Lab 2
Trip Generation (1)

Yufeng Zhang
zhan4879@umn.edu

CEGE-3201: Transportation Engineering

January 31, 2019

Objectives

- ▶ Get familiar with R
- ▶ Build linear models and make predictions in R
- ▶ Use Twin Cities demographic information and TBI 2010 Household Survey results for trip generation

Data sets

- **TAZ2010_wOrigDest.xlsx** is the main data set for today's lab
- It contains aggregated TAZ info and trip data(from TBI!) and demographic data from the US Census
- Use **TAZ2010-MetaData.pdf** as a reference for explanation of variables in the data set.

What is R

R is a free and widely used programming language for statistical computing and graphics.
R studio

- ▶ open-source Integrated Development Environment (IDE) for R
- ▶ Graphical User Interface(GUI), more helpful than console
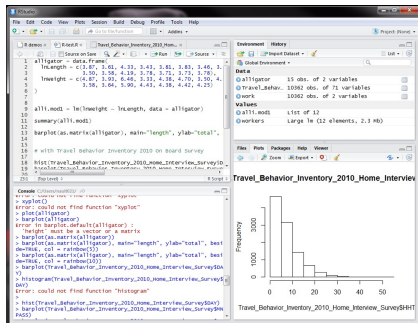
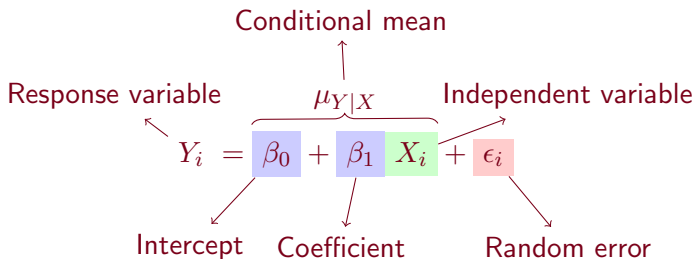In some cases, using R is much easier than Excel...

Figure: R studio interface

What is a linear model

Conditional mean

Response variable $\mu_{Y|X}$ Independent variable

$$Y_i \;=\; \beta_0 \;+\; \beta_1 \; X_i \;+\; \epsilon_i$$

Intercept   Coefficient   Random error

More generally:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \epsilon_i$$

We want to determine what types of sociodemographic and built environment characteristics produce trips and how...

▶ Response variable: Number of trips from/to each TAZ

▶ Independent (explanatory) variables: demographic characteristics (eg. population, race, households,...)

Fitting a linear model
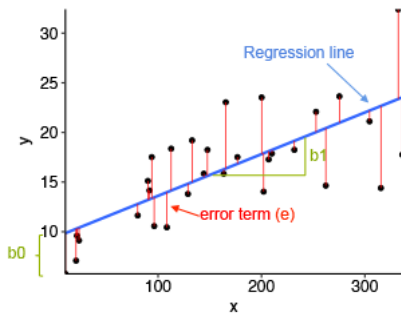
Minimizing total prediction error

Figure: Graphical illustration of linear regression[1]

---

[1]http:
//www.sthda.com/english/articles/39-regression-model-diagnostics/
161-linear-regression-assumptions-and-diagnostics-in-r-essentials

Linear regression in R...

```r
model <- lm(formula, data)
#Build a linear model
model1 <- lm( formula = Otrips ~ POPOVER18 +
    HHTYPE3, data = TAZ2010)
#Summarize regression model1
summary(model1)
```

Linear regression in R...


```
> summary(dest.model)

Call:
lm(formula = destTrips ~ totalEMP + retailEMP + nonretailEMP +
    hhTotal + popTotal + acres)

Residuals:
    Min      1Q  Median      3Q     Max
-6651.1  -174.7   -88.9    45.1  7915.4

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   58.759901  17.901476   3.282  0.00104 **
totalEMP       0.572309   0.010533  54.336  < 2e-16 ***
retailEMP     -0.259121   0.051095  -5.071 4.19e-07 ***
nonretailEMP        NA          NA      NA       NA
hhTotal       -0.308092   0.094519  -3.260  0.00113 **
popTotal       0.151587   0.036652   4.136 3.63e-05 ***
acres         -0.001903   0.002371  -0.803  0.42213
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 589.1 on 3024 degrees of freedom
Multiple R-squared:  0.5491,     Adjusted R-squared:  0.5483
F-statistic: 736.5 on 5 and 3024 DF,  p-value: < 2.2e-16
```

Figure: linear regression model summary

▶ Significance is probability of calculating the specific coefficient estimate given the data
▶ "Pr()": a smaller p-value indicates a smaller probability that the coefficient is 0

Please report information in red in your report.

A couple of reminders

- Two files to help you use R: "Example R Codes.pdf" and "Basic R commands.docx"
- Save the work from this lab in a safe place.
- Double check that labels are readable (especially after saving to .pdf)
- Focus on quality of work, not specific formatting details. Formatting isn't critical, but clarity is.