Regression Analysis: Inference

Statistics and Econometrics

Jiahua Wu

Testing hypotheses about a single population parameter

Example 4.1

Testing a simple null hypothesis is straightforward in R, as the default R output provides the t statistic and p-value for $H_0: \beta_j = 0$ in the columns of "t value" and "Pr>|t|", respectively, assuming a two-sided alternative.

```
load("wage1.RData")
wage.m1 <- lm(log(wage) ~ educ + exper + tenure, data = data)
summary(wage.m1)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper + tenure, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05802 -0.29645 -0.03265  0.28788  1.42809
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.284360   0.104190   2.729  0.00656 **
## educ        0.092029   0.007330  12.555  < 2e-16 ***
## exper       0.004121   0.001723   2.391  0.01714 *
## tenure      0.022067   0.003094   7.133 3.29e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4409 on 522 degrees of freedom
## Multiple R-squared:  0.316,  Adjusted R-squared:  0.3121
## F-statistic: 80.39 on 3 and 522 DF,  p-value: < 2.2e-16
```

If we ever need to run the hypothesis testing manually, then remember that the t statistic is the ratio between point estimate and standard error for the simple null hypothesis. We can find critical value using *qt* or *qnorm* functions. For instance,

```
# find the critical value for 99.5th percentile from a standard norm distribution
qnorm(0.995)
```

```
## [1] 2.575829
```

```
# find the critical value for 99.5th percentile from a t distribution with df = 522
qt(0.995, df = 522)
```

```
## [1] 2.58528
```

In general, *linearHypothesis* in the *car* package is the function to use for hypothesis testing in R. For instance, if we want to test the simple null hypothesis that $H_0: \beta_{exper} = 0$, we can type the following command

```r
linearHypothesis(wage.m1, "exper = 0")
```

```
## Linear hypothesis test
##
## Hypothesis:
## exper = 0
##
## Model 1: restricted model
## Model 2: log(wage) ~ educ + exper + tenure
##
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1    523 102.57
## 2    522 101.46  1    1.1115 5.719 0.01714 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$linearHypothesis$ is implemented based on F test, rather than the usual t test for the simple null hypothesis testing. However, p-value from $linearHypothesis$ is the same as the p-value from a standard t test, assuming a two-sided alternative. In this test, p-value is 0.01714, so we can reject null at 5% significance level but not at 1% significance level.

We can also use $linearHypothesis$ to test a more general form of t test, where the null is $H_0 : \beta_j = a_j$.

```r
linearHypothesis(wage.m1, "exper = 1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## exper = 1
##
## Model 1: restricted model
## Model 2: log(wage) ~ educ + exper + tenure
##
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    523 65011
## 2    522   101  1     64909 333966 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Confidence interval

The built-in function for calculating confidence interval is $confint$.

```r
# calculate 95% confidence interval for the variable educ
confint(wage.m1, 'educ', level = 0.95)
```

```
##           2.5 %    97.5 %
## educ 0.07762921 0.1064288
```

```r
# calculate 95% confidence interval for all parameters in the linear model wage.m1
confint(wage.m1, level = 0.95)
```

```
##                     2.5 %     97.5 %
## (Intercept) 0.0796755842 0.48904353
```

```
## educ        0.0776292137 0.10642876
## exper       0.0007356983 0.00750652
## tenure      0.0159896850 0.02814475
```

Testing a linear combination of parameters

Again, *linearHypothesis* function can help us to test a linear combination of parameters. For instance to test the hypothesis $H_0 : \beta_{educ} - \beta_{exper} = 0$ on slide 33, we can use the following code.

```
linearHypothesis(wage.m1, "educ - exper = 0")
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ - exper = 0
##
## Model 1: restricted model
## Model 2: log(wage) ~ educ + exper + tenure
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    523 132.25
## 2    522 101.46  1    30.798 158.46 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing multiple linear restrictions (Online Material Session 2.6)

Example 4.9

We can use F test for testing exclusion restrictions. SSRs from both restricted and unrestricted models will be used to calculate F statistic. One thing to keep in mind is that we need to take care of missing values in the sample. The exact same sample shall be used to estimate both restricted and unrestricted models for a valid F statistic. For instance, there are missing values for $motheduc$ and $fatheduc$ in this example. Thus we need to remove the observations with missing values before running regressions.

```
load("bwght.RData")
summary(data)
```

```
##      faminc          cigtax         cigprice         bwght
##  Min.   : 0.50   Min.   : 2.00   Min.   :103.8   Min.   : 23.0
##  1st Qu.:14.50   1st Qu.:15.00   1st Qu.:122.8   1st Qu.:107.0
##  Median :27.50   Median :20.00   Median :130.8   Median :120.0
##  Mean   :29.03   Mean   :19.55   Mean   :130.6   Mean   :118.7
##  3rd Qu.:37.50   3rd Qu.:26.00   3rd Qu.:137.0   3rd Qu.:132.0
##  Max.   :65.00   Max.   :38.00   Max.   :152.5   Max.   :271.0
##
##     fatheduc        motheduc         parity           male
##  Min.   : 1.00   Min.   : 2.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:12.00   1st Qu.:12.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :12.00   Median :12.00   Median :1.000   Median :1.0000
##  Mean   :13.19   Mean   :12.94   Mean   :1.633   Mean   :0.5209
##  3rd Qu.:16.00   3rd Qu.:14.00   3rd Qu.:2.000   3rd Qu.:1.0000
```

```
##  Max.   :18.00   Max.   :18.00   Max.   :6.000   Max.   :1.0000
##  NA's   :196     NA's   :1
##      white             cigs            lbwght         bwghtlbs
##  Min.   :0.0000   Min.   : 0.000   Min.   :3.135   Min.   : 1.438
##  1st Qu.:1.0000   1st Qu.: 0.000   1st Qu.:4.673   1st Qu.: 6.688
##  Median :1.0000   Median : 0.000   Median :4.787   Median : 7.500
##  Mean   :0.7846   Mean   : 2.087   Mean   :4.760   Mean   : 7.419
##  3rd Qu.:1.0000   3rd Qu.: 0.000   3rd Qu.:4.883   3rd Qu.: 8.250
##  Max.   :1.0000   Max.   :50.000   Max.   :5.602   Max.   :16.938
##
##      packs           lfaminc
##  Min.   :0.0000   Min.   :-0.6931
##  1st Qu.:0.0000   1st Qu.: 2.6741
##  Median :0.0000   Median : 3.3142
##  Mean   :0.1044   Mean   : 3.0713
##  3rd Qu.:0.0000   3rd Qu.: 3.6243
##  Max.   :2.5000   Max.   : 4.1744
##
```

```r
# remove observations with missing motheduc and fatheduc
data.new <- na.omit(data)
bwght.ur <- lm(bwght ~ cigs + parity + faminc + motheduc + fatheduc, data = data.new)
ur.res <- sum(bwght.ur$residuals^2)

bwght.r <- lm(bwght ~ cigs + parity + faminc, data = data.new)
r.res <- sum(bwght.r$residuals^2)

# calculate F statistic
F.stat <- (r.res - ur.res)/2 / (ur.res/(bwght.ur$df.residual))
F.stat
```

```
## [1] 1.437269
```

```r
# calculate p value
pf(F.stat, 2, bwght.ur$df.residual, lower.tail = FALSE)
```

```
## [1] 0.2379896
```

```r
# Alternatively, we can test it using linearHypothesis
linearHypothesis(bwght.ur, c("motheduc = 0", "fatheduc = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## motheduc = 0
## fatheduc = 0
##
## Model 1: restricted model
## Model 2: bwght ~ cigs + parity + faminc + motheduc + fatheduc
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1187 465167
## 2   1185 464041  2    1125.7 1.4373  0.238
```