TA Session Sep 27 2021

Chunyu Qu

09 27, 2021

The initial line in a code chunk may include various options. For example, echo=FALSE indicates that the code will not be shown in the final document.

You can use include=FALSE to exclude everything in a chunk.

If you only want to suppress messages, use message=FALSE instead.

If you want to block warnings, use warning=FALSE instead.

```
knitr::opts_chunk$set(echo = TRUE)
```

0.Introduction to R and R markdown

(1)Prepare

Download R first from "https://www.r-project.org/"

Download R markdown from "https://www.rstudio.com/products/rstudio/download/"

Some Debug points:
(1) Install Miltex if you have to use it. Alternatively, you can use

```
### install.packages('tinytex')
### tinytex::install_tinytex()
```

(2) To adjust Miltex
"https://github.com/rstudio/rmarkdown/issues/1285"

(2)Format

"https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf"

1. Import the data of CSV into R

```
# Make sure where we are working at first
getwd()
```

```
## [1] "D:/Google Drive/Fordham/2019 Spring/AE/TA/TA1"
```

```r
# Use CSV to read the data
# "Header=true" specifies that this data includes a header row, the names of the title row are now turn
data=read.csv(’401k.csv’,header=TRUE,sep=",")
# data=read.csv(’401k.csv’,header=FALSE,sep=",")

# The database is attached to the R search path. This means that the database is searched by R when eva
attach(data)

# To view the data in spreadsheet form
fix(data)

write.csv(data, file = "MyData.csv")
```

2. Basic Statistics in R

```r
# 1. Shows the types of the data
# for a certain variable
typeof(sole)
```

```
## [1] "integer"
```

```r
typeof(data)
```

```
## [1] "list"
```

```r
# for the whole dataset
str(data)
```

```
## ’data.frame':    1534 obs. of  8 variables:
##  $ prate  : num  26.1 100 97.6 100 82.5 100 100 92.5 100 96.8 ...
##  $ mrate  : num  0.21 1.42 0.91 0.42 0.53 1.82 0.53 0.34 0.22 0.6 ...
##  $ totpart: num  1653 262 166 257 591 ...
##  $ totelg : num  6322 262 170 257 716 ...
##  $ age    : num  8 6 10 7 28 7 31 13 21 10 ...
##  $ totemp : num  8709 315 275 500 933 ...
##  $ sole   : num  0 1 1 0 1 1 1 0 1 1 ...
##  $ ltotemp: num  9.07 5.75 5.62 6.21 6.84 ...
```

```r
# Check for NA
NAcheck=is.na(data) # returns TRUE of x is missing
sum(NAcheck)
```

```
## [1] 0
```

```r
# 2. Means
# (1) One-by-one
mean(prate)
```

```
## [1] 87.36291
```

```r
mean(mrate)
```

```
## [1] 0.7315124
```

```r
mean(totpart)
```

```
## [1] 1354.231
```

```r
mean(totelg)
```

```
## [1] 1628.535
```

```r
mean(age)
```

```
## [1] 13.18123
```

```r
mean(totemp)
```

```
## [1] 3567.321
```

```r
mean(sole)
```

```
## [1] 0.4876141
```

```r
mean(ltotemp)
```

```
## [1] 6.686034
```

```r
# (2) For a bundle
colMeans(data, na.rm = FALSE, dims = 1)
```

```
##        prate        mrate      totpart       totelg          age       totemp
##   87.3629074    0.7315124 1354.2307692 1628.5345502   13.1812256 3567.3213820
##         sole      ltotemp
##    0.4876141    6.6860342
```

```r
# A quick question, why the following does not work?
# mean(data)
# data.n=as.numeric(unlist(data))
# typeof(data.n)

# 3. Variance
varr=var(data)


# 4. Standard Deviation
sd=sqrt(diag(varr))
```

3.Constructing Subsets in R

Suppose we would like to to create a data set which is a subset of the main data file in use. This corresponds to sole = 1 in the data set. We can do this by using the subset command:

```r
datasub=subset(data,sole==1)
detach(data)
attach(datasub)
mean(prate)
```

```
## [1] 90.07487
```

4.Running Regression in R

To run a simple regression in R lm command:

```r
attach(data)
```

```
## The following objects are masked from datasub:
##
##     age, ltotemp, mrate, prate, sole, totelg, totemp, totpart
```

```r
output = lm(prate ~ age + mrate)
summary(output)
```

```
##
## Call:
## lm(formula = prate ~ age + mrate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.162  -8.067   4.787  12.474  18.256
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  80.1191     0.7790  102.85  < 2e-16 ***
## age           0.2432     0.0447    5.44 6.21e-08 ***
## mrate         5.5213     0.5259   10.50  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 15.94 on 1531 degrees of freedom
## Multiple R-squared:  0.09225,    Adjusted R-squared:  0.09106
## F-statistic: 77.79 on 2 and 1531 DF,  p-value: < 2.2e-16
```

Suppose we would just like to run the same regression for only the 401k plans that are older than 8 years old. We can do this by creating a new data set that is a subset of the original one

```r
# Let us make a subset again
older8data = subset(data,age>8)

# Now let us detach the previous whole dataset and attach our current subset of interest
detach(data)
attach(older8data)
```

```
## The following objects are masked from datasub:
##
##     age, ltotemp, mrate, prate, sole, totelg, totemp, totpart
```

```r
outputolder8 = lm(prate ~ age + mrate)
summary(outputolder8)
```

```
##
## Call:
## lm(formula = prate ~ age + mrate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.910  -5.686   4.816   9.831  13.619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 85.56532    1.22439  69.884  < 2e-16 ***
## age          0.07157    0.05356   1.336    0.182
## mrate        4.28636    0.60544   7.080 3.24e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 13.51 on 774 degrees of freedom
## Multiple R-squared:  0.06483,    Adjusted R-squared:  0.06241
## F-statistic: 26.83 on 2 and 774 DF,  p-value: 5.425e-12
```