Hypothesis Testing in R

Justin Smith

2023-10-31

Background

- OLS is a way to estimate unknown parameters
- Gives us a **point estimate**
  - A single number to estimate the parameter
- This estimate is subject to sampling variation
  - You get a different value in each hypothetical sample
- The sampling uncertainty makes it impossible to make definitive statements about the value of the parameter
- But we can make probabilistic statements about the parameter
- To do this
  - Assume a value for the parameter (the null hypothesis)
  - Determine the sampling distribution of our estimator when the null hypothesis is true
  - Figure out where the estimate falls in the distribution
  - Decide whether the null hypothesis is likely false or likely true
- This process is **Hypothesis Testing**

Setup

Population Regression Model

- Recall the population regression is

$$y = \mathbf{x}\boldsymbol{\beta} + u$$

- Where
  - y is the outcome variable
  - $\mathbf{x}$ is a vector of independent variables
  - $\boldsymbol{\beta}$ is the corresponding vector of slopes
  - u is the population residual
- The population regression slope vector is

$$\boldsymbol{\beta} = (\mathbf{E}[\mathbf{x}'\mathbf{x}])^{-1}\mathbf{E}[\mathbf{x}'y]$$

Ordinary Least Squares

- The associated OLS estimator for the population slope vector is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$$

Sampling Distribution of the slope estimator

- With a large sample, the central limit theorem implies

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, n^{-1}[\mathbf{E}(\mathbf{x}'\mathbf{x})^{-1}]\mathbf{E}(u^2\mathbf{x}'\mathbf{x})[\mathbf{E}(\mathbf{x}'\mathbf{x})^{-1}])$$

- When it comes to doing hypothesis tests, we substitute in an estimate for the standard errors

Estimation

- Use the `bwght` data from the `wooldridge` package to estimate a regression

```
library(wooldridge)
library(stargazer)
data <- bwght
reg <- lm(bwght ~faminc + cigs, data = data)
stargazer(reg, type = "text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                 bwght
## ------------------------------------------------
## faminc                         0.093***
##                                 (0.029)
##
## cigs                           -0.463***
##                                 (0.092)
##
## Constant                      116.974***
##                                 (1.049)
##
## ------------------------------------------------
## Observations                     1,388
## R2                               0.030
## Adjusted R2                      0.028
## Residual Std. Error      20.063 (df = 1385)
## F Statistic           21.274*** (df = 2; 1385)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Hypothesis Test About Single Parameter

- The information for a t-test is in the `stargazer` output
  - Displays the standard error and p-value stars
- We can still do this manually for instructional purposes
- One package or hypothesis testing is `lmtest`

```r
library(lmtest)
library(sandwich)
data <- bwght
reg <- lm(bwght ~faminc + cigs, data = data)
coeftest(reg)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 116.974130   1.048984 111.5118 < 2.2e-16 ***
## faminc        0.092765   0.029188   3.1782  0.001515 **
## cigs         -0.463408   0.091577  -5.0603 4.747e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The function `coeftest()` computes individual statistics for hypothesis testing
  - We can compare the t-value to the critical values based on chosen significance level
  - Can also use the p-value
- Problem: by default these use non-robust standard errors
- We can make them robust easily with `coeftest()`

```r
library(sandwich)
data <- bwght
reg <- lm(bwght ~faminc + cigs, data = data)
coeftest(reg, vcov = vcovHC, type = "const")
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 116.974130   1.048984 111.5118 < 2.2e-16 ***
## faminc        0.092765   0.029188   3.1782  0.001515 **
## cigs         -0.463408   0.091577  -5.0603 4.747e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
coeftest(reg, vcov = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 116.974130   1.037207 112.7780 < 2.2e-16 ***
## faminc        0.092765   0.028586   3.2451  0.001202 **
## cigs         -0.463408   0.088759  -5.2209 2.052e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The first gives the non-robust errors
- The second are robust
- It is unfortunately not easy to get these into `stargazer`
- We have to trick it into using them

- A package that helps with this is `estimatr`

```
library(estimatr)
data <- bwght
reg <- lm(bwght ~faminc + cigs, data = data)
stargazer(reg, se = starprep(reg), se_type = "HC1", type = "text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                 bwght
## ------------------------------------------------
## faminc                         0.093***
##                                 (0.029)
##
## cigs                           -0.463***
##                                 (0.089)
##
## Constant                      116.974***
##                                 (1.037)
##
## ------------------------------------------------
## Observations                     1,388
## R2                               0.030
## Adjusted R2                      0.028
## Residual Std. Error      20.063 (df = 1385)
## F Statistic           21.274*** (df = 2; 1385)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
##
## ===
## HC1
## ---
```

- Note that this still does not fix the F-statistic at the bottom

Joint Hypothesis Tests

- Suppose we want to jointly test that family income and cigarettes do not affect birthweight
- To do that, we need the `car` package and the `linearHypothesis()` function

```
library(car)
data <- bwght
reg <- lm(bwght ~faminc + cigs, data = data)
linearHypothesis(reg, c("faminc=0","cigs=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## faminc = 0
## cigs = 0
##
## Model 1: restricted model
## Model 2: bwght ~ faminc + cigs
##
```

```
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1387 574612
## 2   1385 557486  2     17126 21.274 7.942e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We can adjust this for heteroskedasticity

```
library(car)
data <- bwght
reg <- lm(bwght ~faminc + cigs, data = data)
linearHypothesis(reg, c("faminc=0","cigs=0"), white.adjust="hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## faminc = 0
## cigs = 0
##
## Model 1: restricted model
## Model 2: bwght ~ faminc + cigs
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1   1387
## 2   1385  2 22.112 3.524e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```