Heteroskedasticity[1]

1. Consequences of Heteroskedasticity
2. Heteroskedasticity-Robust Inference
3. Testing for Heteroskedasticity
4. Weighted Least Squares Estimation

[1] Wooldridge, Chapter 8.

Homoskedasticity fails whenever the variance of the unobserved factors changes across different segments of the population, where the segments are determined by the different values of the explanatory variables.

In a savings equation, for example, heteroskedasticity is present if the variance of the unobserved factors affecting savings increases with income.

Homoskedasticity is needed to justify the usual t tests, F tests, and confidence intervals for OLS estimation of the linear regression model, even with large sample sizes.

Heteroskedasticity:

- Consequences for ordinary least squares estimation,
- Available remedies when heteroskedasticity occurs, and
- Test for its presence.

Consequences for OLS

Consider the multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon.$$

The OLS estimators $\hat{\beta}_0, \ldots, \hat{\beta}_k$ are unbiasedness and consistent, under the first four Gauss-Markov assumptions.

The homoskedasticity assumption

$$V(\varepsilon | x_1, \ldots, x_k) = \sigma^2,$$

plays no role in showing whether OLS was unbiased or consistent.

If heteroskedasticity does not cause bias or inconsistency in the OLS estimators, why did we introduce it as one of the Gauss-Markov assumptions?

The estimators of the variances, $V(\hat{\beta}_j)$, are biased without the homoskedasticity assumption.

Since the OLS standard errors are based directly on these variances, they are no longer valid for constructing confidence intervals and t statistics.

The usual OLS t statistics do not have t distributions in the presence of heteroskedasticity, and the problem is not resolved by using large sample sizes.

In summary, the statistics we used to test hypotheses under the Gauss-Markov assumptions are not valid in the presence of heteroskedasticity.

We will show how the usual OLS test statistics can be modified so that they are valid, at least asymptotically.

Consider the model with a single independent variable, where the first four Gauss-Markov assumptions hold.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

If the errors contain heteroskedasticity

$$V(\varepsilon_i | x_i) = \sigma_i^2,$$

and knowing that

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

it follows that

$$V(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma_i^2}{\{\sum_{i=1}^{n}(x_i - \bar{x})^2\}^2}$$

White (1980) showed that a valid estimator of $V(\hat{\beta}_1)$, for heteroskedasticity of any form is

$$\widehat{V(\hat{\beta}_1)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \hat{\varepsilon}_i^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}.$$

In what sense is this a valid estimator of $V(\hat{\beta}_1)$?
The law of large numbers (LLN) and the central limit theorem (CLT) play key roles in establishing its validity.

A similar formula works in the general multiple regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i.$$

It can be shown that a valid estimator of $V(\hat{\beta}_j)$, under Assumptions MLR.1 through MLR.4, is

$$\widehat{V(\hat{\beta}_j)} = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{\varepsilon}_i^2}{\left\{\sum_{i=1}^{n} \hat{r}_{ij}^2\right\}^2},$$

where $\hat{r}_{1j}, \ldots, \hat{r}_{nj}$ are the residuals from regressing x_j on all other independent variables.

$\sqrt{\widehat{V(\hat{\beta}_j)}}$ is the heteroskedasticity-robust standard error for $\hat{\beta}_j$ (White, 1980).

wage1.csv

The data-set wage1.csv was provided with: Wooldridge, Jeffrey M. (200x), Introductory Econometrics: A Modern Approach, x. Edition, South Western College Publishing, Mason (Ohio). (Note: x stands for different years/editions) These are data from the 1976 Current Population Survey, collected by Henry Farber and contain the following variables (Obs. 526):

1.	wage	average hourly earnings
2.	educ	years of education
3.	exper	years potential experience
4.	tenure	years with current employer
5.	nonwhite	=1 if nonwhite
6.	female	=1 if female
7.	married	=1 if married
8.	numdep	number of dependents
9.	smsa	=1 if live in SMSA
10.	northcen	=1 if live in north central U.S
11.	south	=1 if live in southern region
12.	west	=1 if live in western region
13.	construc	=1 if work in construc. indus.
14.	ndurman	=1 if in nondur. manuf. indus.
15.	trcommpu	=1 if in trans, commun, pub ut
16.	trade	=1 if in wholesale or retail
17.	services	=1 if in services indus.
18.	profserv	=1 if in prof. serv. indus.
19.	profocc	=1 if in profess. occupation
20.	clerocc	=1 if in clerical occupation
21.	servocc	=1 if in service occupation
22.	lwage	log(wage)
23.	expersq	exper^2
24.	tenursq	tenure^2

R CODE

Wage differences: married men/women, single men/women.
Dependent variable is `lwage`, $n = 526$ and $R^2 = 0.461$

```
data = read.csv("wage1.csv",header=TRUE)
attach(data)
n = nrow(data)

# Dummy variables
marrmale = rep(0,n)
marrfem = rep(0,n)
singfem = rep(0,n)
marrmale[(female==0)&(married==1)]=1
marrfem[(female==1)&(married==1)]=1
singfem[(female==1)&(married==0)]=1

# Multiple regression
X = cbind(1,marrmale,marrfem,singfem,educ,exper,expersq,tenure,tenursq)
reg  = lm(lwage~X-1)
summary(reg)

# Heterokedasticity-robust standard errors
se = rep(0,ncol(X))
i=1
reg1 = lm(X[,1]~X[,-1]-1)
se[i] = sqrt(sum((reg1$res^2)*(reg$res^2))/(sum(reg1$res^2))^2)
for (i in 2:9){
  reg1 = lm(X[,i]~X[,-i])
  se[i] = sqrt(sum((reg1$res^2)*(reg$res^2))/(sum(reg1$res^2))^2)
}
```

Standard errors

coefficient	estimate	OLS s.e.	HR s.e.
intercept	0.3214	0.100009	0.108528
`marrmale`	0.2127	0.055357	0.056651
`marrfem`	-0.1983	0.057836	0.058265
`singfem`	-0.1104	0.055742	0.056626
`educ`	0.0789	0.006695	0.007351
`exper`	0.0268	0.005243	0.005095
`tenure`	0.0291	0.006762	0.006881
`expersq`	-0.00054	0.000110	0.000105
`tenuresq`	-0.00053	0.000231	0.000242

HR F statistic

The HR standard errors provide a method for computing t statistics that are asymptotically t distributed.

Testing

$$H_0 : V(\varepsilon | x_1, x_2, \ldots, x_k) = \sigma^2$$

is the same as testing

$$H_0 : E(\varepsilon^2 | x_1, x_2, \ldots, x_k) = \sigma^2$$

This shows that, in order to test for violation of the homoskedasticity assumption, we want to test whether ε^2 is related (in expected value) to one or more of the explanatory variables.

If H_0 is false, the expected value of ε^2, given the independent variables, can be virtually any function of the x_j.

A simple approach is to assume a linear function:

$$\varepsilon^2 = \delta_0 + \delta_1 x_1 + \cdots \delta_k x_k + \nu,$$

so the null hypothesis of homoskedasticity is

$$H_0 : \delta_1 = \delta_2 = \cdots = \delta_k = 0.$$

The F statistic depend on the $R^2_{\hat{\varepsilon}^2}$ from regression

$$\hat{\varepsilon}^2 = \delta_0 + \delta_1 x_1 + \cdots \delta_k x_k + \text{error},$$

and is computed as

$$F = \frac{R^2_{\hat{\varepsilon}^2}/k}{(1 - R^2_{\hat{\varepsilon}^2})/(n-k-1)}.$$

This F statistic has (approximately) an $F_{k,n-k-1}$ distribution under the null hypothesis of homoskedasticity.

hprice1.txt

```
Contains data from hprice1.txt
  obs:            88
 vars:            10
-------------------------------------------------------------
variable        variable label
-------------------------------------------------------------
price           house price, $1000s
assess          assessed value, $1000s
bdrms           number of bdrms
lotsize         size of lot in square feet
sqrft           size of house in square feet
colonial        =1 if home is colonial style
lprice          log(price)
lassess         log(assess
llotsize        log(lotsize)
lsqrft          log(sqrft)
-------------------------------------------------------------
```

R CODE

```
data = read.table("hprice1.txt",header=TRUE)
attach(data)
n = nrow(data)

reg1 = lm(price ~ lotsize+sqrft+bdrms)
reg2 = lm(lprice ~ llotsize+lsqrft+bdrms)

summary(reg1)
summary(reg2)

e1sq = reg1$res^2
e2sq = reg2$res^2

R2.e1 = summary(lm(e1sq~lotsize+sqrft+bdrms))$r.sq
R2.e2 =summary(lm(e2sq~llotsize+lsqrft+bdrms))$r.sq

F1 = R2.e1/(1-R2.e1)*(84/3)
F2 = R2.e2/(1-R2.e2)*(84/3)

pval1 = 1-pf(F1,3,84)
pval2 = 1-pf(F2,3,84)

rbind(c(R2.e1,R2.e2),
c(F1,F2),
c(pval1,pval2))
```

Regression on levels

```
> summary(reg1)
lm(formula = price ~ lotsize + sqrft + bdrms)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.177e+01  2.948e+01  -0.739  0.46221
lotsize      2.068e-03  6.421e-04   3.220  0.00182 **
sqrft        1.228e-01  1.324e-02   9.275 1.66e-14 ***
bdrms        1.385e+01  9.010e+00   1.537  0.12795
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 59.83 on 84 degrees of freedom
Multiple R-squared:  0.6724,Adjusted R-squared:  0.6607
F-statistic: 57.46 on 3 and 84 DF,  p-value: < 2.2e-16
```

Computing errors

$$\hat{\varepsilon} = \texttt{price} + 21.77 - 0.00207\texttt{lotsize} - 0.123\texttt{sqrft} - 13.85\texttt{bdrms},$$

and fitting

$$\hat{\varepsilon}^2 = \delta_0 + \delta_1\texttt{lotsize} + \delta_2\texttt{sqrft} + \delta_3\texttt{bdrms} + \nu,$$

leads to $R^2_{\hat{\varepsilon}^2} = 0.160140744$.

The HR F statistic for the null hypothesis

$$H_0 : \delta_1 = \delta_2 = \delta_3 = 0,$$

is

$$F = \frac{0.160140744/3}{0.8398593/84} = 5.338919368,$$

with p-value of 0.002047744.

Conclusion: There is strong evidence against the null hypothesis.

Regression on logs

```
> summary(reg2)
lm(formula = lprice ~ llotsize + lsqrft + bdrms)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.29704    0.65128  -1.992   0.0497 *
llotsize     0.16797    0.03828   4.388 3.31e-05 ***
lsqrft       0.70023    0.09287   7.540 5.01e-11 ***
bdrms        0.03696    0.02753   1.342   0.1831
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1846 on 84 degrees of freedom
Multiple R-squared:  0.643,Adjusted R-squared:  0.6302
F-statistic: 50.42 on 3 and 84 DF,  p-value: < 2.2e-16
```

Computing errors

$$\hat{\varepsilon} = \texttt{price} + 1.30 - 0.168\texttt{llotsize} - 0.700\texttt{lsqrft} - 0.037\texttt{bdrms},$$

and fitting

$$\hat{\varepsilon}^2 = \delta_0 + \delta_1\texttt{llotsize} + \delta_2\texttt{lsqrft} + \delta_3\texttt{bdrms} + \nu,$$

leads to $R^2_{\hat{\varepsilon}^2} = 0.04799136$.

The HR F statistic for the null hypothesis

$$H_0 : \delta_1 = \delta_2 = \delta_3 = 0,$$

is

$$F = \frac{0.04799136/3}{0.9520086/84} = 1.41149767,$$

with p-value of 0.24514631.

Conclusion: There is not strong evidence against the null hypothesis, so we fail to reject the null.

Known heteroskedasticity

Suppose that

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

where

$$\begin{aligned} V(\varepsilon_i | x_{i1}, \ldots, x_{ik}) &= E(\varepsilon_i^2 | x_{i1}, \ldots, x_{ik}) \\ &= \sigma^2 h(x_{i1}, \ldots, x_{ik}) \\ &\equiv \sigma^2 h_i. \end{aligned}$$

Therefore,

$$V\left(\frac{\varepsilon_i}{\sqrt{h_i}} \mid x_{i1}, \ldots, x_{ik} \right) = \sigma^2,$$

If $\varepsilon_i^* = \varepsilon_i / \sqrt{h_i}$, then

$$\varepsilon_1^*, \ldots, \varepsilon_n^* \ \text{ iid } \ (0, \sigma^2)$$

It is easy to see that

$$\frac{y_i}{\sqrt{h_i}} = \beta_0 \frac{1}{\sqrt{h_i}} + \beta_1 \frac{x_{i1}}{\sqrt{h_i}} + \cdots + \beta_k \frac{x_{ik}}{\sqrt{h_i}} + \frac{\varepsilon_i}{\sqrt{h_i}},$$

is an homoskedastic regression and OLS can be used to compute $\hat{\beta}_0, \ldots, \hat{\beta}_k$ and respective standard errors.

Alternatively,

$$y_i^* = \beta_0 x_{i0} + \beta_1 x_{i1}^* + \cdots + \beta_k x_{ik}^* + \varepsilon_i^*$$

with $x_{i0} = 1/\sqrt{h_i}$ and $V(\varepsilon_i^*) = \sigma^2$

OLS vs GLS

The ordinary least square (OLS)estimation of

$$y_i^* = \beta_0 + \beta_1 x_{i1}^* + \cdots + \beta_k x_{ik}^* + \varepsilon_i^* \qquad \varepsilon_i^* \sim (0, \sigma^2),$$

yields $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, the generalized least square (GLS) estimates of of $\beta_0, \beta_1, \ldots, \beta_k$.

The GLS estimators are used to account for heteroskedasticity in the errors.

The GLS estimators for correcting heteroskedasticity are called weighted least squares (WLS) estimators. This name comes from the fact that the $\hat{\beta}_j$ minimize the weighted sum of squared residuals, where each squared residual is weighted by $1/h_i$.

Unknown heteroskedasticity

There are many ways to model heteroskedasticity, but we will study one particular, fairly flexible approach. Assume that

$$V(\varepsilon|x_1, \ldots, x_k) = \sigma^2 \exp\{\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k\}$$

where $x_1, x_2, \ldots, x_k$ are the independent variables appearing in the regression model, and the δ_j are unknown parameters.

In the notation of the previous slides

$$h(x_1, \ldots, x_k) = \exp\{\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k\}.$$

Algorithm

1. Run the regression of y on $x_1, x_2, \ldots, x_k$ and obtain the residuals, $\hat{\varepsilon}$.
2. Create $\log(\hat{\varepsilon}^2)$ by first squaring the OLS residuals and then taking the natural log.
3. Run the regression of $\log(\hat{\varepsilon}^2)$ on $x_1, x_2, \ldots, x_k$ and obtain the fitted values, $\hat{g}$.
4. Exponentiate the fitted values $\hat{h} = \exp(\hat{g})$.
5. Estimate the equation

$$y = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k + \varepsilon$$

by WLS, using weights $1/\hat{h}$.

smoke.txt

```
Obs:   807

educ: years of schooling
cigpric: the per-pack price of cigarettes (in cents)
white: =1 if white
age: measured in years
income: annual income
cigs: number of cigarettes smoked per day
restaurn: =1 if state with restaurant smoking restrictions
lincome: log(income)
agesq: age^2
lcigpric: log(cigprice)
```

R CODE

```
data = read.table("smoke.txt",header=TRUE)
attach(data)
n = nrow(data)

reg = lm(cigs~lincome+lcigpric+educ+age+agesq+restaurn)
summary(reg)

esq   = reg$res^2
R2.e  = summary(lm(esq~lincome+lcigpric+educ+age+agesq+restaurn))$r.sq
Ftest = R2.e/(1-R2.e)*((n-7)/6)
pval  = 1-pf(Ftest,6,n-7)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.639868  24.078661  -0.151  0.87988
lincome      0.880269   0.727784   1.210  0.22682
lcigpric    -0.750854   5.773343  -0.130  0.89656
educ        -0.501498   0.167077  -3.002  0.00277 **
age          0.770694   0.160122   4.813 1.78e-06 ***
agesq       -0.009023   0.001743  -5.176 2.86e-07 ***
restaurn    -2.825085   1.111794  -2.541  0.01124 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 13.4 on 800 degrees of freedom
Multiple R-squared:  0.05274,Adjusted R-squared:  0.04563
F-statistic: 7.423 on 6 and 800 DF,  p-value: 9.499e-08
```

```
> R2.e
[1] 0.03997326

> Ftest
[1] 5.551687

> pval
[1] 1.18881e-05
```

which is very strong evidence of heteroskedasticity.

GLS

```
lesq = log(esq)
g    = lm(lesq~lincome+lcigpric+educ+age+agesq+restaurn)$fit
hhat = exp(g)

cigs1 = cigs/sqrt(hhat)
ones1 = rep(1,n)/sqrt(hhat)
lincome1 = lincome/sqrt(hhat)
lcigpric1 = lcigpric/sqrt(hhat)
educ1=educ/sqrt(hhat)
age1 = age/sqrt(hhat)
agesq1 = agesq/sqrt(hhat)
restaurn1 = restaurn/sqrt(hhat)

reg.gls = lm(cigs1~ones1+lincome1+lcigpric1+educ1+age1+agesq1+restaurn1-1)
```

The weighted least squares estimates are

```
            Estimate Std. Error t value Pr(>|t|)
ones1      5.6353434 17.8031310   0.317 0.751678
lincome1   1.2952413  0.4370119   2.964 0.003128 **
lcigpric1 -2.9402848  4.4601431  -0.659 0.509934
educ1     -0.4634462  0.1201586  -3.857 0.000124 ***
age1       0.4819474  0.0968082   4.978 7.86e-07 ***
agesq1    -0.0056272  0.0009395  -5.990 3.17e-09 ***
restaurn1 -3.4610662  0.7955046  -4.351 1.53e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```