Rotman

INTRO TO R PROGRAMMING

R Tutorial (RSM358) – Session 2 (Optional Materials)

September 18, 2024 Prepared by Jay Cao / TDMDAL

Website: https://tdmdal.github.io/r-intro-2024-fall/

Rotman School of Management
UNIVERSITY OF TORONTO

Linear Regression - Housing Price & Clean Air

- Manipulate data
  - Load data
  - Create new columns
  - Filter columns and rows
- Build models
  - Multiple linear regressions
- Report and graph
  - Plot a few graphs
  - Report regression results

```
Obs:   506

1. price      median housing price, $
2. crime      crimes committed per capita
3. nox        nitrous oxide, parts per 100 mill.
4. rooms      avg number of rooms per house
5. dist       weighted dist. to 5 employ centers
6. radial     accessibiliy index to radial hghwys
7. proptax    property tax per $1000
8. stratio    average student-teacher ratio
9. lowstat    % of people 'lower status'
```

Choice 1: Use Only Base R packages

- Manipulate data
  - Load data (`read.csv()`)
  - Create new columns (`base R data frame manipulation`)
  - Filter columns and rows (`base R data frame manipulation`)

- Build models
  - Multiple regression (`lm()` from `stats` library in R base)

- Report and graph
  - Base R plot system, `plot()`
  - Base R `summary()` function

- A Note on Predictive Analysis
  - Train and test (or validation) split
  - Predict on test data and obtain evaluation measures of interest

Choice 2: Use 3-party Packages (**Optional**)

- Manipulate data (`tidyverse` eco-system)
  - Load data (`read_csv()` from the `readr`)
  - Create new columns (`mutate()` from `dplyr`)
  - Filter columns and rows (`select()` and `filter()` from `dplyr`)

- Build models
  - Multiple regression (`lm()` from `stats` library in R base)

- Report and graph
  - Graph using ggplot2 and some of its extensions
  - Build a publication-ready table (`huxreg()` from `huxtable` library)

Load a CSV file

- Choice1: `read.csv()` from Base R's utils library (load into dataframe)

```
read.csv(file)
```

```
e.g. hprice <- read.csv("hprice.csv")
```

- Choice 2: `read_csv()` from `tidyverse`'s `readr` library (load into tibble/dataframe)

```
read_csv(file)
```

```
e.g. hprice <- read_csv("hprice.csv")
```

Data Frame Manipulation – Base R vs dplyr

Data Process Operation	Base R	dplyr

Create a new column variable (from other column varibles)	`df$z <- df$x + df$y, or`
	
	`df["z"] <- df["x"] + df["y"], or`
	
	`transform()`
		`mutate(df, z = x + y)`

Filter rows based on conditions	`df[which(x), , drop = FALSE], or`
	
	`subset()`
		`filter(df, x)`

Select column variables	`df[c("x", "y")], or`
	
	`subset()`
		`select(df, x, y)`

…	…	…

Source: https://dplyr.tidyverse.org/articles/base.html

Data Manipulation: `dplyr` basics

- Filter observations (rows): **`filter()`**

`filter(my_dataframe, condition1, …)`

`e.g., hprice_reg <- filter(hprice, price > 20000)`

- Create new variables: **`mutate()`**

`mutate(my_dataframe, new_var1 = expression1, …)`

`e.g., hprice_reg <- mutate(hprice_reg, lprice = log(price))`

- Select variables (columns): **`select()`**

`select(my_dataframe, var1, …)`

`e.g., hprice_reg <- select(hprice_reg, lprice, rooms)`

Ref. Base R vs dplyr data frame manipulation.

Data Manipulation: Data Pipe (%>%)

```
hprice_reg <- filter(hprice, price > 20000)
hprice_reg <- mutate(hprice_reg, lprice = log(price))
hprice_reg <- select(hprice_reg, lprice, rooms)
```

```
hprice_reg <- hprice %>%
  filter(price > 20000) %>%
  mutate(lprice = log(price)) %>%
  select(lprice, rooms)
```

Regression

- Multiple regressions: `lm()` from `stats` library in base R

```
my_model <- lm(y ~ x1 + x2, data)
```

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$$

```
my_model <- lm(y ~ x1 + x2 + I(x1 * x2), data)
```

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon_i$$

- Regression result summary: `summary()`

Ref. https://faculty.chicagobooth.edu/richard.hahn/teaching/FormulaNotation.pdf

Report

- Choice 1: Summary table (Base R)
  - Summary for lm(): `summary(my_model)`

- Choice 2: publication-ready table: `huxreg()` from `huxtable` library

```
huxtable(my_model1, my_model2, …)
```

Ref. https://hughjonesd.github.io/huxtable/huxreg.html

Read the Regression Report

```
Call:
lm(formula = lprice ~ lnox + rooms + I(rooms^2) +
stratio, data = hprice_reg)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67205 -0.11678  0.01795  0.11597  0.59801

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.17845    0.47124  27.965  < 2e-16 ***
lnox        -0.58449    0.04594 -12.724  < 2e-16 ***
rooms       -0.69766    0.14221  -4.906 1.30e-06 ***
…
```

```
…
I(rooms^2)   0.07211    0.01129   6.385 4.29e-10 ***
stratio     -0.03929    0.00426  -9.223  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1833 on 448 degrees of freedom
Multiple R-squared:  0.6188, Adjusted R-squared:  0.6154
F-statistic: 181.8 on 4 and 448 DF,  p-value: < 2.2e-16
```

Interpret Regression Result (Coefficients)

- $y = \hat{\beta}_0 + \widehat{\boldsymbol{\beta}}_{\mathbf{1}} x_1 + \hat{\beta}_2 x_2$ (x_1 is continuous)
- $y = \hat{\beta}_0 + \widehat{\boldsymbol{\beta}}_{\mathbf{1}} x_1 + \hat{\beta}_2 x_2$ (x_1 is categorical, say, 0 or 1)

- $\log(y) = \hat{\beta}_0 + \widehat{\boldsymbol{\beta}}_{\mathbf{1}} x_1 + \hat{\beta}_2 x_2$ (y is log-transformed)
- $y = \hat{\beta}_0 + \widehat{\boldsymbol{\beta}}_{\mathbf{1}} \log(x_1) + \hat{\beta}_2 x_2$ (x_1 is log-transformed)
- $\log(y) = \hat{\beta}_0 + \widehat{\boldsymbol{\beta}}_{\mathbf{1}} \log(x_1) + \hat{\beta}_2 x_2$ (y and x_1 are log-transformed)

- $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$ (an interactive term)

Ref. https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-variables-are-log-transformed/

A Note on Predictive Analysis

- Causal vs predictive analysis

- Training and test (validation) data split

- Three Steps
  1. randomly split the data into training and test set.
  2. train/estimate a model on training set.
  3. Evaluate the estimated model on test set, i.e., predict on the test set, and obtain evaluation measures of interest.