Hypothesis Testing

Zhidong Chen

Setup

```
# Read data
lawsch85 <- read.csv("data/LAWSCH85.csv")

# Load packages
library(car)
```

Linear Hypothesis Testing

1. Single parameter T-test and Joint Significance F-test

```
# See what variables we have in the data
names(lawsch85)
```

```
##  [1] "rank"    "salary"  "cost"    "LSAT"    "GPA"     "libvol"  "faculty"
##  [8] "age"     "clsize"  "north"   "south"   "east"    "west"    "lsalary"
## [15] "studfac" "top10"   "r11_25"  "r26_40"  "r41_60"  "llibvol" "lcost"
```

```
# Run full unrestricted linear model
model.ur <- lm(lsalary ~ LSAT +
                         GPA  +
                         lcost +
                         rank,
               data = lawsch85)

# Pull a summary of the model resutls
summary(model.ur)
```

```
##
## Call:
## lm(formula = lsalary ~ LSAT + GPA + lcost + rank, data = lawsch85)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.313018 -0.080990 -0.007693  0.078588  0.296160
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.560139   0.538439  15.898  < 2e-16 ***
## LSAT         0.006873   0.004029   1.706  0.09037 .
```

```
## GPA          0.257218   0.091890   2.799  0.00589 **
## lcost        0.036032   0.032753   1.100  0.27328
## rank        -0.003731   0.000324 -11.515  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.115 on 132 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.8329, Adjusted R-squared:  0.8279
## F-statistic: 164.5 on 4 and 132 DF,  p-value: < 2.2e-16
```

Single coefficent T-test and F-test of Joint significance of all slope parameters are included in the results of summary function

2.Test linear combination of parameters

~ Method 1: Use "car::linearHypothesis()"

```
# Examples of different linear hypothesis testing
linearHypothesis(model.ur, c("LSAT = 0", "GPA = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## LSAT = 0
## GPA = 0
##
## Model 1: restricted model
## Model 2: lsalary ~ LSAT + GPA + lcost + rank
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    134 2.1024
## 2    132 1.7462  2   0.35618 13.462 4.778e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# or lht(model.ur, c("LSAT = 0", "GPA = 0"))
linearHypothesis(model.ur, "LSAT + GPA = 0", test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## LSAT  + GPA = 0
##
## Model 1: restricted model
## Model 2: lsalary ~ LSAT + GPA + lcost + rank
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
```

```
## 1    133 1.8615
## 2    132 1.7462  1   0.11529 8.7153 0.003736 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(model.ur, "2*LSAT + 1*GPA = 0")
```

```
## Linear hypothesis test
##
## Hypothesis:
## 2 LSAT  + GPA = 0
##
## Model 1: restricted model
## Model 2: lsalary ~ LSAT + GPA + lcost + rank
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    133 1.8741
## 2    132 1.7462  1    0.1279 9.6687 0.002297 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(model.ur, "2*LSAT + 1*GPA - rank= 0")
```

```
## Linear hypothesis test
##
## Hypothesis:
## 2 LSAT  + GPA - rank = 0
##
## Model 1: restricted model
## Model 2: lsalary ~ LSAT + GPA + lcost + rank
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    133 1.8781
## 2    132 1.7462  1    0.1319 9.9711 0.001971 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

~ Method 2: Run restricted model, then use "anova(model.ur, model.r)"

```
model.r1 <- lm(lsalary ~ llibvol + lcost + rank, data = lawsch85)

# If you we use "anova(model.ur, model.r1)" directly, we get error message:
# Error in anova.lmlist(object, ...) :
# models were not all fitted to the same size of dataset

# To correct for the error, we need to remove the missing values.
# Missing values are coded as "NA" in R.
# is.na(object) tests weather each item of the "object" has a missing value.
```

```r
# It returns logical values, "TRUE" or "FALSE".
# "!" in front of is.na(object) is the negate sign.
```

```r
# Remove missing values of GPA and LSAT
lawsch85.subset <- lawsch85[!is.na(lawsch85$GPA) & !is.na(lawsch85$LSAT), ]

# Unrestricted linear model
model.ur2 <- lm(lsalary ~ LSAT +
                        GPA +
                        llibvol +
                        lcost +
                        rank,
                data = lawsch85.subset)

# Restricted linear Model
model.r2 <- lm(lsalary ~ llibvol +
                        lcost +
                        rank,
                data = lawsch85.subset)

# Hypothesis testing using F-test
anova(model.ur2, model.r2)
```

Solution:

```
## Analysis of Variance Table
##
## Model 1: lsalary ~ LSAT + GPA + llibvol + lcost + rank
## Model 2: lsalary ~ llibvol + lcost + rank
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    130 1.6427
## 2    132 1.8942 -2  -0.25151 9.9518 9.518e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A note about "car::lht()": In function lht(), the default test type is an asymptotic Chi-square test. When the model we put in is a lm model such as in all the above examples, the default test type changes to F-test automatically.

Caution! When the model we put in is a different type of model, such as the felm model, to conduct a F-test, we add an option to your function: test = "F".

```r
# For illurstration
linearHypothesis(model.ur2, c("LSAT = 0", "GPA = 0"), test="F")
```

```
## Linear hypothesis test
##
```

```
## Hypothesis:
## LSAT = 0
## GPA = 0
##
## Model 1: restricted model
## Model 2: lsalary ~ LSAT + GPA + llibvol + lcost + rank
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    132 1.8942
## 2    130 1.6427  2   0.25151 9.9518 9.518e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.Recover F-statistic using regression results

F-statistic formula using Sum of Squared Residuals:SSR

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

Plug in the values from above regressions:

$$\begin{aligned} F &= \frac{(1.8942 - 1.6427)/2}{1.6427/130} \\ &= 9.951604 \end{aligned}$$