P5 code

My focus for this assignment

Data dictionary: https://nhts.ornl.gov/tables09/CodebookBrowser.aspx (https://nhts.ornl.gov/tables09/CodebookBrowser.aspx)

Filter for a particular trip purpose and/or by a particular population → predict the likelihood that a short trip (< 1.5mi) will take place by walking (mode choice)

- **Research question:** are adults (18-65) living in urban areas more likely to walk for short trips (< 1.5mi) to run errands compared to adults (18-65) living in rural areas?
- Filter the population to only include people ages 18-65
- Filter the WHYTO or WHYFROM variables to include the trip purposes of interest
    - 11: buy goods (groceries, clothes, appliances, gas)
    - 12: buy services (dry cleaners, banking, service a car, pet care)
    - 13: buy meals (go out for a meal, snack, takeout)
    - 14: other general errands (post office, library)
- Predictors
    - Trip distance (TRPMILES)
    - Gender (R_SEX)
    - Age (R_AGE)
    - Urban/rural (URBRUR)
    - Population density of the trip destination (DBPPOPDN)

Setup

Load libraries

```
library(tidyverse)
library(here)
library(knitr)
library(srvyr)
library(tidycensus)
library(jtools)
```

Load data

```
trips <- here("P5_mode_choice", "trippub.csv") |>
  read_csv(show_col_types = FALSE)
```

```
people <- here("P5_mode_choice", "perpub.csv") |>
  read_csv(show_col_types = FALSE)
```

Data cleaning

Filter the trip purposes

```
# Filter the dataset to only include the trip purposes I'm interested in studying (11, 12, 13, 14)
trip_purposes <- c(11, 12, 13, 14)

visit_trips <- trips |>
  filter(WHYTO %in% trip_purposes |
           WHYFROM %in% trip_purposes)
```

Filter the population

```
# Filter the dataset to only include adults age 18-65
sr_visit_trips <- visit_trips |>
  filter(R_AGE >= 18 & R_AGE <= 65)
```

Filter for short distance trips

```
short_sr_visit_trips <- sr_visit_trips |>
  filter(TRPMILAD < 1.5)

nrow(short_sr_visit_trips)
```

```
## [1] 79511
```

We have a sample of 79,511 trips after all the filtering

Generate the outcome variable

```
# Create a binary outcome variable of whether or not the mode of transport was walking

short_sr_visit_trips <- short_sr_visit_trips |>
  mutate(walk = TRPTRANS == "01")

short_sr_visit_trips |>
  mutate(Mode = factor(ifelse(walk, "Walk", "Other mode"),
                       levels = c("Walk", "Other mode"))) |>
  group_by(Mode) |>
  summarise(`Number of trips` = n()) |>
  mutate(`Percent of trips` =
           paste0(round(100*`Number of trips`/sum(`Number of trips`)), "%")) |>
  kable()
```

Mode	Number of trips	Percent of trips
Walk	18548	23%
Other mode	60963	77%

Incorporate survey weights

We’re incorporating survey weights to re-estimate the percent of short trips to run errands for adults 18-65 that take place via walking vs. another mode

```
short_sr_visit_trips |>
  srvyr::as_survey_design(weights = WTTRDFIN) |>
  mutate(Mode = factor(ifelse(walk, "Walk", "Other mode"),
                       levels = c("Walk", "Other mode"))) |>
  group_by(Mode) |>
  srvyr::survey_tally(vartype = "ci") |>
  mutate(`Estimated percent of trips` =
           paste0(round(100*n/sum(n)),"%"),
         `Lower estimate (95% confidence)` =
           paste0(round(100*n_low/sum(n)),"%"),
         `Upper estimate (95% confidence)` =
           paste0(round(100*n_upp/sum(n)),"%")) |>
  select(Mode,
         `Estimated percent of trips`,
         `Lower estimate (95% confidence)`,
         `Upper estimate (95% confidence)`) |>
  kable()
```

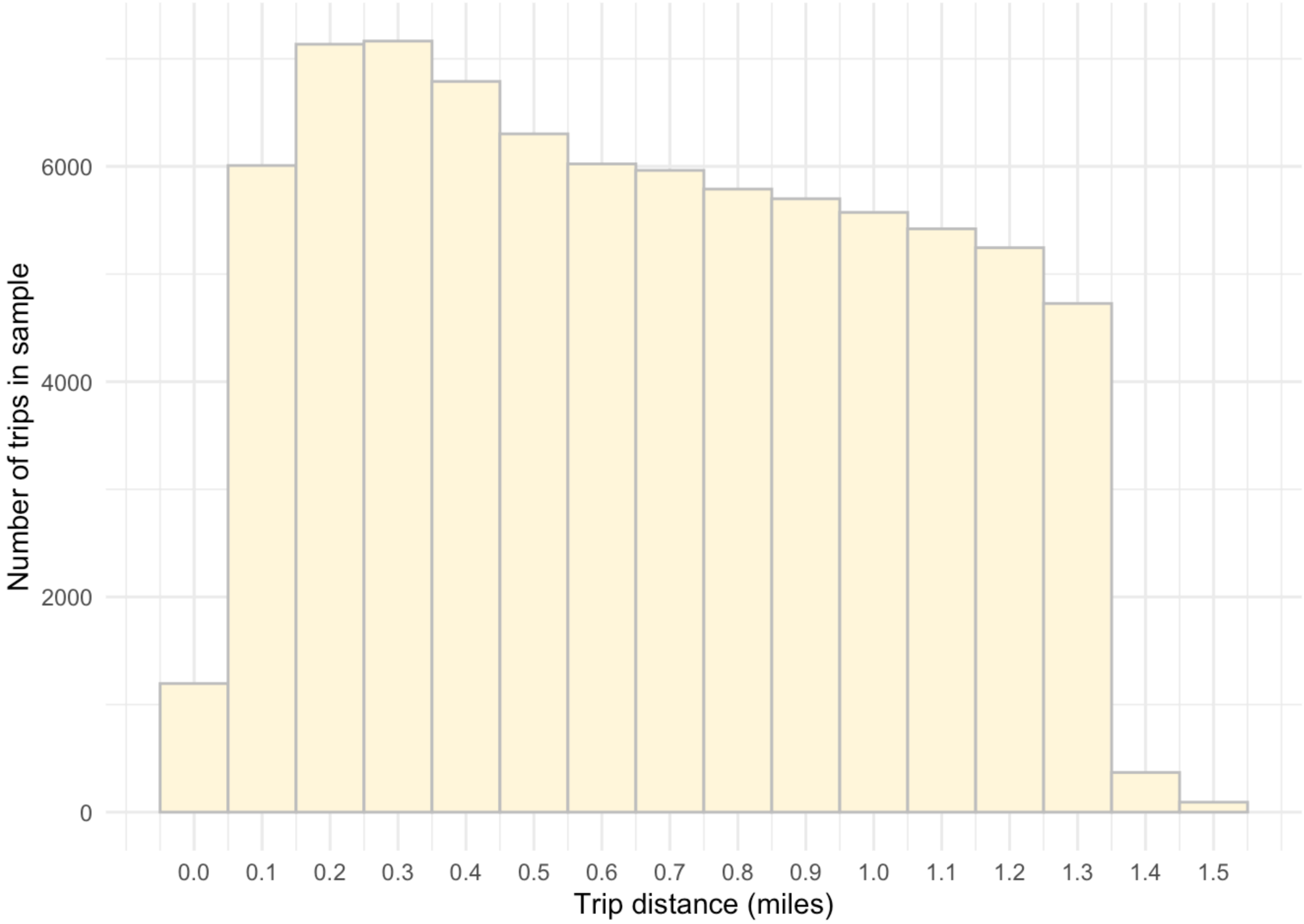
Mode	Estimated percent of trips	Lower estimate (95% confidence)	Upper estimate (95% confidence)
Walk	29%	28%	30%
Other mode	71%	70%	72%

Examining and cleaning predictors

Trip distance (TRPMILES)

```
# Remove NAs (coded as negative values)
sample_trips <- short_sr_visit_trips |>
  filter(TRPMILES >=0)

ggplot(sample_trips) +
  geom_histogram(aes(x = TRPMILES),
                 color = "gray",
                 fill = "cornsilk",
                 binwidth = 0.1) +
  scale_x_continuous(name = "Trip distance (miles)",
                     breaks = seq(0, 1.5, by=0.1)) +
  scale_y_continuous(name = "Number of trips in sample") +
  theme_minimal()
```



Gender (R_SEX)

```
# Remove NAs (coded as negative values)
# Male = 0 FALSE; Female = 1 TRUE
sample_trips <- sample_trips |>
  filter(R_SEX > 0) |>
  mutate(R_SEX_logical = R_SEX == "02")

table(sample_trips$R_SEX_logical)
```

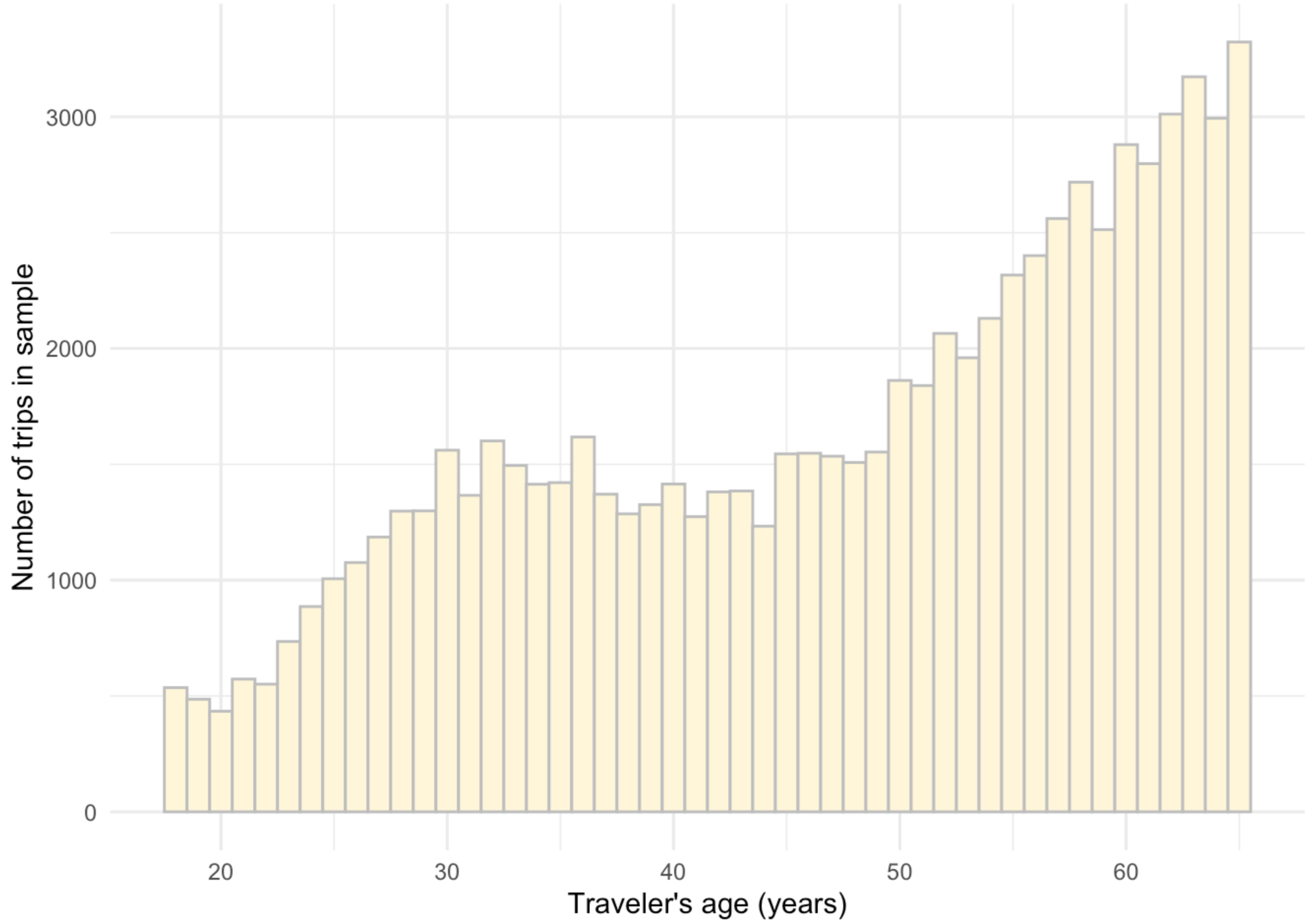
```
##
## FALSE  TRUE
## 34489 44960
```

```
prop.table(table(sample_trips$R_SEX_logical))
```

```
##
##     FALSE      TRUE
## 0.4341024 0.5658976
```

Age (R_AGE)

```
ggplot(sample_trips) +
  geom_histogram(aes(x = R_AGE),
                 color = "gray",
                 fill = "cornsilk",
                 binwidth = 1) +
  scale_x_continuous(name = "Traveler's age (years)",
                     breaks = seq(0, 100, by = 10)) +
  scale_y_continuous(name = "Number of trips in sample") +
  theme_minimal()
```



Urban/rural (URBRUR)

```
# Remove NAs (coded as negative values)
# Urban = 0 FALSE; Rural = 1 TRUE
sample_trips <- sample_trips |>
  filter(URBRUR > 0) |>
  mutate(URBRUR_logical = URBRUR == "02")

table(sample_trips$URBRUR_logical)
```

```
##
## FALSE  TRUE
## 66640 12809
```

```
prop.table(table(sample_trips$URBRUR_logical))
```

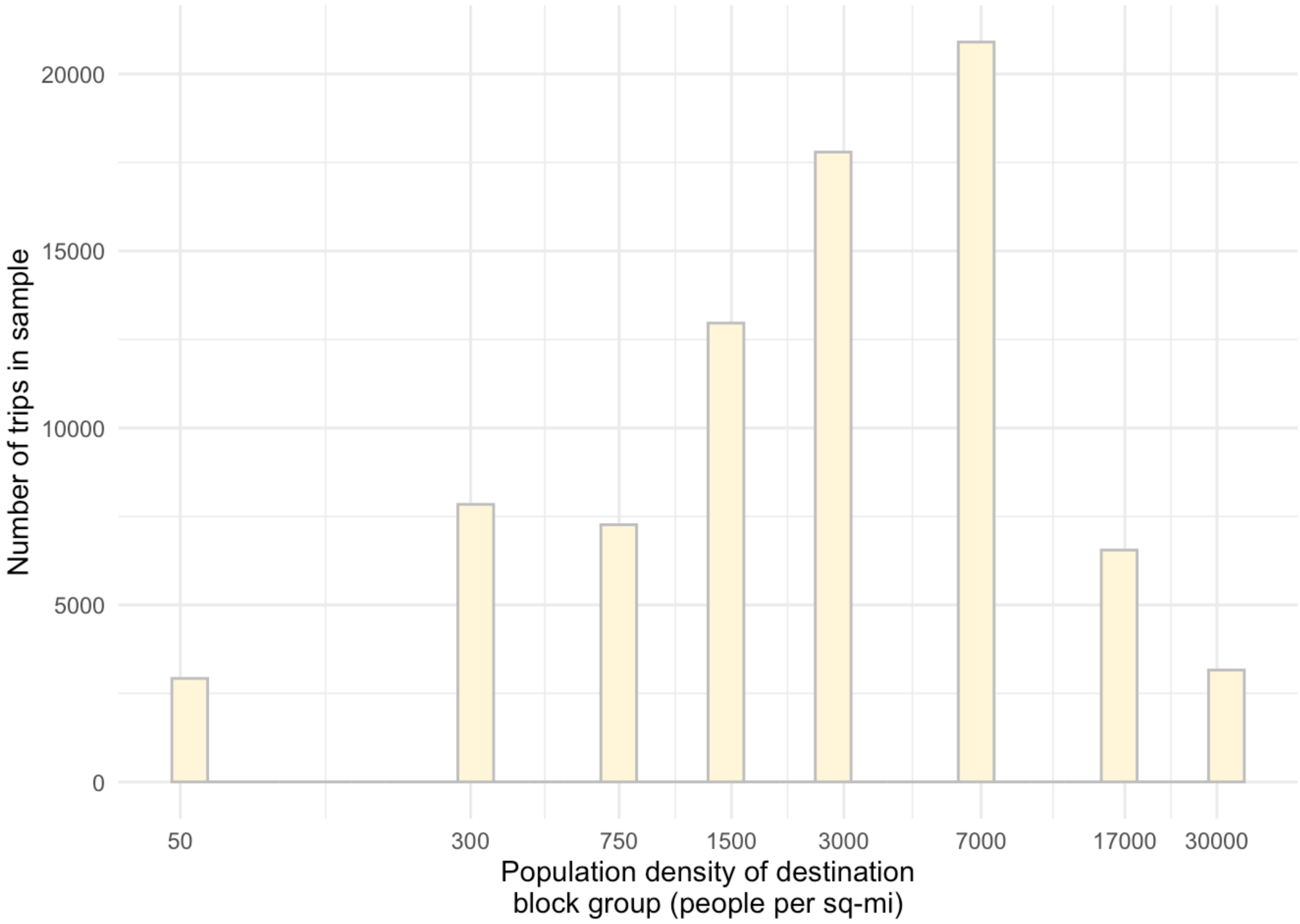
```
##
##     FALSE      TRUE
## 0.8387771 0.1612229
```

Population density of the trip destination (DBPPOPDN)

```
# Remove NAs (coded as negative values)
sample_trips <- sample_trips |>
  filter(DBPPOPDN > 0)

density_values <- c(50, 300, 750, 1500, 3000, 7000, 17000, 30000)

ggplot(sample_trips) +
  geom_histogram(aes(x = DBPPOPDN),
                 color = "gray",
                 fill = "cornsilk",
                 bins = 30) +
  scale_x_continuous(breaks = density_values,
                 labels = density_values,
                 transform = "log",
                 name = paste0("Population density of destination\n",
                              "block group (people per sq-mi)")) +
  scale_y_continuous(name = "Number of trips in sample") +
  theme_minimal()
```



Final sample size after removing NAs

```
nrow(sample_trips)
```

```
## [1] 79398
```

The final sample size is 79,398 trips after removing NAs

Logistic regression model

We will estimate a logistic regression model to predict the likelihood that a trip in our cleaned/filtered dataset will take place by walking.

```
model <- glm(walk ~
               TRPMILES +
               R_SEX_logical +
               R_AGE +
               URBRUR_logical +
               DBPPOPDN,
             data = sample_trips,
             family = "binomial")

coeff_labels <- c("Trip distance (miles)" = "TRPMILES",
                  "Sex (female)" = "R_SEX_logicalTRUE",
                  "Age (years)" = "R_AGE",
                  "Rural" = "URBRUR_logicalTRUE",
                  "Block-group population density at destination" = "DBPPOPDN")

export_summs(model,
             robust = "HC3",
             coefs = coeff_labels,
             error_format = "(p = {p.value})",
             error_pos = "right",
             digits = 4)
```

	Model 1	
Trip distance (miles)	-3.9965 ***	(p = 0.0000)
Sex (female)	-0.2343 ***	(p = 0.0000)
Age (years)	-0.0178 ***	(p = 0.0000)
Rural	-0.4781 ***	(p = 0.0000)
Block-group population density at destination	0.0001 ***	(p = 0.0000)
N	79398	
AIC	60300.4050	
BIC	60356.0984	

Pseudo R2	0.4214

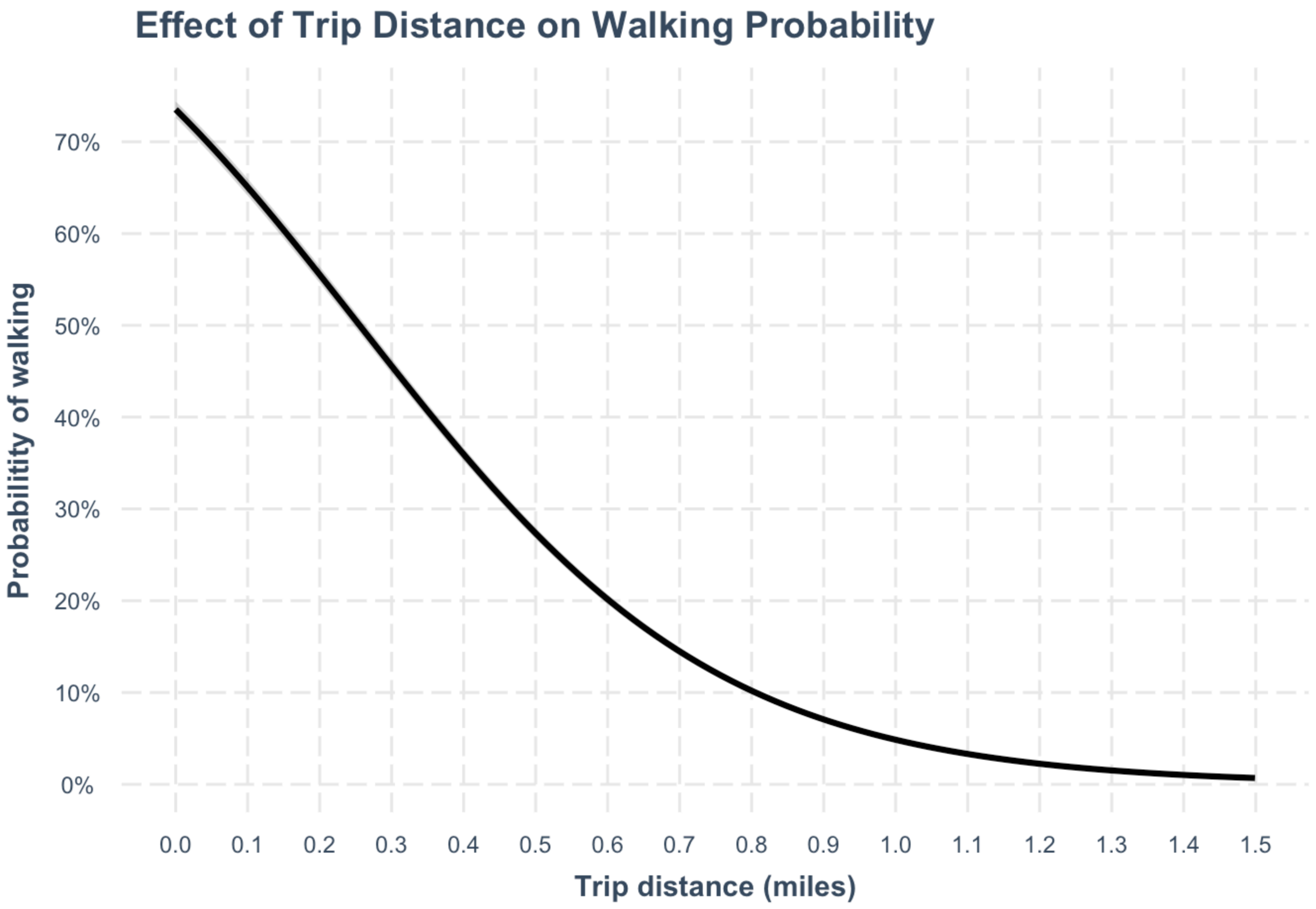
Standard errors are heteroskedasticity robust. *** p < 0.001; ** p < 0.01; * p < 0.05.

Visualizing the impact of each predictor

Trip distance (TRPMILES)

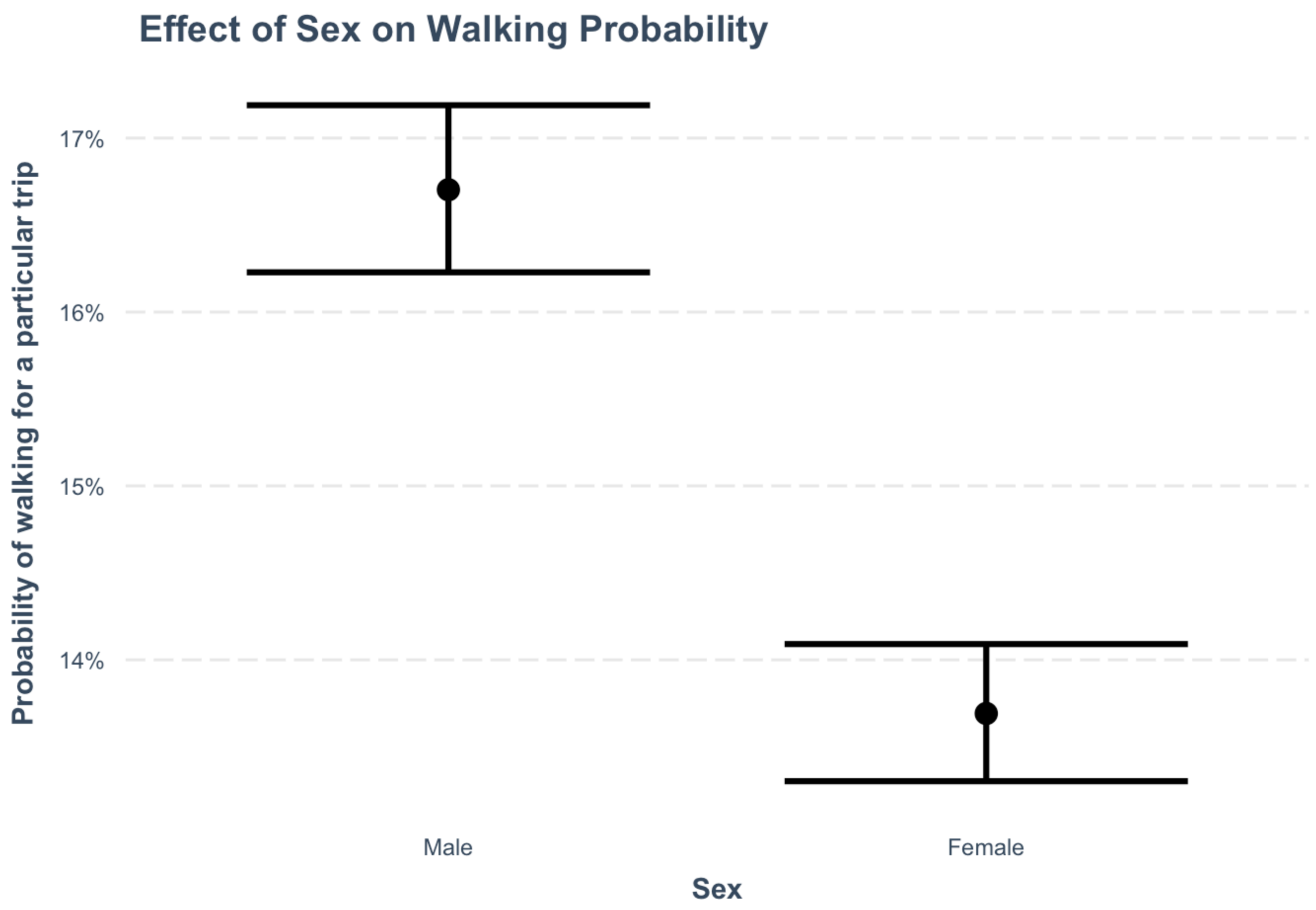
Effect plot of the predicted probabilities of walking across the range of trip distances, holding all other predictors at their average (if continuous) or reference (if categorical) values

```
effect_plot(model, pred = "TRPMILES", interval = TRUE) +
  scale_x_continuous(name = "Trip distance (miles)",
                     breaks = seq(0, 1.5, by  =0.1)) +
  scale_y_continuous(name = "Probabilitity of walking",
                     breaks = breaks <- seq(0, 1, by = 0.1),
                     labels = paste0(breaks*100, "%")) + ggtitle("Effect of Trip Distanc
e on Walking Probability")
```



Gender (R_SEX_logical)

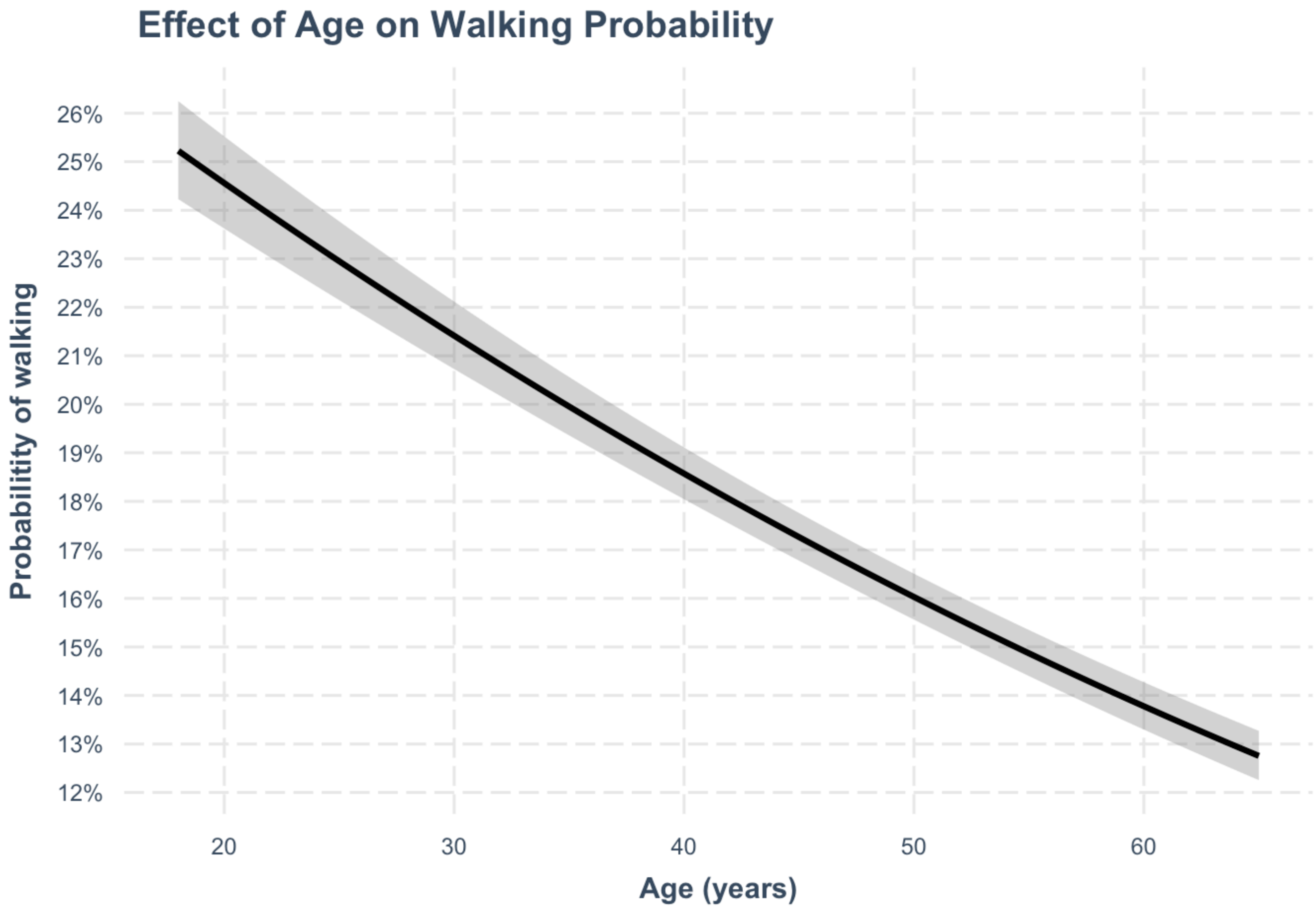
```
effect_plot(model = model, pred = "R_SEX_logical", interval = TRUE) +
  scale_y_continuous(name = "Probability of walking for a particular trip",
                     breaks = breaks <- seq(0, 1, by=0.01),
                     labels = paste0(breaks*100, "%")) +
  scale_x_discrete(name = paste0("Sex"),
                   labels = c("Male", "Female")) + ggtitle("Effect of Sex on Walking Pro
bability")
```



Age (R_AGE)

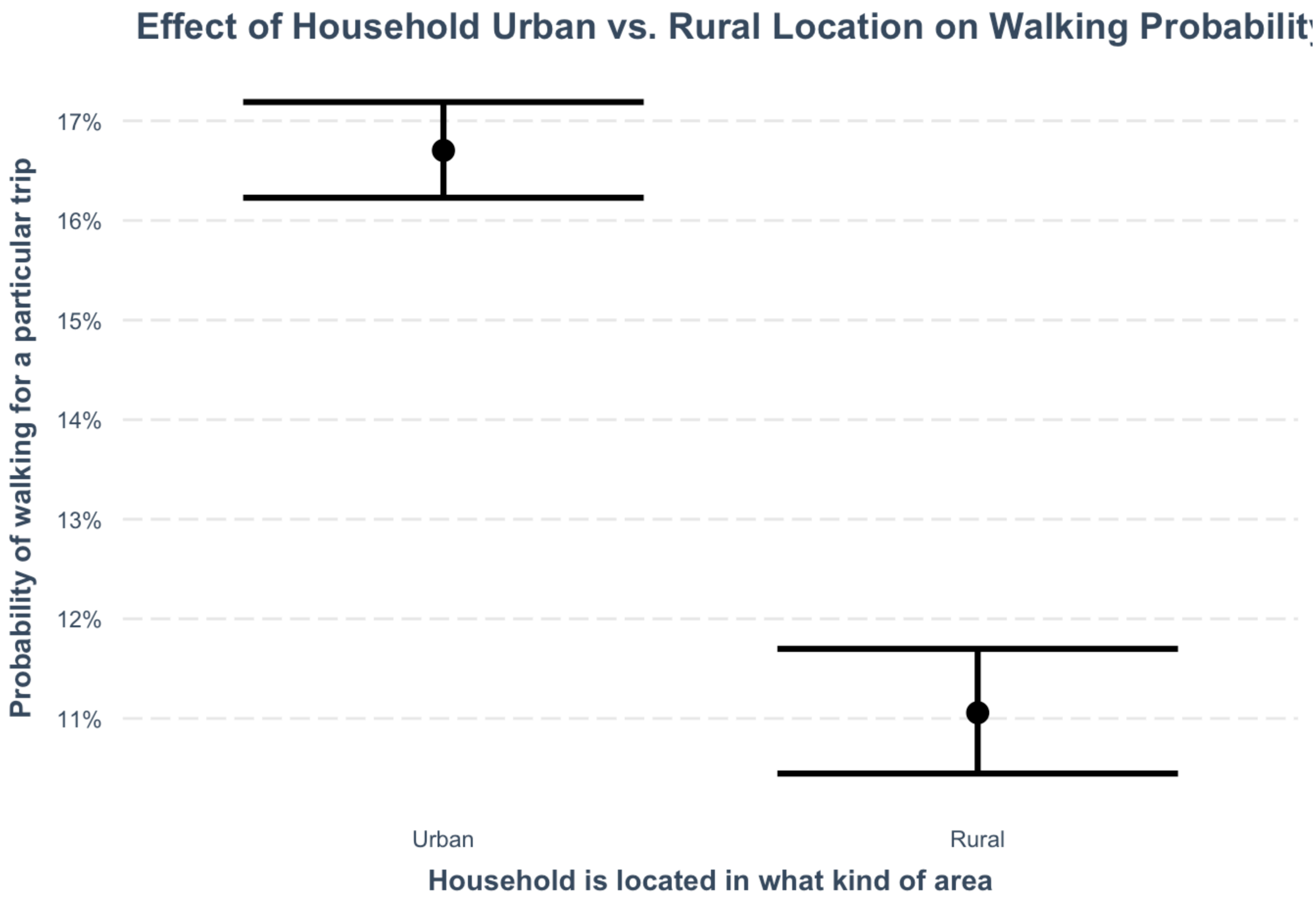
Effect plot of the predicted probabilities of walking across the range of ages, holding all other predictors at their average (if continuous) or reference (if categorical) values

```
effect_plot(model, pred = "R_AGE", interval = TRUE) +
  scale_x_continuous(name = "Age (years)",
                     breaks = seq(0, 70, by  =10)) +
  scale_y_continuous(name = "Probabilitity of walking",
                     breaks = breaks <- seq(0, 1, by = 0.01),
                     labels = paste0(breaks*100, "%")) + ggtitle("Effect of Age on Walki
ng Probability")
```



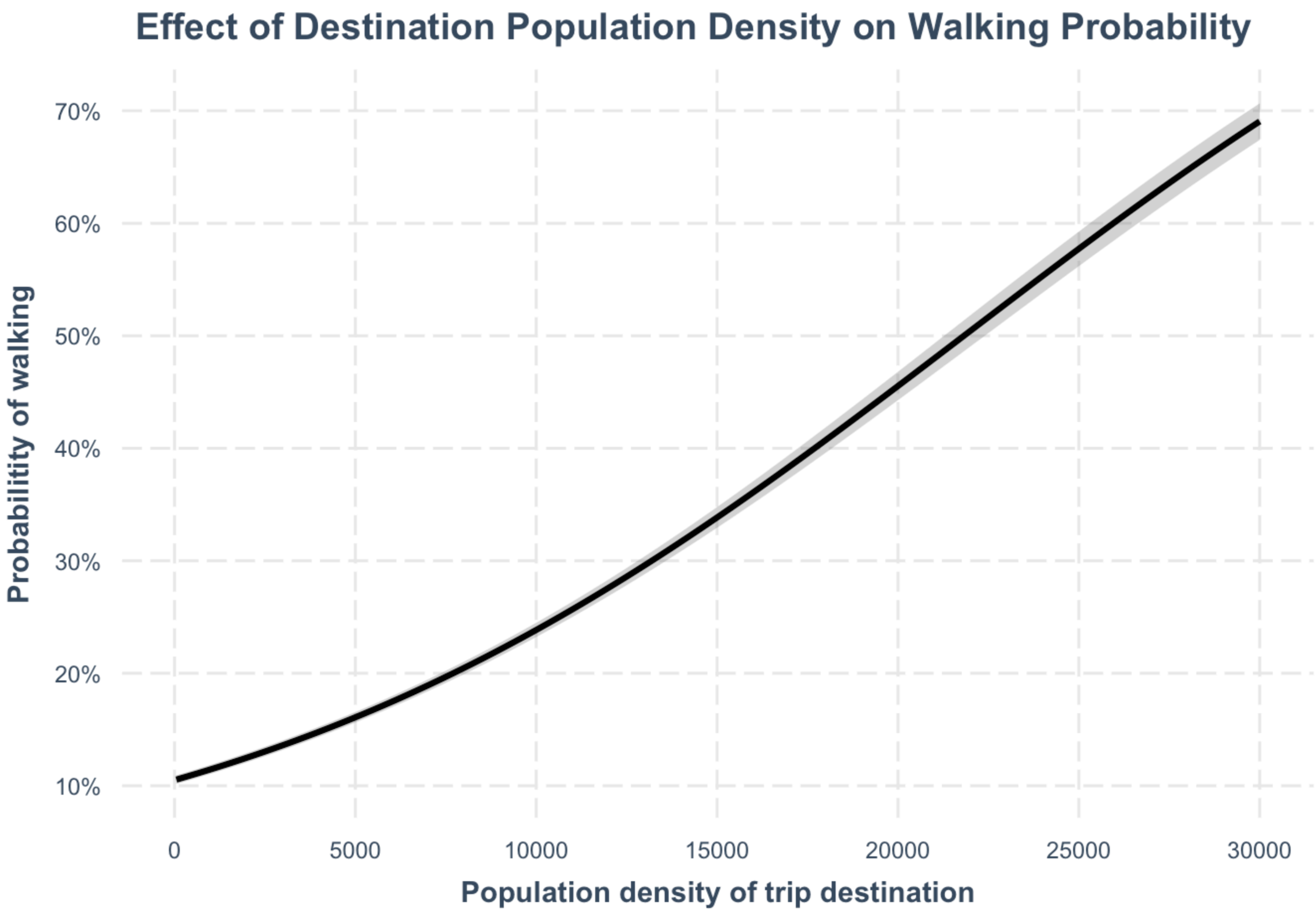
Urban/rural (URBRUR)

```
effect_plot(model = model, pred = "URBRUR_logical", interval = TRUE) +
  scale_y_continuous(name = "Probability of walking for a particular trip",
                     breaks = breaks <- seq(0, 1, by=0.01),
                     labels = paste0(breaks*100, "%")) +
  scale_x_discrete(name = paste0("Household is located in what kind of area"),
                   labels = c("Urban", "Rural")) + ggtitle("Effect of Household Urban v
s. Rural Location on Walking Probability")
```



Population density of the trip destination (DBPPOPDN)

```
effect_plot(model, pred = "DBPPOPDN", interval = TRUE) +
  scale_x_continuous(name = "Population density of trip destination",
                     breaks = seq(0, 30000, by  =5000)) +
  scale_y_continuous(name = "Probabilitity of walking",
                     breaks = breaks <- seq(0, 1, by = 0.1),
                     labels = paste0(breaks*100, "%")) + ggtitle("Effect of Destination
Population Density on Walking Probability")
```


Effect of Destination Population Density on Walking Probability
Probability of walking
70%
60%
50%
40%
30%
20%
10%
0
5000
10000
15000
20000
25000
30000
Population density of trip destination