_1\beta_2})$$

```r
## Plot distribution of implied discount rate
ggplot(data = data.frame(x = c(0, 0.5)), aes(x)) +
  stat_function(fun = dlnorm, n = 1001,
                args = list(mean = model_2$coefficients[7] -
                              model_2$coefficients[8],
                            sd = sqrt(model_3_vcov[1] + model_3_vcov[2] -
                                        2 * model_3_vcov[3]))) +
  xlab(NULL) +
  ylab(NULL)
```

Implied Discount Rate Plot



Announcements

Reading for next time

- Train textbook, Chapter 10

Upcoming

- Problem Set 4 will be posted soon, due November 21