# GEOG 738 Discrete Choice Analysis - Exercise 1

Haoran Xu

2025-01-21

## 1. Define "model".

A model is a simplified representation of a real-world system or problem that include exploring relationships between different things.

## 2. Why are models on a 1-to-1 scale undesirable?

- Models are for simplifying things but a 1-to-1 scale model would essentially recreate the original system, offering no reduction in complexity, which might make itself hard to demonstrate or analyze.
- Building a 1-to-1 scale a model, either physical or digital, would cost a lot of material and energy.
- 1-to-1 scale models might not be convenient or feasible for model experimentation or adjustments.

## 3. Invoke data set Mode from package mlogit. To do this you need to first load the package. This is a data set with choices about mode of transportation. This is done as follows:

```
library(mlogit)
data("Mode")
```

## 4. Describe this data set. How many variables are there and of which type (i.e., categorical/quantitative)?

There are nine variables and only `choice` is a nominal categorical variable coded as factor, while the other eight variables are all quantitative variables in ratio scale coded as numeric.

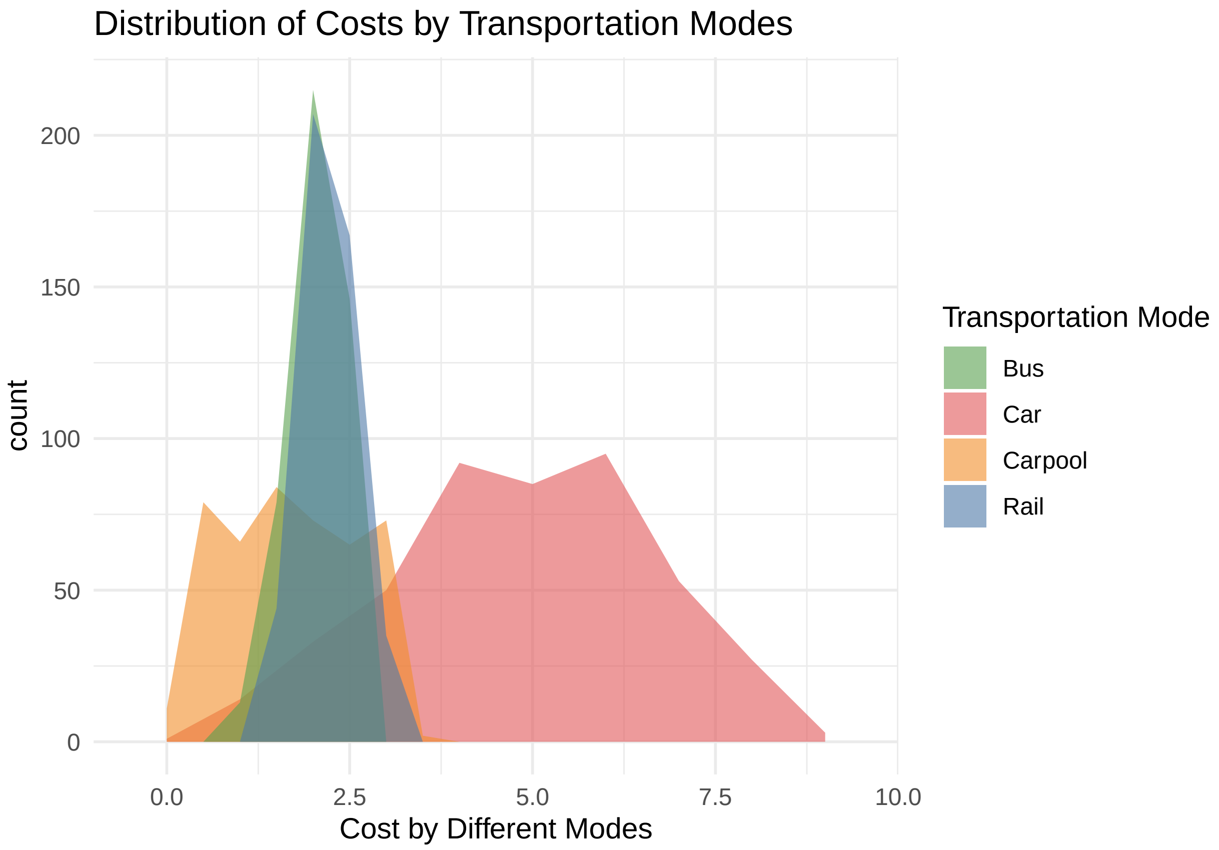
## 5. How many different modes of transportation are in this data set? What is the most popular mode? What is the least popular mode?

```
summary(Mode)
```

There are four modes of transportation in this data set, which are car, carpool, bus, and rail. Car is the most popular mode while carpool is the least popular.

## 6. In general, what is the most expensive mode? The least expensive?

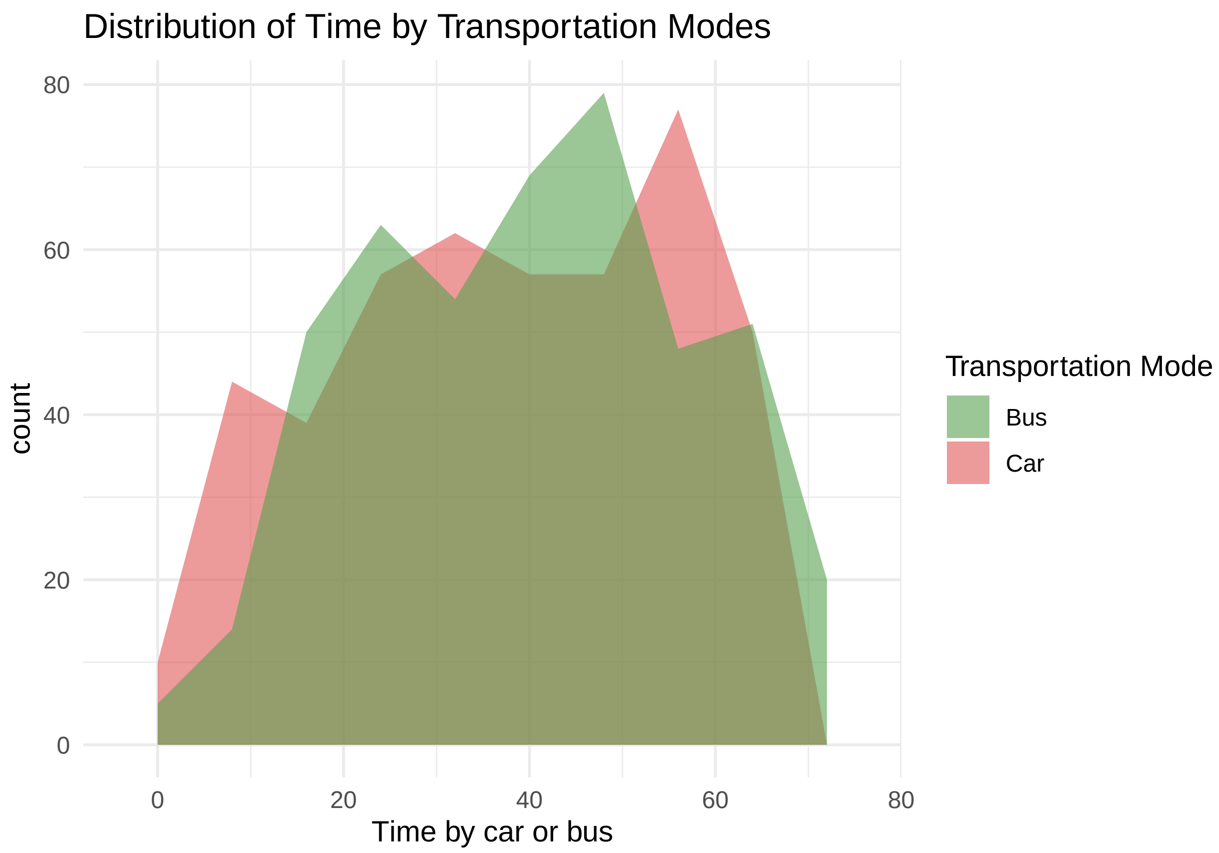
```
library(knitr)
library(ggplot2)

ggplot(data = Mode) +
  geom_area(aes(x = cost.car,
                fill = "Car"),
          stat = "bin",
          binwidth = 1,
          color = NA,
          alpha = 0.6) +
geom_area(aes(x = cost.carpool,
              fill = "Carpool"),
          stat = "bin",
          binwidth = 0.5,
          color = NA,
          alpha = 0.6) +
geom_area(aes(x = cost.bus,
              fill = "Bus"),
          stat = "bin",
          binwidth = 0.5,
          color = NA,
          alpha = 0.6) +
geom_area(aes(x = cost.rail,
              fill = "Rail"),
          stat = "bin",
          binwidth = 0.5,
          color = NA,
          alpha = 0.6) +
scale_fill_manual(values = c("Car" = "#E15759",
                             "Carpool" = "#F28E2B",
                             "Bus" = "#59A14F",
                             "Rail" = "#4E79A7")) +
labs(x = "Cost by Different Modes",
     fill = "Transportation Mode",
     title = "Distribution of Costs by Transportation Modes") +
theme_minimal() +
theme(legend.position = "right")
```



In general, traveling by car is the most expansive mode with the highest mean, average, and quantiles. Carpooling appears to be the least expansive mode with lowest median and mean, while cost of bus or rail are also few as carpooling. But mode of carpooling has higer standard deviation with wider spans.

### 7. Create a plot showing the univariate distributions of time by car and time by bus. Discuss.

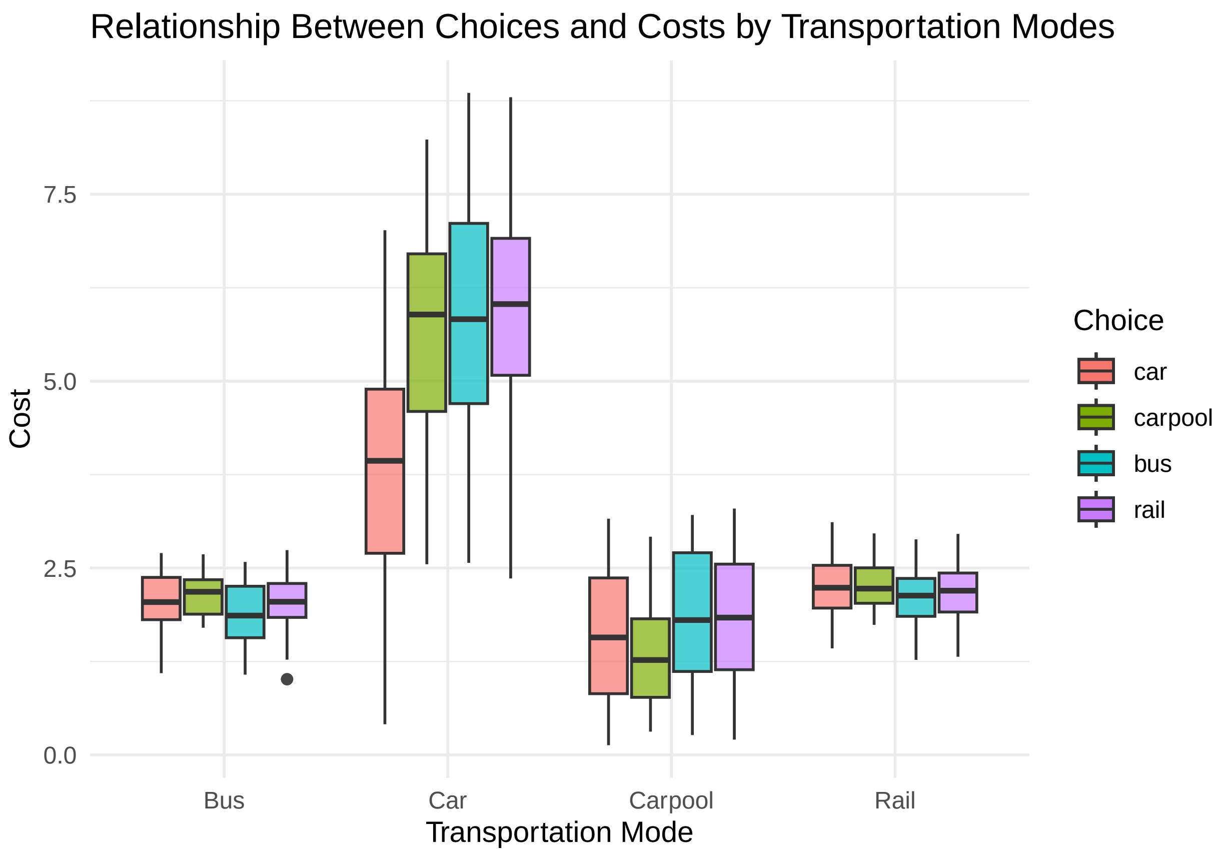
```
ggplot(data = Mode) +
  geom_area(aes(x = time.car, fill = "Car"),
            stat = "bin",
            binwidth = 8,
            color = NA,
            alpha = 0.6) +
  geom_area(aes(x = time.bus, fill = "Bus"),
            stat = "bin",
            binwidth = 8,
            color = NA,
            alpha = 0.6) +
  scale_fill_manual(values = c("Car" = "#E15759",
                             "Bus" = "#59A14F")) +
  labs(x = "Time by car or bus",
       fill = "Transportation Mode",
       title = "Distribution of Time by Transportation Modes") +
  theme_minimal() +
  theme(legend.position = "right")
```



The time of cars or buses range span from 1 min to 76 min. Time by car generally appear to have (with `binwidth` set as 8) three peaks which appear around 8 min, 32 min, and 56 min while time by bus has two peaks in 23 min and 48 min. The mean and median of time by bus is slightly higher than time by car, which is indicated in the graph with having shorter tais in the low end and longer tail in the high end.

## 8. How do choices relate to cost by the different modes?

```r
ggplot() +
  geom_boxplot(data = Mode, aes(x = "Car", y = cost.car, fill = choice), alpha = 0.7) +
  geom_boxplot(data = Mode, aes(x = "Carpool", y = cost.carpool, fill = choice), alpha = 0.7) +
  geom_boxplot(data = Mode, aes(x = "Bus", y = cost.bus, fill = choice), alpha = 0.7) +
  geom_boxplot(data = Mode, aes(x = "Rail", y = cost.rail, fill = choice), alpha = 0.7) +
  labs(
    title = "Relationship Between Choices and Costs by Transportation Modes",
    x = "Transportation Mode",
    y = "Cost",
    fill = "Choice"
  ) +
  theme_minimal()
```



The boxplot visualization shows how choices (i.e., car, carpool, bus, rail) relate to the cost for different transportation modes in three aspects:

1. Cost-Choice Trade-off

Choices like carpool, bus, and rail are associated with lower costs, while choosing a car results in the highest costs. This indicates individuals looking to minimize transportation expenses are more likely to choose public transportation or carpooling.

2. Stability of Public Transportation

Modes like bus and rail show lower and more stable costs, likely making them attractive options for cost-conscious individuals. Comparatively, carpooing also have lower costs but with more unstable variations.

3. Variability in Cars

Cars exhibit the highest cost and highest variability, possibly due to external factors like fuel costs, tolls, and varying trip distances. Besides, for people who choose cars, cars proved to be an effective mode for lowering their costs compared to other three modes.