Inference with Heteroskedasticity

R Tutorials for Applied Statistics

Note on required packages: The following code requires the packages `lmtest`, `sandwich`, and `tidyverse`. The `sandwich` package contains procedures to estimate regression error variance that may change with the explanatory variables. The `lmtest` package contains procedures to conduct hypothesis tests when there is heteroskedasticity. If you have not done so, download and install the package `lmtest`, `sandwich`, and `tidyverse`. When the packages are installed, load these libraries.

```
# This only needs to be executed once for your machine
install.packages("lmtest")

# This only needs to be executed once for your machine
install.packages("sandwich")

# This only needs to be executed once for your machine
install.packages("tidyverse")

# This needs to be executed every time you load R
library("lmtest")

# This needs to be executed every time you load R
library("sandwich")

# This needs to be executed every time you load R
library("tidyverse")
```

---

1 Introduction

Homoskedasticity is the property that the variance of the error term of a regression (estimated by the variance of the residual in the sample) is the same across different values for the explanatory variables, or the same across time for time series models.

Heteroskedasticity is the property when the variance of the error term changes predictably with one or more of the explanatory variables.

2 Example: Factors affecting monthly earnings

Let us examine a data set that explores the relationship between total monthly earnings (`MonthlyEarnings`) and a number of factors that may influence monthly earnings including including each person's IQ (`IQ`), a measure of knowledge workplace environment (`Knowledge`), years of education (`YearsEdu`), years experience (`YearsExperience`), and years at current job (`Tenure`).

The code below downloads a CSV file that includes data on the above variables from 1980 for 663 individuals and assigns it to a data set called `df`.

```
load(url("http://murraylax.org/datasets/wage2.RData"))
```

The following call to `lm()` estimates a multiple regression predicting monthly earnings based on the five explanatory variables given above. The call to `summary()` displays some summary statistics from the regression.

```
lmwages <- lm(MonthlyEarnings
              ~ IQ + Knowledge + YearsEdu + YearsExperience + Tenure,
              data = df)
```

3 Heteroskedasticity robust standard errors

White's correction for heteroskedasticity is a method for using OLS estimates for coefficients and predicted values (which are unbiased), but fixing the estimates for the variances of the coefficients (which are biased). It uses as an estimate for the *possibly changing* variance the squared residuals estimated from OLS which we computed and graphed above.

In R, we can compute the White heteroskedastic variance/covariance matrix for the coefficients with the call below to `vcovHC` from the `sandwich` package. VCOVHC stands for Variance / Covariance Heteroskedastic Consistent.

```
vv <- vcovHC(lmwages, type="HC1")
```

The first parameter in the call above is our original output from our call to `lm()` above. The second parameter `type="HC1"` tells the function to use the White estimate for the variance covariance matrix which uses as an estimate for the changing variance the squared residuals from the OLS call.

4 Hypothesis testing

We can use our estimate for the variance / covariance to properly compute our standard errors, t-statistics, and p-values for the coefficients:

```
coeftest(lmwages, vcov = vv)
```

```
##
## t test of coefficients:
##
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     -512.5397   135.5678 -3.7807 0.0001706 ***
## IQ                 3.8887     1.1352  3.4256 0.0006517 ***
## Knowledge          8.0476     2.4932  3.2278 0.0013093 **
## YearsEdu          46.3084     8.6877  5.3303 1.349e-07 ***
## YearsExperience   13.4454     4.1497  3.2401 0.0012550 **
## Tenure             3.3915     3.0319  1.1186 0.2637138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first parameter to `coeftest` is the result of our original call to `lm()` and the second parameter is the updated variance / covariance matrix to use.

Let's compare the result to the OLS regression output:

```
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -849.51 -244.91  -41.28  191.41 2225.88
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -512.540    139.180  -3.683 0.000250 ***
## IQ                 3.889      1.204   3.230 0.001299 **
## Knowledge          8.048      2.246   3.582 0.000366 ***
## YearsEdu          46.308      8.833   5.243 2.13e-07 ***
## YearsExperience   13.445      4.064   3.309 0.000989 ***
## Tenure             3.392      3.016   1.124 0.261262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 369.6 on 657 degrees of freedom
## Multiple R-squared:  0.1796, Adjusted R-squared:  0.1733
## F-statistic: 28.76 on 5 and 657 DF,  p-value: < 2.2e-16
```

The coefficients are exactly the same, but the estimates for the standard errors, the t-statistics, and p-values are slightly different.

One reason the standard errors and p-values are only slightly different is that there may actually be very little heteroskedasticity. The differences in variance of the error term may be small even for large differences in the predicted value for xx.

The White heteroskedastic robust standard errors are valid for **either homoskedasticity or heteroskedasticity**, so it is always safe to use these estimates if you are not sure if the homoskedasticity assumption holds.

5 Confidence Intervals

The function `confint()` computes confidence intervals for the coefficients, but assumes homoskedasticity. Here is an example:

```
confint(lmwages, conf.level=0.95)
```

```
##                       2.5 %      97.5 %
## (Intercept)     -785.831826 -239.247619
## IQ                 1.524802    6.252556
## Knowledge          3.636458   12.458672
## YearsEdu          28.964873   63.651868
## YearsExperience    5.465652   21.425242
## Tenure            -2.531316    9.314351
```

Unfortunately, there R includes no similar method to compute confidence intervals for coefficients with heteroskedastic-robust standard errors. You can compute the confidence intervals manually based on the estimates of the variances from the `vcovHC()` function. I created such a function, which you can download and load into memory using the following call:

```
source(url("https://murraylax.org/code/R/confintHC.R"))
```

We now have a *function* in our environment called `confintHC()`. We use this function in the code below to estimate 95% confidence intervals on the coefficients using the White heteroskedasticity estimates for the variances.

```
confintHC(lmwages, type="HC1", conf.level=0.95)
```

```
##                 Coefficient  2.5% Limit 97.5% Limit
## (Intercept)     -512.539723 -778.738116 -246.341329
## IQ                 3.888679    1.659646    6.117712
## Knowledge          8.047565    3.152031   12.943099
## YearsEdu          46.308370   29.249347   63.367394
## YearsExperience   13.445447    5.297253   21.593641
## Tenure             3.391518   -2.561851    9.344887
```

6 Joint Test for Multiple Restrictions

Let us allow for an interactions effect between IQ and years of experience and IQ and years of education. The interaction effect with years experience allow for the possibility that people with a higher IQ get a greater return to years experience, possibly because people with a higher IQ learn more with each additional year of experience. The interaction effect with years education allows for the possibility that people with a higher IQ get greater benefits for each additional year of education.

```
lmwages <- lm(MonthlyEarnings
              ~ IQ + Knowledge + YearsEdu + YearsExperience + Tenure +
                IQ:YearsExperience + IQ:YearsEdu,
              data = df)
```

Now suppose we want to test whether or not IQ has any influence on monthly earnings. The variable IQ appears multiple times: once on its own and twice in interaction effects. The null and alternative hypotheses for this test is given by,

$$H_0 : \beta_{IQ} = 0, \quad \beta_{IQxYearsEdu} = 0, \quad \beta_{IQxYearsExperience} = 0$$

H0:βIQ=0, βIQxYearsEdu=0, βIQxYearsExperience=0

$$H_A : \text{At least one of these is not equal to zero}$$

HA:At least one of these is not equal to zero

The standard F-test for multiple restrictions that compares the sum of squared explained between an unrestricted and (nested) restricted model (often called the *Wald test*) *assumes homoskedasticity*, and so this test is not appropriate if heteroskedasticity is present.

There is a heteroskedasticity-robust version of the test which can be estimated with the `waldtest()` function. The default output of `waldtest()` is the same as `anova()`, but it has the additional power that we can specify the correct variance / covariance matrix.

First we estimate the restricted model, which does not include IQ in any of the terms:

```
lmwages_res <- lm(MonthlyEarnings ~ Knowledge + YearsEdu + YearsExperience + Tenure, data = d
```

Then we call `waldtest()`, passing both the restricted and unrestricted regression outputs, and we include our estimate for the variance / covariance robust estimator for the unrestricted model:

```
vv <- vcovHC(lmwages, type="HC1")
waldtest(lmwages, lmwages_res, vcov=vv)
```

```
## Wald test
##
## Model 1: MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure + IQ:YearsExperience + IQ:YearsEdu
## Model 2: MonthlyEarnings ~ Knowledge + YearsEdu + YearsExperience + Tenure
##   Res.Df Df      F   Pr(>F)
## 1    655
## 2    658 -3 4.0298 0.007423 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 0.007 which is below 0.05. We reject the null hypothesis and conclude that there is sufficient statistical evidence that IQ does help explain monthly earnings.