An Brief Introduction to R

Frank Qu

9/24/2017

目录

1 Learning Object

- Introduction to R and RStudio
- Reading data: importing datasets, data types, defining variable classes
- Manipulating data: cleaning, manipulate, package dplyr
- Analyzing data: statistical properties, regression model, limited dependent variables
- Visualizing data: built-in plotting functions and ggplot2 package

2 Getting Stared With R

- Not only a statistical programming language, but a computing environment for statistical computing and graphics.
- Powerful Programming and Extending Capability
- Multiple Platforms
- Very excellent graphics
- A big but not a determinate advantage: FREE Open Source

2.1 Installing

2.1.1 Installing R (skip)

2.1.2 Using IDE: RStudio (skip)

- The most popular IDE for R
- Also Free(for basic version)

- Combine with `Markdown` and `Latex` to make scientific writings or presentation easier
- Download it from here: [RStudio]{https://www.rstudio.com/products/rstudio/download/}

2.2 Using R as `Stata`: Packages

- Many researchers provide their own R programs through the R project webpage.
- Many packages are already preinstalled in the basic R installation.
- They can be directly activated from RStudio.
- Or they are activated by issuing a command in the Console.

```
#install.packages("AER",repos = "http://mirrors.xmu.edu.cn/CRAN/")
#library("AER")
#install.packages("haven",repos = "http://mirrors.xmu.edu.cn/CRAN/")
```

2.3 Where to get help

- The online help in R describes all basic R commands as well as commands in active packages.
- search the online help from the Help pane in RStudio.
- Alternatively, using the command

```
?load
# or
help("load")
# or
??load
# or
help.search("read")
```

3 Basic Data Management in R

3.1 Working Directory

- R will look for data or save data in the drive and working directory.
- The working directory is specified depending on the operation system

```
getwd()
```

```
## [1] "/Users/byelenin/Dropbox/R/R_Class/Intro_Metrics"
```

3.2 Changing the Working Directory

```
setwd("/Users/byelenin/Dropbox/R/R_Class/Metrics/Lec1/")
getwd()
```

```
## [1] "/Users/byelenin/Dropbox/R/R_Class/Metrics/Lec1"
```

3.3 Importing Data: From STATA

- R will look for data or save data in the drive and working directory.
- The working directory is specified depending on the operation system
- imports data from STATA

```
#install.packages("haven",repos = "http://mirrors.xmu.edu.cn/CRAN/")
library(haven)
caschool_data <- read_dta("/Users/byelenin/Dropbox/R/R_Class/Metrics/Lec1/caschool.d
View(caschool_data)
```

3.4 Importing Data: From CSV

```
caschool_csv <- read.csv("/Users/byelenin/Dropbox/R/R_Class/Metrics/Lec1/caschool.cs
View(caschool_csv)
```

3.5 Summary the Data

```
summary(caschool_data)
```

```
##     observat        dist_cod        county            district
##  Min.   :  1.0   Min.   :61382   Length:420         Length:420
##  1st Qu.:105.8   1st Qu.:64308   Class :character   Class :character
##  Median :210.5   Median :67760   Mode  :character   Mode  :character
##  Mean   :210.5   Mean   :67473
##  3rd Qu.:315.2   3rd Qu.:70419
##  Max.   :420.0   Max.   :75440
##    gr_span             enrl_tot          teachers          calw_pct
##  Length:420         Min.   :   81.0   Min.   :   4.85   Min.   : 0.000
##  Class :character   1st Qu.:  379.0   1st Qu.:  19.66   1st Qu.: 4.395
##  Mode  :character   Median :  950.5   Median :  48.56   Median :10.520
##                     Mean   : 2628.8   Mean   : 129.07   Mean   :13.246
##                     3rd Qu.: 3008.0   3rd Qu.: 146.35   3rd Qu.:18.981
##                     Max.   :27176.0   Max.   :1429.00   Max.   :78.994
##     meal_pct         computer         testscr         comp_stu
##  Min.   :  0.00   Min.   :   0.0   Min.   :605.5   Min.   :0.00000
##  1st Qu.: 23.28   1st Qu.:  46.0   1st Qu.:640.0   1st Qu.:0.09377
##  Median : 41.75   Median : 117.5   Median :654.5   Median :0.12546
##  Mean   : 44.71   Mean   : 303.4   Mean   :654.2   Mean   :0.13593
##  3rd Qu.: 66.86   3rd Qu.: 375.2   3rd Qu.:666.7   3rd Qu.:0.16447
##  Max.   :100.00   Max.   :3324.0   Max.   :706.8   Max.   :0.42083
##     expn_stu         str            avginc           el_pct
##  Min.   :3926   Min.   :14.00   Min.   : 5.335   Min.   : 0.000
##  1st Qu.:4906   1st Qu.:18.58   1st Qu.:10.639   1st Qu.: 1.941
##  Median :5215   Median :19.72   Median :13.728   Median : 8.778
##  Mean   :5312   Mean   :19.64   Mean   :15.317   Mean   :15.768
##  3rd Qu.:5601   3rd Qu.:20.87   3rd Qu.:17.629   3rd Qu.:22.970
##  Max.   :7712   Max.   :25.80   Max.   :55.328   Max.   :85.540
##     read_scr        math_scr
##  Min.   :604.5   Min.   :605.4
##  1st Qu.:640.4   1st Qu.:639.4
##  Median :655.8   Median :652.5
##  Mean   :655.0   Mean   :653.3
##  3rd Qu.:668.7   3rd Qu.:665.9
##  Max.   :704.0   Max.   :709.5
```

3.6 Variables

```
#install.packages("dplyr")
names(caschool_data)
```

```
##  [1] "observat" "dist_cod" "county"   "district" "gr_span"  "enrl_tot"
##  [7] "teachers" "calw_pct" "meal_pct" "computer" "testscr"  "comp_stu"
## [13] "expn_stu" "str"      "avginc"   "el_pct"   "read_scr" "math_scr"
```

3.6.1 Keeping Variables

```
caschool_data_small <- select(caschool_data,observat,testscr,str,expn_stu,el_pct)
head(caschool_data_small)
```

```
## # A tibble: 6 x 5
##   observat testscr      str expn_stu    el_pct
##      <dbl>   <dbl>    <dbl>    <dbl>     <dbl>
## 1        1  690.80 17.88991 6384.911  0.000000
## 2        2  661.20 21.52466 5099.381  4.583333
## 3        3  643.60 18.69723 5501.955 30.000002
## 4        4  647.70 17.35714 7101.831  0.000000
## 5        5  640.85 18.67133 5235.988 13.857677
## 6        6  605.55 21.40625 5580.147 12.408759
```

3.6.2 Generate new variable

```
caschool_data_small$logexp <- log(caschool_data$expn_stu)

caschool_data_small$el_high <- caschool_data$el_pct>=50



head(caschool_data_small)
```

```
## # A tibble: 6 x 7
##   observat testscr      str expn_stu    el_pct   logexp el_high
##      <dbl>   <dbl>    <dbl>    <dbl>     <dbl>    <dbl>   <lgl>
## 1        1  690.80 17.88991 6384.911  0.000000 8.761693   FALSE
## 2        2  661.20 21.52466 5099.381  4.583333 8.536874   FALSE
## 3        3  643.60 18.69723 5501.955 30.000002 8.612859   FALSE
## 4        4  647.70 17.35714 7101.831  0.000000 8.868108   FALSE
## 5        5  640.85 18.67133 5235.988 13.857677 8.563311   FALSE
## 6        6  605.55 21.40625 5580.147 12.408759 8.626970   FALSE
```

3.7 Descriptive Statistics

- summary a variable

```
summary(caschool_data_small$testscr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   605.5   640.0   654.5   654.2   666.7   706.8
```

- if the dataframe is attached, simply

```
attach(caschool_data_small)
summary(testscr)
```

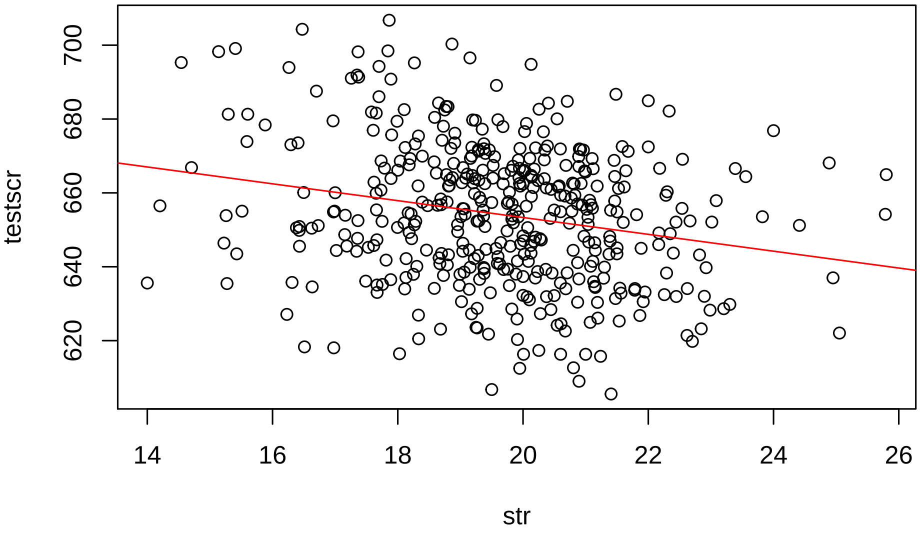
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   605.5   640.0   654.5   654.2   666.7   706.8
```

4 Plot

4.1 Scatter Plot

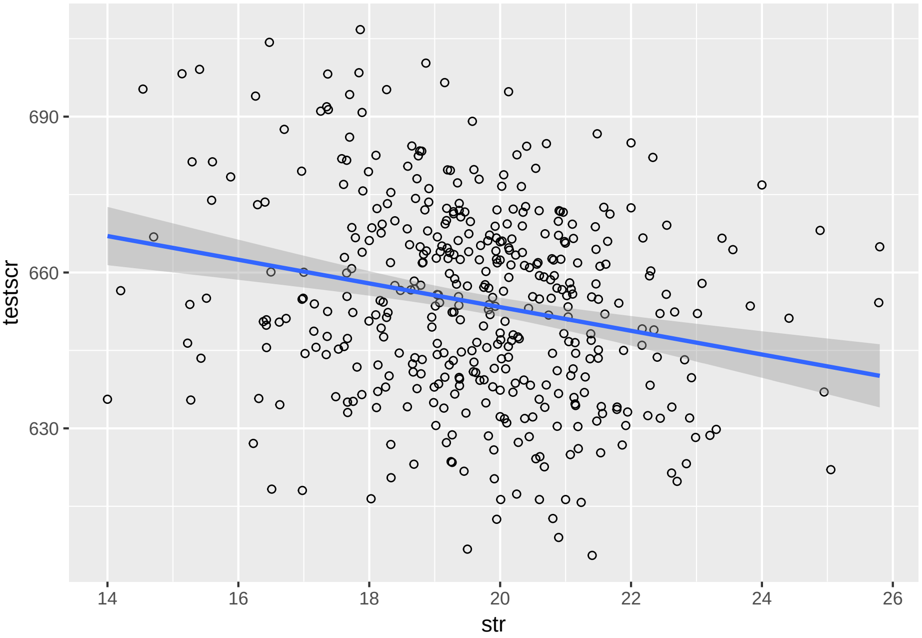
- Draw a scatter plot of the variable “testscr” against “str”:

```
plot(str, testscr)
abline(lm(testscr ~ str , data = caschool_data_small),col = "red")
```



- ggplot2

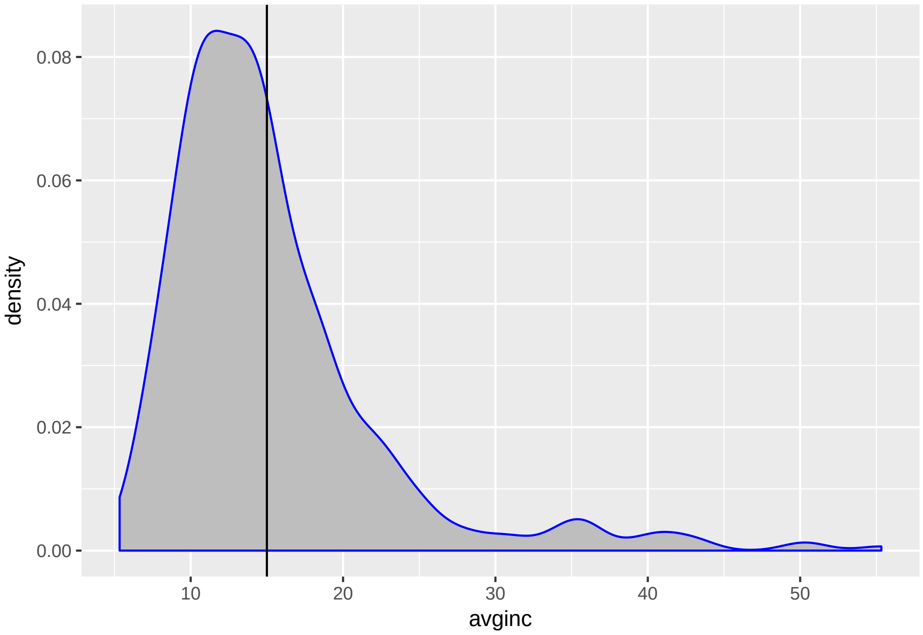
```
library("ggplot2")
ggplot(data =caschool_data_small,aes(x=str, y=testscr)) +
    geom_point(shape=1) +     # Use hollow circles
    geom_smooth(method=lm)    # Add linear regression line
```



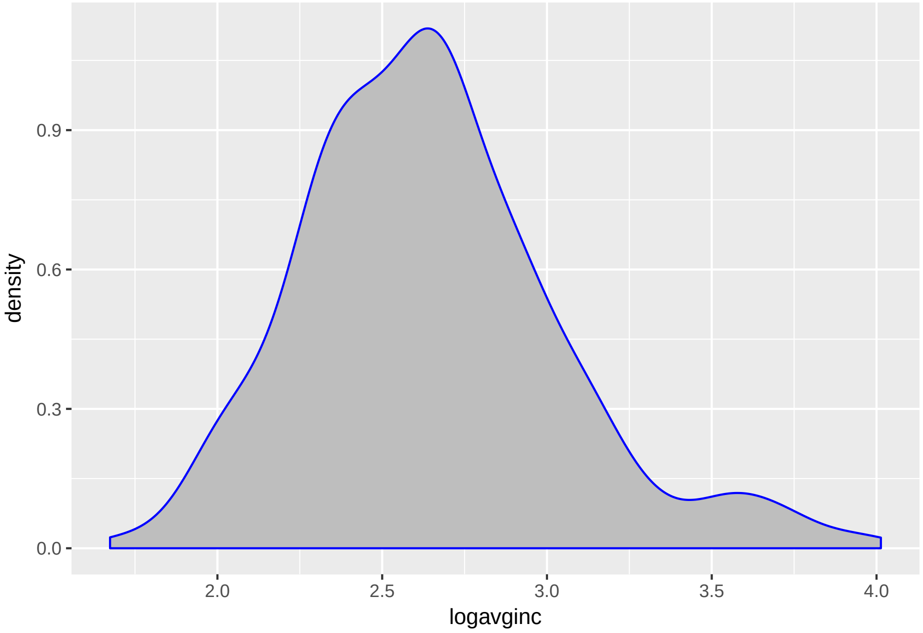
```
                                     #  (by default includes 95% confidence region)
```

4.2 A kdensity distribution of income

```
caschool_data$inc <- with(caschool_data,avginc >=15)
ggplot(caschool_data,aes(x=avginc))+
  geom_density(fill="grey",color ="blue")+
  geom_vline(xintercept = 15)
```



```
caschool_data$logavginc <- log(caschool_data$avginc)
ggplot(caschool_data,aes(x=logavginc))+
  geom_density(fill="grey",color ="blue")
```



5 T-test in R

5.1 single sample

- t-test for scores

```
summary(caschool_data_small$testscr)
```

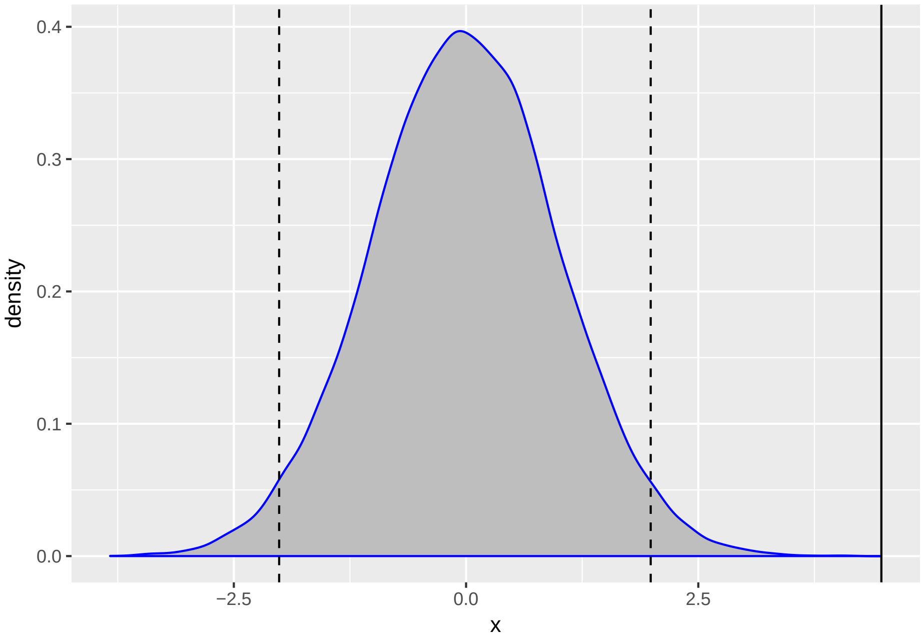
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   605.5   640.0   654.5   654.2   666.7   706.8
```

```
t.test(caschool_data_small$testscr,alternative = "two.sided",mu = 650)
```

```
##
##  One Sample t-test
##
## data:  caschool_data_small$testscr
## t = 4.4708, df = 419, p-value = 1.005e-05
## alternative hypothesis: true mean is not equal to 650
## 95 percent confidence interval:
##  652.3291 655.9840
## sample estimates:
## mean of x
##  654.1565
```

- Construct t-Statistics

```
randT <- rt(30000,df=NROW(testscr)-1) # build a distribution
scoreTtest <- t.test(caschool_data_small$testscr,alternative = "two.sided",mu = 650)
ggplot(data.frame(x=randT)) +
         geom_density(aes(x=x),fill = "grey",color ="blue") +
         geom_vline(xintercept = scoreTtest$statistic) +
         geom_vline(xintercept = mean(randT) + c(-2,2)*sd(randT),linetype = 2)
```



5.2 T-test for the difference between two means

```
t.test(testscr~el_high,data = caschool_data_small)
```

```
##
##  Welch Two Sample t-test
##
## data:  testscr by el_high
## t = 16.419, df = 47.709, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  26.19422 33.50602
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            656.1466            626.2964
```

6 Rstudio for run commands and processing markdown files

6.1 R script similar to Stata dofile

- A script is a text file with a set of R commands that can be executed jointly.
- Script files are convenient because they automate tasks relative to type each command in the Console
- Open a R script from the top-left corner or File, New File, R Script etc.

6.2 Rmarkdown documents

- generate a new Rmarkdown

7 Online Resource

7.1 R tutorial

- Datacamp
- Big Data University

7.2 Markdown

7.3 Latex