9 – Beyond OLS: Heteroskedasticity

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

A1 The population model is linear in the βs.

A2 There is no perfect multicollinearity between the X variables.

A3 The random error term, ϵ, has mean zero.

A4 ϵ is identically and independently distributed.

A5 ϵ and X are independent.

A6 ϵ is Normally distributed.

These assumptions assure that OLS 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 OLS. OLS is simple, and easy to use, but is often thought of a starting point in econometric modelling since the above assumptions are often unreasonable.

In this section, we will consider that assumption A4 is violated in a particular way. Specifically, we consider what happens where the error term, ϵ, is *not* identically distributed.

9.1 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.

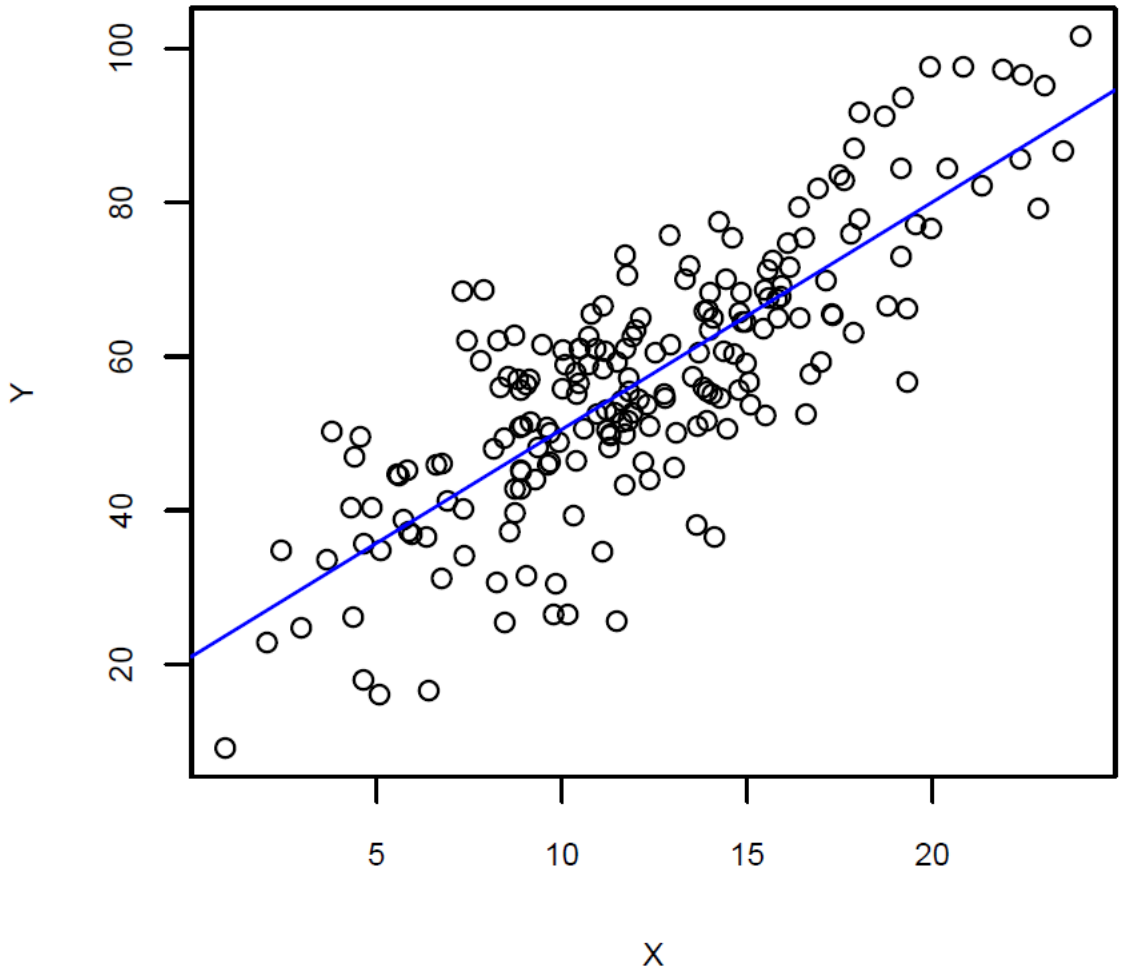
Stated mathematically, homoskedasticity means:

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

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 9.1.

Figure 9.1: Homoskedasticity. The average squared vertical distance from the data points to the OLS estimated line is the same, regardless of the value of X.



9.2 Heteroskedasticity

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:

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

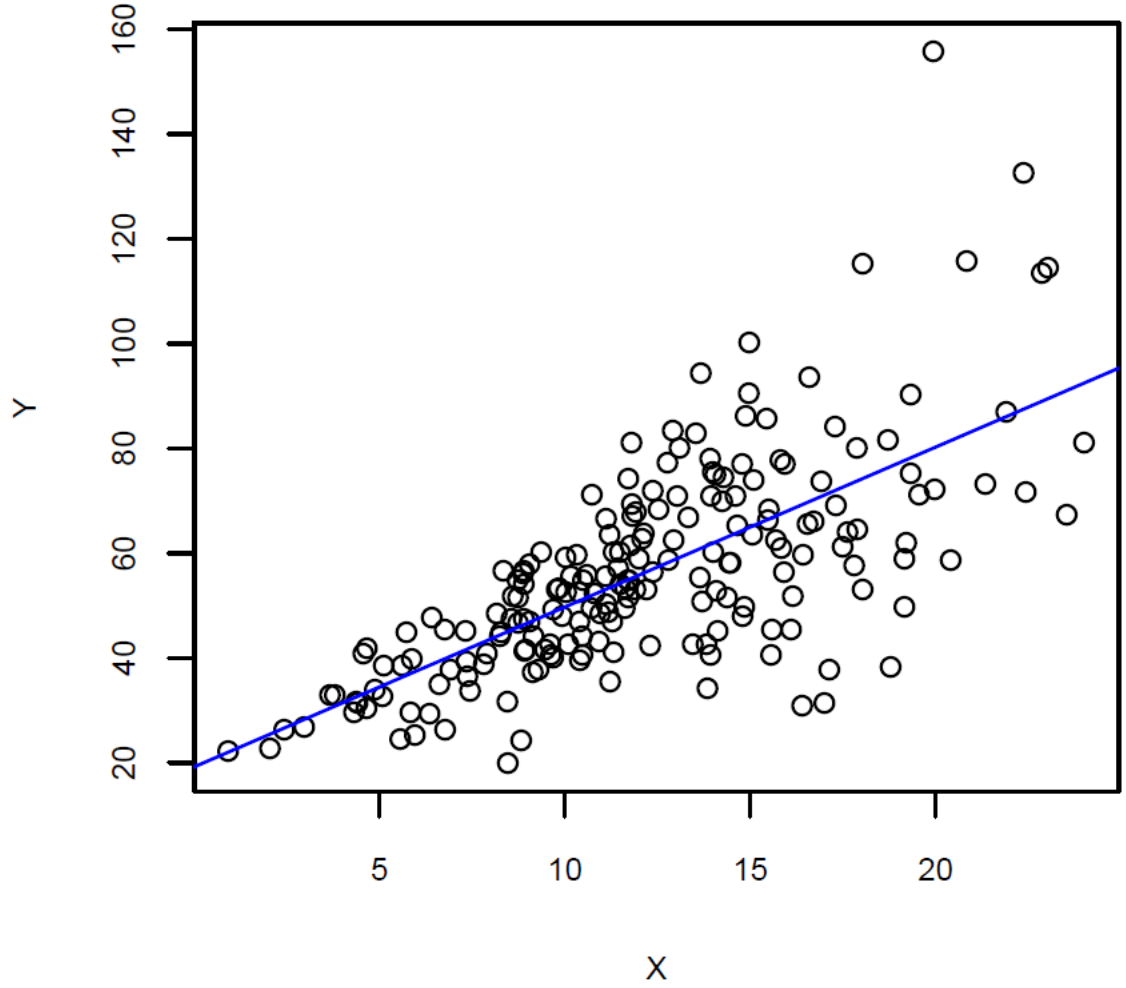
or

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

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. See figure 9.2, where the larger the value of X is, the larger the variance of ϵ.

Figure 9.2: Heteroskedasticity. The squared vertical distance of a data point from the OLS estimated line is influenced by X.



9.2.1 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:

(i) The OLS estimator is no longer efficient.

(ii) The estimator for the variance of the OLS estimator is inconsistent.

The inefficency of OLS is arguably a smaller problem than the inconsistency of the variance estimator. (ii) 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 (equation 5.7):

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

is only correct under homoskedasticity.

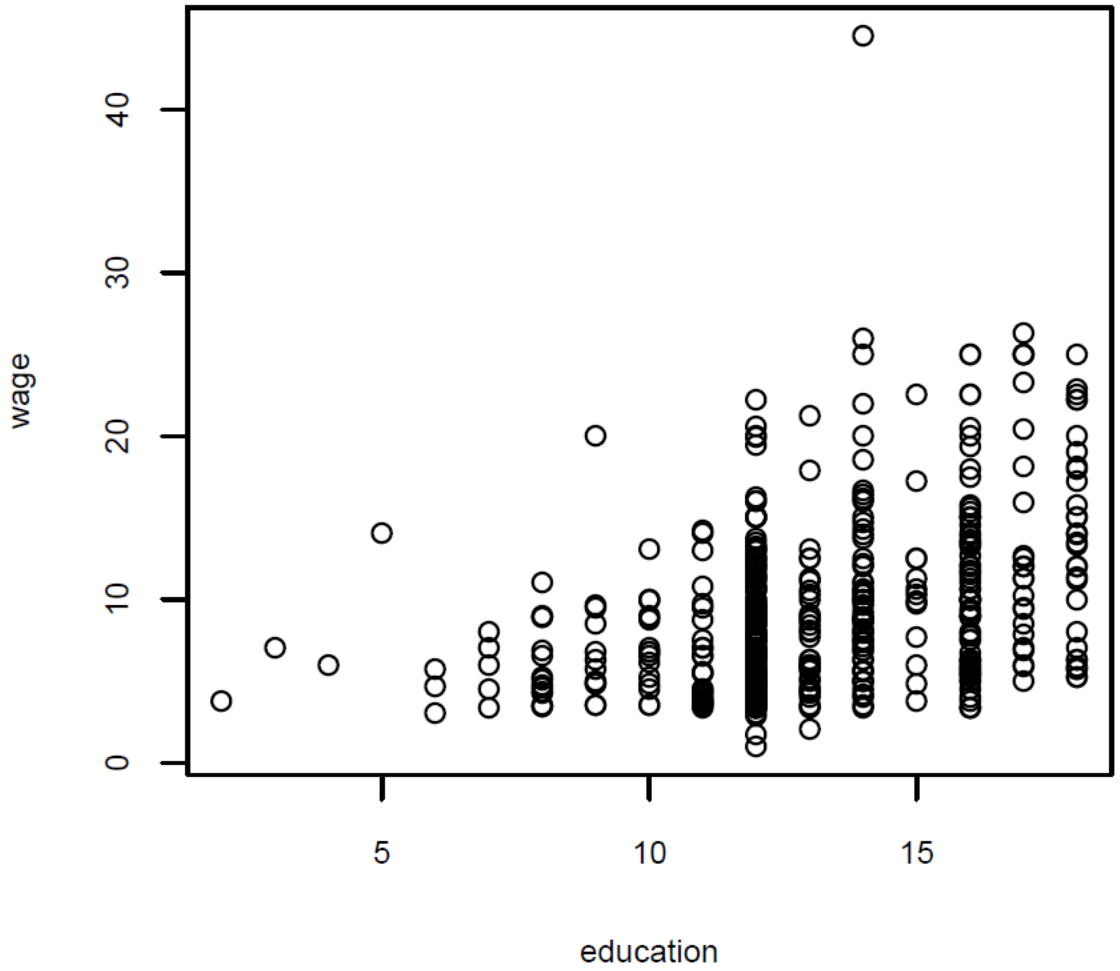
To fix problem (i), the inefficiency of OLS, we must use a different estimator, such as Generalized Least Squares (GLS). GLS is not discussed here. To fix (ii), the more important problem of the inconsistency of the standard errors, the formula for $\text{Var}[b_1]$ must be updated to take into account the possibility of heteroskedasticity.

Updating the formula to allow for heteroskedasticity in the estimation of the standard errors gives what is typically referred to as *robust standard errors*.

9.2.2 Heteroskedasticity in the CPS data

It may be the case that the variance in wages depends on education. The reasoning is that individuals who have not completed highschool (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. Figure 9.3 illustrates this point.

Figure 9.3: Heteroskedasticity in the CPS data. The variance in `wage` may be increasing as `education` increases.



```
summary(lm(wage ~ education + gender + age + experience))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.9574     6.8350  -0.286    0.775
education      1.3073     1.1201   1.167    0.244
genderfemale  -2.3442     0.3889  -6.028 3.12e-09 ***
age           -0.3675     1.1195  -0.328    0.743
experience     0.4811     1.1205   0.429    0.668
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.458 on 529 degrees of freedom
Multiple R-squared:  0.2533,    Adjusted R-squared:  0.2477
F-statistic: 44.86 on 4 and 529 DF,  p-value: < 2.2e-16
```

the standard errors, t-statistics, and associated p-values are all wrong under heteroskedasticity. To estimate the *robust* standard errors (which will update the t-statistics and p-values as well), we can use the following commands in R:

```
results <- lm(wage ~ age + education + gender + experience)
coeftest(results, vcov = vcovHC(results, "HC1"))
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  -1.95744    1.53006 -1.2793  0.201345
age          -0.36749    0.12384 -2.9675  0.003138 **
education     1.30727    0.12452 10.4983 < 2.2e-16 ***
genderfemale -2.34416    0.39543 -5.9282  5.53e-09 ***
experience    0.48107    0.13502  3.5629  0.000400 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that the estimated βs have not changed, but that the standard errors have changed quite dramatically, leading to very different conclusions about the statistical significance of the X variables.