Het

Friday, March 21, 2025 10:30 AM

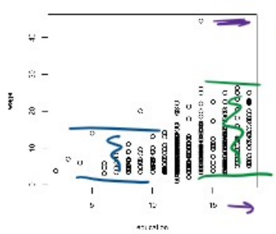
ch9

ECON 3040 - Heteroskedasticity

Ryan T. Godwin

University of Manitoba

Figure: Possible heteroskedasticity in the CPS data. The variance in wage may be increasing as education increases. The reasoning is that individuals who have not completed high school (or university) are precluded from many high paying jobs (doctors, lawyers, etc.). However, having many years of education does not preclude individuals from low-paying jobs. The spread in wages is higher for highly educated individuals.



var(wage | educ) is not constant

1. Define heterosked./homosked.
   ↳ in a sentence/formula
2. What are the implications?
   ↳ s.e.'s are wrong (inconsistent) ⇒ hyp. testing is wrong
   ↳ LS estimator is inefficient
3. Testing for het.
4. How to fix?

We have been assuming homoskedasticity

The estimators that we have used so far have good statistical properties provided that the following assumptions hold:

1. The population model is linear in the βs.
2. There is no perfect multicollinearity between the X variables.
3. The random error term, ϵ, has mean zero.
4. ϵ is identically and independently distributed. → homoskedasticity
5. ϵ and X are independent.
6. ϵ is Normally distributed.

These ensure LS is unbiased, efficient, and consistent, and that hypothesis testing is valid. A violation of one or more of these assumptions might lead us to estimators beyond LS.

We will consider that assumption 4 is violated in a particular way. Specifically, we consider what happens where the error term, ϵ, is *not* identically distributed.

Homoskedasticity

If assumption A4 is satisfied, then ϵ is identically distributed. This means that all of the ϵ_i have the same variance. That is, all of the random effects that determine Y, outside of X, have the same dispersion. The term *homoskedasticity* (same dispersion) refers to this situation of identically distributed error terms.
Stated mathematically, homoskedasticity means:

$$\text{Var}(\epsilon_i|X) = \sigma^2, \forall i$$

The variance of ϵ is constant, even conditional on knowing the value of X.
Homoskedasticity means that the squared vertical distance of each data point from the (population or estimated) line is, on average, the same. The values of the X variables do not influence this distance (the variance of the random unobservable effects are not determined by any of the values of X). See figure 2.

$wage = \beta_0 + \beta_1 educ + \dots + \epsilon$

$Var(wage) \propto var(\epsilon)$

$\cancel{E(\epsilon)=0}$

$\boxed{var(\epsilon) = \sigma^2}$

$var(\epsilon) \neq \sigma^2$

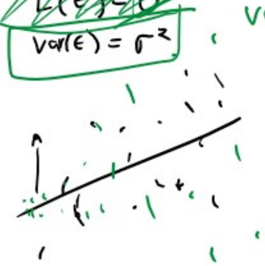
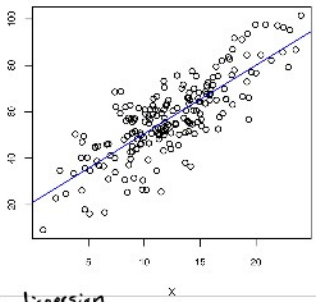


Figure: Homoskedasticity. The average squared vertical distance from the data points to the OLS estimated line is the same, regardless of the value of X.



differing dispersion

Heteroskedasticity

c

Heteroskedasticity refers to the situation where the variance of the error term ϵ is not equal for all observations. The term heteroskedasticity means *differing dispersion*. Mathematically:

$$\text{Var}[\epsilon_i|X_i] \neq \sigma^2, \ \forall i$$

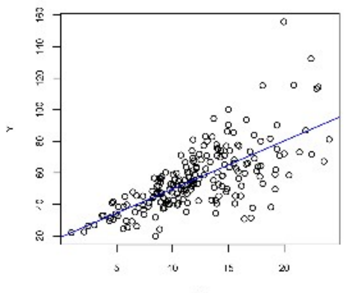
or

$$\text{Var}[\epsilon_i|X_i] = \sigma_i^2$$

Each observation can have its own variance, and the value of X may influence this variance.

Heteroskedasticity means that the squared vertical distance of each data point from the estimated regression line is not the same on average, and may be influenced by one or more of the X variables.

Figure: Heteroskedasticity. The squared vertical distance of a data point from the OLS estimated line is influenced by X.



The implications of heteroskedasticity

Heteroskedasticity is a violation of A.4, since each ϵ_i is not identically distributed. Heteroskedasticity has two main implications for the estimation procedures we have been using in this book:

1. The OLS estimator is no longer efficient. not so bad | fix → GLS weighted LS ← not in this class
2. The estimated standard errors are inconsistent.

The inefficiency of OLS is arguably a smaller problem than the inconsistency of the variance estimator. The second issue means that the estimated standard errors in our regression output are wrong, leading to the incorrect t-statistics and confidence intervals. Hypothesis testing, in general, is invalid. The problem arises because the formula that is the basis for estimating the standard errors in OLS:

$$\text{Var}[b_1] = \frac{\sigma_\epsilon^2}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

$\sigma_\epsilon^2 \rightarrow s_e^2$

assuming homoskedasticity

$$t = \frac{b_k - \beta_{k,0}}{\sqrt{\widehat{\text{Var}}[b_k]}}$$

is only correct under homoskedasticity.

Fixing heteroskedasticity - robust standard errors

To fix the more important problem of the inconsistency of the standard errors, the formula for $\text{Var}[b_1]$ must be updated to take into account the possibility of heteroskedasticity.

homosked. $\frac{\sum e_i}{n-k}$

Updating the formula to allow for heteroskedasticity in the estimation of the standard errors gives what is typically referred to as *robust standard errors*. In R, we will use the code:

```
install.packages('lmtest')
library(lmtest)
install.packages('sandwich')
library(sandwich)
```

→ formula looks like sandwich $X e_i X$

to install and load a package that can estimate the robust standard errors, and then use

```
coeftest(my.lm.model, vcov = vcovHC(my.lm.model, "HC1"))
```

to estimate the correct standard errors and updated t-statistics and p-values, where `my.lm.model` is the least-squares regression that we

Assume: X and ϵ are independent

↳ can't test e and X are always correlated

Assume: $\text{var}(\epsilon_i)$ = constant

↳ can use e_i's

Assume: $\epsilon_i \sim N$

↳ can test e_i in

to install and load a package that can estimate the robust standard errors, and then use

```
coeftest(my.lm.model, vcov = vcovHC(my.lm.model, "HC1"))
```

to estimate the correct standard errors and updated t-statistics and p-values, where `my.lm.model` is the least-squares regression that we have estimated using the `lm()` command.

↳ can use e_i

Assume: $\epsilon_i \sim N$

↳ can test e_i in Jarque-Bera

Testing for heteroskedasticity

There are several (approximately) equivalent tests for heteroskedasticity, but we'll focus on the most famous: White's[1] test. In White's test, the null hypothesis is that there is homoskedasticity, and the alternative is heteroskedasticity. That is:

$$H_0 : \text{var}[\varepsilon_i] = \sigma^2 \quad \text{homosked.}$$
$$H_A : \text{var}[\varepsilon_i] \neq \sigma_i^2 \quad \text{heterosked.}$$

[1] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: Journal of the Econometric Society, 817-838.

Take a simple population model with two regressors. Remember that the population model and the estimated model are (respectively):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$
$$y = b_0 + b_1 x_1 + b_2 x_2 + e$$

The residual e is the counterpart to the unobservable error term ε! Sometimes, we can use the residuals to test assumptions or properties of the error term. For example, we can look to see if the *residuals* are homoskedastic or heteroskedastic, in order to infer those propertied about the error term. That is, if e looks homoskedastic, we will conclude that so is ε.

White's test tries to explain differences in the size of the squared residuals from a least-squares model by regressing them on the original x variables, and the squares and cross products of the xs. If the R^2 from this regression is high, then we conclude that there is some pattern to the size of the residuals, and reject the null hypothesis of homoskedasticity.

H_0: homosked → $\text{var}(\epsilon_i) = \sigma^2$

H_A: heterosked → $\text{var}(\epsilon_i) \neq \sigma^2$

1st: estimate pop. model (get e_i)

2nd: (auxilliary regression): $e_i^2 = \alpha_0 + \alpha_1 X_1 + \alpha X_2 + \alpha_3 X_1^2 + \alpha_4 X_2^2 + \alpha_5 X_1 X_2 + \varepsilon_i$

↳ if R^2 is large ⇒ reject homosked.

To test for heteroskedasticity in the population model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

we would estimate it by LS, for example by using `lm(y ~ x1 + x2)`. We then get the squared residuals from this regression, and estimate the following equation by LS:

$$\hat{e}^2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \times x_2) + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon \qquad (1)$$

Equation 1 is looking for any approximate way to explain variation in the size of the squared residuals. If the estimated model from equation 1 fits well (in terms of the R-squared), then there is some explanation for the variance in the error term, and the error term is heteroskedastic. White's test statistic is the nR^2 from this auxiliary regression, and the p-value for the test comes from the Chi-square distribution. As usual, if the p-value is small, we reject the null of homoskedasticity, in favour of heteroskedasticity.

explanation for the variance in the error term, and the error term is heteroskedastic. White's test statistic is the nR^2 from this auxiliary regression, and the p value for the test comes from the Chi square distribution. As usual, if the p value is small, we reject the null of homoskedasticity, in favour of heteroskedasticity.

To test for heteroskedasticity in R, we need to install and load a package:

```
install.packages('skedastic')
library(skedastic)
```

and then use:

```
white(my.lm.model, interactions = TRUE)
```

where `my.lm.model` is the model we have estimated by LS. If we find heteroskedasticity, then we need to use heteroskedastic robust standard errors (such as White's standard errors).

Heteroskedasticity in food expenditure data

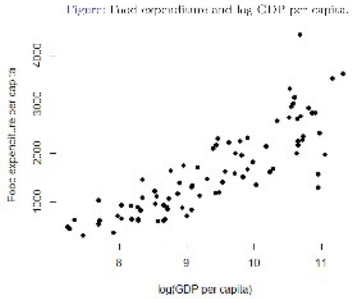
Download a data set on food expenditure by country, in 2016:

```
food <- read.csv('https://rtgodwin.com/data/foodexp.csv')
```

The variables are `foodexp` - food expenditure per capita (in US dollars), and `GDPpercap` - GDP per capita. There are 84 countries in the sample. Plot the data, taking the log of GDP per capita (see Figure 4):

```
plot(log(food$GDPpercap), food$foodexp, pch=16, xlab="log(
    GDP per capita)",
     ylab="Food expenditure per capita")
```

Figure: Food expenditure and log GDP per capita.



Estimate the population model

The following model for food expenditure:

$$foodexp = \beta_0 + \beta_1 \log(GDPpercap) + \epsilon$$

can be estimated in R using:

```
food.mod <- lm(foodexp ~ log(GDPpercap), data=food)
summary(food.mod)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4737.68     451.35  -10.50   <2e-16 ***
log(GDPpercap)    677.40      47.81   14.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
    1

Residual standard error: 479 on 82 degrees of freedom
Multiple R-squared:   0.71, Adjusted R-squared:  0.7065
F-statistic: 200.8 on 1 and 82 DF,  p-value: < 2.2e-16
```

Test for heteroskedasticity

If heteroskedasticity is present in this data, then the standard errors, t-statistics, and p-values, are all wrong! Hypothesis testing, and any conclusions we draw, may be incorrect due to the heteroskedasticity. To test for heteroskedasticity, we can use White's test:

$e^2 = \beta_1 x + \beta_2 x^2 + \varepsilon$

```
install.packages('skedastic')
library(skedastic)
white(food.mod)
```

options: interactions / squares

```
  statistic p.value parameter method       alternative
      <dbl>   <dbl>     <dbl> <chr>        <chr>
1      11.6 0.00304         2 White's Test greater
```

The test statistic from the White test is 11.6, with an associated p-value of 0.00304. We reject the null hypothesis of homoskedasticity.

To see what the function white() is doing, we'll calculate the White test statistic and p-value "by hand":

3. Testing

↳ 2nd stage regression of e^2 on stuff

```
food.resid.sq <- food.mod$residuals ^ 2
summary(lm(food.resid.sq ~ log(GDPpercap) + I(log(GDPpercap)
      ^ 2), data=food))
```

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           4703163    3680579   1.278    0.205
log(GDPpercap)       -1121179     795218  -1.410    0.162
I(log(GDPpercap)^2)     67733      42508   1.593    0.115

Residual standard error: 444800 on 81 degrees of freedom
Multiple R-squared:  0.138, Adjusted R-squared:  0.1167
F-statistic: 6.486 on 2 and 81 DF,  p-value: 0.002442
```

The test statistic is $nR^2 = 84 \times 0.138 = 11.6$ (same as from the white() command). The p-value can be found from:

```
1 - pchisq(84 * 0.138, 2)
```

```
0.003039889
```

which is the same from the white() command.

White's heteroskedastic consistent standard errors

To recalculate the standard errors, t-statistics, and p-values, we can use the coeftest() function:

```
install.packages('lmtest')
library(lmtest)
install.packages('sandwich')
library(sandwich)
coeftest(food.mod, vcov = vcovHC(food.mod, "HC1"))
```

"heteroskedastic consistent"

```
t test of coefficients:

               Estimate Std. Error t value  Pr(>|t|)
(Intercept)   -4737.082    476.616 -9.9423 9.705e-16 ***
log(GDPpercap)  677.398     54.089 12.5384 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
    1
```

this changes!

Notice that the estimated βs have not changed, but that the standard errors have changed, t-statistics, and p-values have changed.

estimated βs are unchanged

Heteroskedastic errors have a pretty severe consequence; hypothesis testing may be invalid. The prevalence of heteroskedasticity in many economics data has led to the common practice of erring on the side of caution. Heteroskedastic robust standard errors are often used, if heteroskedasticity is suspected. Note that homoskedasticity is a special case of heteroskedasticity, so the downside of using the robust estimator when it is not needed, is small.