Heteroskedasticity —Chapter 8 of Wooldridge's textbook

(Jing Li, Miami University)

Big Picture

In this lecture you will learn

1. Homoskedasticity and Heteroskedasticity
2. Robust Standard Error
3. Weighted Least Squares Estimator

Homoskedasticity

Consider a simple regression

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad (i = 1, \ldots, n) \tag{1}$$

Homoskedasticity is the assumption that the variance of u is constant

$$var(u_i|x_i) = var(y_i|x_i) = \sigma^2 = \texttt{constant} \qquad (\texttt{Homoskedasticity}) \tag{2}$$

Under the homoskedasticity assumption we can show that

1. the conventional standard error of $\hat{\beta}_1$ is

$$se(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{TSS_x}} = \sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}} \tag{3}$$

   By default R function **lm** uses formula (3) to compute the standard error, t value, p-value and etc. They are all wrong if homoskedasticity fails.

2. OLS is the best unbiased linear estimator (BLUE), a result called Gauss-Markov Theorem

Heteroskedasticity

1. We can imagine a situation in which homoskedasticity is invalid (see the picture in next page).

2. Heteroskedasticity is present when the variance of error term varies across observations

$$var(u_i|x_i) = var(y_i|x_i) = h(x_i) = \sigma_i^2 \neq \texttt{constant} \qquad (\texttt{Heteroskedasticity}) \quad (4)$$

Under heteroskedasticity we can show that

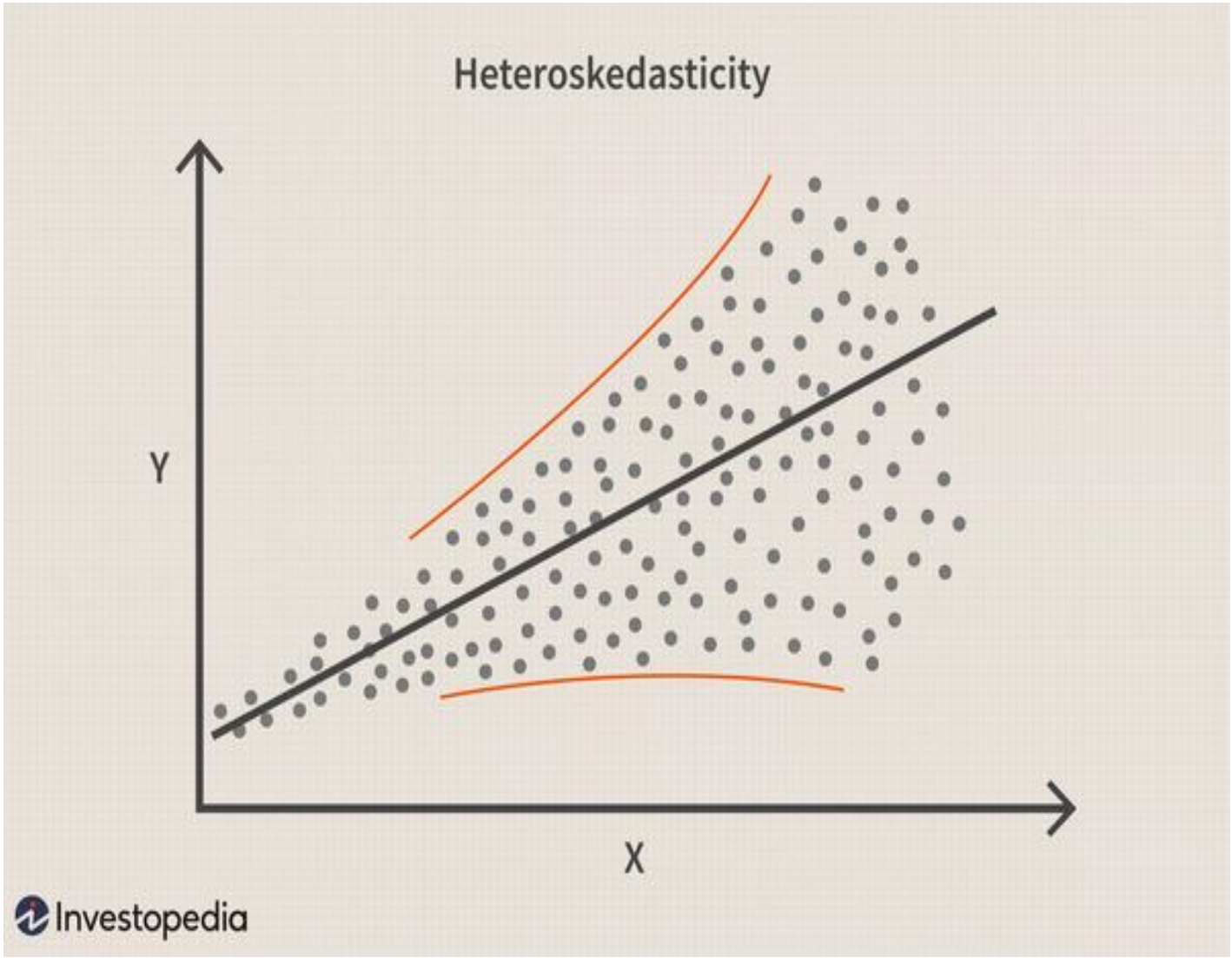
(a) the new standard error of the slope coefficient estimate is

$$se(\hat{\beta}_1) = \sqrt{\frac{\sum(x_i - \bar{x})^2 \sigma_i^2}{(TSS_x)^2}} \quad (\texttt{heteroskedasticity-robust standard error}) \quad (5)$$

R function **coeftest** uses formula (5) to compute the heteroskedasticity-robust standard error, t value, p-value and etc. They are correct no matter whether homoskedasticity holds.

(b) OLS is no longer BLUE. Instead Weighted Least Squares (WLS) is BLUE

A Picture of Heteroskedasticity



Example 1

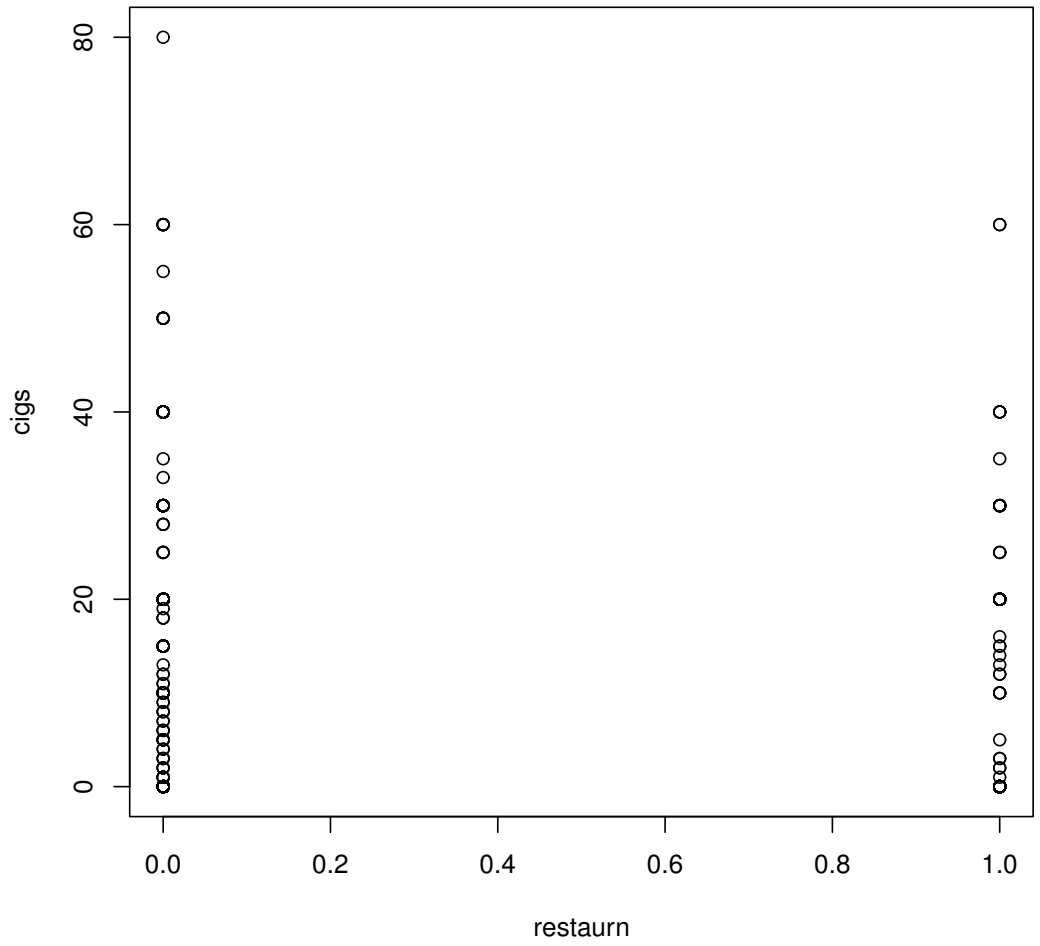
```
> rm(list = ls())
> ad = "https://www.fsb.miamioh.edu/lij14/411_smoke.txt"
> data = read.table(url(ad), header=T)
> attach(data)
> summary(lm(cigs~lincome+lcigpric+educ+age+agesq+restaurn))$coef
               Estimate   Std. Error    t value     Pr(>|t|)
(Intercept) -3.6398415 24.078660000 -0.1511646 8.798840e-01
lincome      0.8802682  0.727783176  1.2095198 2.268205e-01
lcigpric    -0.7508586  5.773342707 -0.1300561 8.965547e-01
educ        -0.5014982  0.167077202 -3.0015959 2.769188e-03
age          0.7706936  0.160122326  4.8131549 1.776147e-06
agesq       -0.0090228  0.001743033 -5.1764942 2.861971e-07
restaurn    -2.8250848  1.111793538 -2.5410157 1.124091e-02
```

Remarks

1. We use Smoke data to illustrate heteroskedasticity
2. We regress cigs (number of cigarettes) onto lincome (log income), lcigpric (log price), education, age, age squared and a dummy variable that equals one when there is a smoking ban in restaurants
3. We find neither income nor price matters (t values = 1.21, -0.13). This unexpected result raises a red flag. Something may be wrong
4. R command **lm** uses formula (3) to compute standard error, t value, and p-value. They are all wrong if homoskedasticity fails. We are unsure if there is heteroskedasticity

Example 1—continued

```
> plot(cigs~restaurn)
```



Informal Check I of Homoskedasticity

1. We next consider an informal check (eye-ball econometrics) of the homoskedasticity assumption
2. We plot cigs against smoking-ban dummy variable
3. We find that the conditional distribution of cigs when restaurn=0 is wider or has greater dispersion than the conditional distribution when restaurn=1
4. This finding indicates that the variance at least depends on restaurn, so is not constant. In short, it is very likely that for this problem homoskedasticity fails and heteroskedasticity is present

Example 1—continued

```
> sd(cigs[restaurn==0])
[1] 14.21497
> sd(cigs[restaurn==1])
[1] 11.88069
```

1. We compare the conditional standard deviation of cigs when restaurn=0 to when restaurn=1
2. It's clear the two conditional standard deviations 14.2149 and 11.8806 are different. Again, this finding suggests possible heterskedasticity
3. This method is informal because we do not know whether that difference is statistically significant

Breusch-Pagan (BP) Test for Homoskedasticity

1. BP Test is the formal way to test the null hypothesis of homoskedasticity
2. BP tests involves three steps
   (a) Step I: run original regression and save residual $\hat{u}$

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \texttt{error term}$$

   (b) Step II: run an auxiliary regression given by

$$\hat{u}^2 = c_0 + c_1 x_1 + \ldots c_k x_k + \texttt{error term} \tag{6}$$

   Note the dependent variable is squared residual

   (c) Step III: compute

$$\texttt{BP Test} = nR^2_{\texttt{StepIIregression}} \sim \chi^2(k) \tag{7}$$

   We reject homoskedasticity if the p-value is less than 0.05

3. Intuitively, we reject homoskedasticity when the auxiliary regression has a big R-squared, that is, when some of x variables can explain or matters for variance

Example 1—continued

```
> m1 = lm(cigs~lincome+lcigpric+educ+age+agesq+restaurn)
> uhat = resid(m1)
> uhatsq = uhat^2
> summary(lm(uhatsq~lincome+lcigpric+educ+age+agesq+restaurn))$coef
                Estimate  Std. Error    t value     Pr(>|t|)
(Intercept) -636.3030617 652.4945618 -0.9751852 3.297632e-01
lincome       24.6384840  19.7218020  1.2493019 2.119201e-01
lcigpric      60.9765528 156.4486861  0.3897543 6.968220e-01
educ          -2.3842257   4.5275346 -0.5266057 5.986134e-01
age           19.4174795   4.3390682  4.4750344 8.747747e-06
agesq         -0.2147895   0.0472335 -4.5473978 6.271968e-06
restaurn     -71.1813778  30.1278907 -2.3626406 1.838368e-02
```

Remarks

1. We use command **resid** to save the residual of original regression
2. Next we generate squared residual, the dependent variable in the auxiliary regression
3. The auxiliary regression (6) indicates that age, age squared and restaurn are significant (t values = 4.48, -4.55, -2.36). So variances change when those variables change

Example 1—continued

```
> library(lmtest)
> bptest(m1)
studentized Breusch-Pagan test
data:  m1
BP = 32.258, df = 6, p-value = 1.456e-05
```

1. The BP test follows chi-squared distribution with 6 degrees of freedom. For this problem BP test is 32.25
2. The p value is less than 0.05, so we reject homoskedasticity
3. We conclude that for this problem homoskedasticity is invalid and heterskedasticity is present

Road Map

1. If homoskedasticity is not rejected by BP test, we can trust the result reported by R function **lm**
2. If homoskedasticity is rejected, there are two options
   (a) Option A: we still use OLS, but we must use R function **coeftest** to report the heteroskedasticity-robust standard error, t value and p value
   (b) Option B: we use WLS, which is more efficient than OLS in the presence of heteroskedasticity

Example 1—continued

```
> library("lmtest")
> library("sandwich")
> coeftest(m1, vcov = vcovHC(m1, type = "HC0"))
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -3.6398415 25.5051188 -0.1427  0.886555
lincome      0.8802682  0.5934209  1.4834  0.138368
lcigpric    -0.7508586  6.0091689 -0.1250  0.900593
educ        -0.5014982  0.1616877 -3.1016  0.001992 **
age          0.7706936  0.1376831  5.5976 2.987e-08 ***
agesq       -0.0090228  0.0014558 -6.1977 9.167e-10 ***
restaurn    -2.8250848  1.0036512 -2.8148  0.005001 **
```

Remarks

1. With function **coeftest**, R reports heteroskedasticity-robust standard error, t value, and p value
2. For instance, the standard error of coefficient of lincome is 0.7277 based on formula (3), the robust standard error based on formula (5) is 0.5934
3. We see t values and p values change accordingly
4. Note that the coefficient estimates remain unchanged—heteroskedasticity has nothing to do with $\hat{\beta}$, it matters for $var(\hat{\beta})$

Weighted Least Squares (WLS)

1. We consider weighted least squares method if there is heteroskedasticity
2. Consider a simple regression

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad (var(u_i|x_i) = h(x_i)) \tag{8}$$

The idea of WLS is to transform the model so that the new error is homoskedastic

$$\frac{y_i}{\sqrt{h(x_i)}} = \beta_0 \frac{1}{\sqrt{h(x_i)}} + \beta_1 \frac{x_i}{\sqrt{h(x_i)}} + \frac{u_i}{\sqrt{h(x_i)}} \tag{9}$$

$$var\left(\frac{u_i}{\sqrt{h(x_i)}}|x_i\right) = 1 = \texttt{constant} \tag{10}$$

3. WLS is just the OLS applied to the transformed regression (9). Because the error is now homoskedastic, WLS is BLUE
4. $\frac{1}{\sqrt{h(x_i)}}$ is called weight. In practice we need to estimate $h(x_i)$ before applying WLS. The details are in the textbook

Example 1—continued

```
> luhatsq = log(uhatsq)
> m3 = lm(luhatsq~lincome+lcigpric+educ+age+agesq+restaurn)
> ghat = fitted(m3)
> hhat = exp(ghat)
> we = 1/hhat
> wls_m = lm(cigs~lincome+lcigpric+educ+age+agesq+restaurn, weights=we)
> summary(wls_m)$coef
               Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)  5.63546183 1.780314e+01  0.3165432 7.516728e-01
lincome      1.29523990 4.370118e-01  2.9638560 3.128261e-03
lcigpric    -2.94031229 4.460144e+00 -0.6592415 5.099304e-01
educ        -0.46344637 1.201587e-01 -3.8569532 1.240386e-04
age          0.48194788 9.680823e-02  4.9783772 7.855907e-07
agesq       -0.00562721 9.394801e-04 -5.9897061 3.174847e-09
restaurn    -3.46106414 7.955050e-01 -4.3507763 1.532145e-05
```

Remarks

1. The details of obtaining WLS can be found in the textbook
2. In this case, log income becomes significant (t-value=2.964) after we apply WLS
3. This finding confirms that WLS can be more efficient (have smaller standard error) than OLS in the presence of heteroskedasticity