Data Appendix to "The role of homeownership in transit use in the Allston-Brighton neighborhood of Boston"

Victoria Tse

Contents


```
## Warning: package 'summarytools' was built under R version 3.6.3
```

1 Appendix description

This Data Appendix documents the data used in "The role of homeownership in transit use in the Allston-Brighton neighborhood of Boston". It was prepared in a Rmarkdown document that contains both the documentation and the R code used to prepare the data used in the final estimation. It also includes descriptive statistics for both the original data and the final dataset, with a discussion of any issues of note.

The datasets used directly by the final analysis are saved in `processed-data/` at the end of this file.

2 Raw data

Each dataset you use will have its own documentation section. The next subsection in this document (Dataset description) is a template. You can copy this section and paste it into your document each time you need to add a section for a new dataset. Note that each line in the Dataset description section ***must*** *end with two spaces.* This section documents the datasets used in this analysis.

2.1 Dataset description

Citation:

Massachusetts Department of Transportation. (2012, June). Massachusetts Travel Survey.

DOI:

N/A

Date Downloaded:

3/7/20, from files provided by Susan Sayre

Filename(s):

- HH.txt
- PER.txt
- PLACE.txt

Unit of observation:

Each unit is an individual person

Dates covered:

6/14/10 - 10/24/11

2.1.1 To obtain a copy

To obtain a copy, be a part of the Smith College institution and ask Professor Sayre! Also, agree to the necessary privacy measures to ensure the safety of the data. If not a part of Smith College as an institution, contact MassDOT.

2.1.2 Importable version (if necessary)

N/A

Filename(s):

- raw_data/filename.csv

- HH.txt
- PER.txt
- PLACE.txt
- VEH.txt

2.1.3 Variable descriptions

- **RIBUS:** Use transit on a regular basis during the week

  1 - Yes

  0- No

- **HHVEH:** Number of household vehicles in working condition

  Range 1-7 8 - 8 or more

- **HISP:** Hipanic or Latino

  1- Yes

  0- No

- **RACE:** Ethnicity or race

  1 White alone

  2 Black or African American alone

  3 Indigenous/ Alaska Native alone

  4 Asian alone

  5 Native Hawaiian/ Pacific Islander alone

  6 Some other race alone

  7 Two or more races

- **OWN:** Homeownership

  1 Own/Mortgaged

  0 Renter

- **INCOME:** Household Income in 2009

  1 Less than $15,000

  2 $15,00-$24,999

  3 $25,000-$34,999

  4 $35,000- $49,999

  5 $50,000-$74,999

  6 $75,000-$99,999

  7 $100,00- $149,999

  8 $150,000 or more

- **AGE:** Age in years

- **LIC:** Valid Driver's License

  1- Yes

  2- No

- **TRANS:** Have a transit pass?

  1- Yes

  0- No

- **EMPLY:** Employed

  1- Yes

  0- No

- **EDUCA:** Level of Education

  1- Not a high school graduate, 12th or less (Includes very young children)

  2-High school diploma or GED

  3- Some college credit but no degree

  4- Associate or technical school degree

  5- Bachelor's or undergraduate degree

  6- Graduate Degree (Includes professional degrees, MD, DDs, JD)

- **ENROL:** Student

  1- Yes

  0- No

- **BIKFA:** Bike facilities available at work/school? [ENROL = 1, WORK=1]

  1- Yes

  0- No

2.1.4 Data import code and summary

Once you've described the variables, enter an R chunk by selecting Code -> Insert Chunk, or Ctrl+Alt+I, give it a name to describe the dataset you are importing. After importing, export a dataframe summary using the command.

```
hh_data <- read_csv(file.path("raw-data","mts-survey-data","HH.txt"), guess_max = 15000)
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   MPO = col_character(),
##   O_RESTY = col_character(),
##   O_OWN = col_character(),
##   HCITY = col_character(),
##   HSTATE = col_character(),
##   HZIP = col_character(),
##   HTOWN = col_character(),
##   HPUMA10 = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
per_data <- read_csv(file.path("raw-data","mts-survey-data","PER.txt"), guess_max = 15000)
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   O_PASST = col_character(),
##   O_WKSTAT = col_character(),
##   O_WMODE = col_character(),
##   WTRSB = col_character(),
##   O_FLEXP = col_character(),
##   O_EDUCA = col_character(),
##   O_SCHOL = col_character(),
##   O_SMODE = col_character(),
##   STRSB = col_character(),
##   O_TYPDY = col_character(),
##   O_TYPPL = col_character(),
##   O_NOGO = col_character(),
##   WCITY = col_character(),
##   WSTATE = col_character(),
##   WZIP = col_character(),
##   WTOWN = col_character(),
##   WPUMA10 = col_character(),
##   SCITY = col_character(),
##   SSTATE = col_character(),
```

```
##   SZIP = col_character()
##   # ... with 2 more columns
## )

## See spec(...) for full column specifications.
```

```
place_data <- read_csv(file.path("raw-data","mts-survey-data","PLACE.txt"), guess_max = 15000)
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   O_TPURP = col_character(),
##   O_TPUR2 = col_logical(),
##   O_MODE = col_character(),
##   O_PRKTY = col_character(),
##   PRKLC = col_character(),
##   HOV = col_character(),
##   O_FARE = col_character(),
##   ROUTE = col_character(),
##   CITY = col_character(),
##   STATE = col_character(),
##   ZIP = col_character(),
##   TOWN = col_character(),
##   PUMA10 = col_character()
## )

## See spec(...) for full column specifications.

## Warning: 46 parsing failures.
##   row     col           expected       actual                                 file
## 36226 O_TPUR2 1/0/T/F/TRUE/FALSE VOLUNTEERING 'raw-data/mts-survey-data/PLACE.txt'
## 42538 O_PRKCS a double           DK/RF        'raw-data/mts-survey-data/PLACE.txt'
## 43711 O_PRKCS a double           DK/RF        'raw-data/mts-survey-data/PLACE.txt'
## 57718 O_PRKCS a double           DK/RF        'raw-data/mts-survey-data/PLACE.txt'
## 61029 O_PRKCS a double           DK/RF        'raw-data/mts-survey-data/PLACE.txt'
## ..... ....... .................. ............ ....................................
## See problems(...) for more details.
```

```
export_summary_table(dfSummary(dataset_name))
```

While it will make your resulting file long, you should not modify the chunk options to suppress printing of code and output. I would likely not include this in the documentation for an actual paper I was submitting, but including them here will let me read your code and the output message from R and may help identify data import concerns early in the process. Since these files will exist only electronically, their length is less of a concern. If you like to print out files to proofread and want me to help you shorten the printed versions, let me know. We can temporarily modify the chunk options for printing and restore them before you submit the assignment.

```
# Data Processing and Combination
*This section should include a discussion of the processing and merging steps needed to create your bas
```

```r
Dummy_1 <- merge(hh_data, per_data,
                              by = "SAMPN")
```

```r
Preliminary_data <- merge(Dummy_1, place_data,
                 by ="SAMPN")
```

```r
  Final_data_half1 <- Preliminary_data[ which(Preliminary_data$ZIP=='02134'), ]
```

```r
  Final_data_half2 <- Preliminary_data[ which(Preliminary_data$ZIP=='02135'), ]
```

```r
Final_data <- rbind(Final_data_half1, Final_data_half2)
```

```r
Modified_final_data <- Final_data %>%

mutate(RIBUS = case_when(RIBUS== 1 ~ 1,
                         RIBUS== 2 ~ 0),
       HISP = case_when(HISP== 1 ~ 1,
                        HISP == 2 ~ 0),
       OWN = case_when(OWN== 1 ~ 1,
                       OWN== 2 ~ 0),
       LIC = case_when(LIC== 1 ~ 1,
                       LIC== 2 ~ 0),
       TRANS = case_when(TRANS== 1 ~ 1,
                         TRANS== 2~ 0),
       EMPLY = case_when(EMPLY== 1 ~ 1,
                         EMPLY== 2 ~ 0),
       ENROL = case_when(ENROL == 1 ~ 1,
                         ENROL== 2 ~ 0),
       BIKFA = case_when(BIKFA == 1 ~ 1,
                         BIKFA == 2 ~ 0),
       RACE_factor = as.factor(case_when(RACE == 1 ~ "White",
                              RACE== 2 ~ "Black",
                              RACE== 3 ~ "Indigenous",
                              RACE==  4 ~ "Asian",
                              RACE== 5 ~ "Pacific Islander",
                              RACE== 6 ~ "Other",
                              RACE== 7 ~ "Multiracial")),
      INCOME_factor = as.factor(case_when(INCOME == 1 ~ "15",
                                INCOME== 2 ~ "15-24",
                                INCOME== 3 ~ "25-34",
                                INCOME==  4 ~ "35-49",
                                INCOME== 5 ~ "50-74",
                                INCOME== 6 ~ "75-99",
                                INCOME== 7 ~ "100-149",
                                INCOME == 8 ~ "150" )),
  EDUCA_factor = as.factor(case_when(EDUCA== 1 ~ "K-12",
                              EDUCA== 2 ~ "HS-GED",
                              EDUCA== 3 ~ "Some_college",
                              EDUCA==  4 ~ "Associates",
                              EDUCA== 5 ~ "Bachelors",
                              EDUCA== 6 ~ "Graduate" )))
```

```r
Summary_statistics <- Modified_final_data %>%
  select(RIBUS, OWN, HHVEH, LIC, TRANS, BIKFA, RACE, HISP, INCOME, EDUCA, ENROL, EMPLY, AGE)

stargazer(Summary_statistics, type = "latex",
          title = "Summary Statistics",
          header = F,
          digits = 2)
```

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
RIBUS	4,206	0.66	0.47	0	0	1	1
OWN	4,182	0.66	0.47	0.00	0.00	1.00	1.00
HHVEH	4,206	1.36	0.95	0	1	2	5
LIC	3,132	0.81	0.39	0.00	1.00	1.00	1.00
TRANS	3,113	0.59	0.49	0.00	0.00	1.00	1.00
BIKFA	2,781	0.54	0.50	0.00	0.00	1.00	1.00
RACE	4,206	1.84	1.95	1	1	1	9
HISP	4,194	0.06	0.24	0.00	0.00	0.00	1.00
INCOME	4,206	13.42	25.73	1	5	7	99
EDUCA	4,206	3.77	2.16	1	1	6	9
ENROL	4,182	0.29	0.46	0.00	0.00	1.00	1.00
EMPLY	3,138	0.73	0.44	0.00	0.00	1.00	1.00
AGE	4,206	36.98	23.93	0	15	54	99

```r
Regression_1 <- lm(RIBUS ~ OWN, data = Modified_final_data)

stargazer(Regression_1, type = "latex",
          intercept.bottom = F,
          header = F)
```

Table 2:

	Dependent variable:
	RIBUS
Constant	0.802^{***}
	(0.012)
OWN	-0.211^{***}
	(0.015)
Observations	4,182
R^2	0.044
Adjusted R^2	0.044
Residual Std. Error	0.462 (df = 4180)
F Statistic	194.168^{***} (df = 1; 4180)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
Regression_2 <- lm(RIBUS ~ OWN + HHVEH + LIC + TRANS + BIKFA, data = Modified_final_data)


stargazer(Regression_2, type = "latex",
          intercept.bottom = F,
          header = F)
```

Table 3:

	Dependent variable:
	RIBUS
Constant	0.735^{***}
	(0.032)
OWN	-0.182^{***}
	(0.022)
HHVEH	-0.065^{***}
	(0.012)
LIC	-0.003
	(0.025)
TRANS	0.242^{***}
	(0.021)
BIKFA	-0.033^{*}
	(0.019)
Observations	2,042
R^2	0.194
Adjusted R^2	0.192
Residual Std. Error	0.427 (df = 2036)
F Statistic	98.170^{***} (df = 5; 2036)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
Regression_3 <- lm(RIBUS ~ OWN + factor(RACE) + HISP, data = Modified_final_data)

stargazer(Regression_3, type = "latex",
          intercept.bottom = F,
          header = F)
```

```
Regression_4 <- lm(RIBUS ~  OWN + factor(EDUCA), data = Modified_final_data)

stargazer(Regression_4, type = "latex",
          intercept.bottom = F,
          header = F)
```

Table 4:

	Dependent variable:
	RIBUS
Constant	0.819^{***}
	(0.014)
OWN	-0.241^{***}
	(0.016)
factor(RACE)2	0.225^{***}
	(0.039)
factor(RACE)4	-0.117^{***}
	(0.030)
factor(RACE)6	0.251^{***}
	(0.041)
factor(RACE)7	-0.014
	(0.044)
factor(RACE)9	0.296^{***}
	(0.046)
HISP	-0.259^{***}
	(0.037)
Observations	4,170
R^2	0.078
Adjusted R^2	0.077
Residual Std. Error	0.455 (df = 4162)
F Statistic	50.471^{***} (df = 7; 4162)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5:

	Dependent variable:
	RIBUS
Constant	0.814^{***}
	(0.017)
OWN	-0.217^{***}
	(0.015)
factor(EDUCA)2	0.072^{**}
	(0.029)
factor(EDUCA)3	0.028
	(0.032)
factor(EDUCA)4	-0.093^{**}
	(0.041)
factor(EDUCA)5	-0.089^{***}
	(0.020)
factor(EDUCA)6	0.029
	(0.018)
factor(EDUCA)8	0.349
	(0.230)
factor(EDUCA)9	-0.121^{*}
	(0.072)
Observations	4,182
R^2	0.058
Adjusted R^2	0.056
Residual Std. Error	0.460 (df = 4173)
F Statistic	31.873^{***} (df = 8; 4173)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
Regression_5 <- lm(RIBUS ~  OWN + factor(INCOME), data = Modified_final_data)

stargazer(Regression_5, type = "latex",
          intercept.bottom = F,
          header = F)
```

Table 6:

	Dependent variable:
	RIBUS
Constant	0.795^{***}
	(0.036)
OWN	-0.196^{***}
	(0.018)
factor(INCOME)2	-0.105^{**}
	(0.047)
factor(INCOME)3	-0.001
	(0.054)
factor(INCOME)4	-0.004
	(0.047)
factor(INCOME)5	0.156^{***}
	(0.041)
factor(INCOME)6	-0.014
	(0.040)
factor(INCOME)7	0.034
	(0.041)
factor(INCOME)8	-0.117^{***}
	(0.043)
factor(INCOME)99	-0.093^{**}
	(0.044)
Observations	4,182
R^2	0.074
Adjusted R^2	0.072
Residual Std. Error	0.456 (df = 4172)
F Statistic	37.028^{***} (df = 9; 4172)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
Regression_6 <- lm(RIBUS ~  OWN + EMPLY + ENROL, data = Modified_final_data)

stargazer(Regression_6, type = "latex",
```

```
          intercept.bottom = F,
          header = F)
```

Table 7:

	Dependent variable:
	RIBUS
Constant	0.806^{***}
	(0.019)
OWN	-0.239^{***}
	(0.017)
EMPLY	-0.0004
	(0.019)
ENROL	0.073^{***}
	(0.025)
Observations	3,098
R^2	0.067
Adjusted R^2	0.066
Residual Std. Error	0.455 (df = 3094)
F Statistic	74.339^{***} (df = 3; 3094)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
Regression_8 <- lm(RIBUS ~ OWN + AGE, data = Modified_final_data)

stargazer(Regression_8, type = "latex",
          intercept.bottom = F,
          header = F)
```

```
  Homeowner_data <- Modified_final_data[ which(Modified_final_data$OWN=='1'), ]
```

```
  Renter_data <- Modified_final_data[ which(Modified_final_data$OWN=='0'), ]
```

```
Summary_statistics <- Homeowner_data %>%
  select(RIBUS, OWN, HHVEH, LIC, TRANS, BIKFA, RACE, HISP, INCOME, EDUCA, ENROL, EMPLY, AGE)

stargazer(Summary_statistics, type = "latex",
          title = "Homeowner Summary Statistics",
          header = F,
          digits = 2)
```

```
Summary_statistics <- Renter_data %>%
  select(RIBUS, OWN, HHVEH, LIC, TRANS, BIKFA, RACE, HISP, INCOME, EDUCA, ENROL, EMPLY, AGE)

stargazer(Summary_statistics, type = "latex",
```

Table 8:

	Dependent variable:
	RIBUS
Constant	0.852^{***}
	(0.016)
OWN	-0.206^{***}
	(0.015)
AGE	-0.001^{***}
	(0.0003)
Observations	4,182
R^2	0.050
Adjusted R^2	0.049
Residual Std. Error	0.461 (df = 4179)
F Statistic	109.176^{***} (df = 2; 4179)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 9: Homeowner Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
RIBUS	2,769	0.59	0.49	0	0	1	1
OWN	2,769	1.00	0.00	1	1	1	1
HHVEH	2,769	1.67	0.89	0	1	2	5
LIC	1,918	0.90	0.30	0.00	1.00	1.00	1.00
TRANS	1,918	0.52	0.50	0.00	0.00	1.00	1.00
BIKFA	1,830	0.53	0.50	0.00	0.00	1.00	1.00
RACE	2,769	1.72	1.94	1	1	1	9
HISP	2,757	0.01	0.09	0.00	0.00	0.00	1.00
INCOME	2,769	13.69	24.92	1	6	8	99
EDUCA	2,769	3.68	2.25	1	1	6	9
ENROL	2,754	0.31	0.46	0.00	0.00	1.00	1.00
EMPLY	1,924	0.75	0.43	0.00	1.00	1.00	1.00
AGE	2,769	38.09	26.12	0	12	56	99

```
          title = "Renter Summary Statistics",
          header = F,
          digits = 2)
```

Table 10: Renter Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
RIBUS	1,413	0.80	0.40	0	1	1	1
OWN	1,413	0.00	0.00	0	0	0	0
HHVEH	1,413	0.74	0.76	0	0	1	3
LIC	1,198	0.67	0.47	0.00	0.00	1.00	1.00
TRANS	1,179	0.72	0.45	0.00	0.00	1.00	1.00
BIKFA	927	0.58	0.49	0.00	0.00	1.00	1.00
RACE	1,413	2.04	1.96	1	1	2	7
HISP	1,413	0.16	0.37	0	0	0	1
INCOME	1,413	13.06	27.43	1	3	6	99
EDUCA	1,413	3.98	1.95	1	2	6	8
ENROL	1,404	0.27	0.45	0.00	0.00	1.00	1.00
EMPLY	1,198	0.69	0.46	0.00	0.00	1.00	1.00
AGE	1,413	34.92	18.84	0	23	48	99

3 Analysis Variables

In depth descriptions of the variables and models can be found in the paper. The variables I adjusted were RIBUS, OWN, TRANS, LIC, BIKFA, HISP, EMPLY, ENROL. Each of them expressed "no" with a "2", which I adjusted to a "0".

4 Discussion of Data

Some of the data may be slightly skewed by anomalous variables. For example, INCOME, despite having a 1-8 range, has a mean of 13.42 due to a 99 outcome for "Don't Know".