

1 The problem.

Recall

$$y_i = b_0 + b_1 x_i + e_i$$

We typically assume the e_i 's are drawn from the same distribution, so that

$$E(e_i) = 0 \text{ for all } i$$

and the variance is homogeneous as well:

$$\text{Var}(e_i) = E[(e_i - E(e_i))^2] = \sigma_i^2 = \sigma^2$$

In other words, all error terms have the same variance. They are drawn from the same distribution.

Almost always, we cling with vigor to the first assumption, but there is a pretty literature on the impact of violations of the second one. The problem of heteroskedasticity (or heteroscedasticity) arises when the assumption of homogeneous variance is violated.

If this is violated

1. Estimates of b_0 and b_1 are still unbiased and consistent.

Proof: For simplicity, consider the OLS estimate of the slope from data in deviations form:

$$\hat{b}_1 = \frac{\sum x_i \cdot y_i}{\sum x_i^2} = \frac{\sum x_i (b_0 + b_1 x_i + e_i)}{\sum x_i^2} = \frac{b_1 \sum x_i^2 + \sum x_i \cdot e_i}{\sum x_i^2} = b_1 + \frac{\sum x_i \cdot e_i}{\sum x_i^2}$$

Usually the textbook will then use the following argument to show that \hat{b}_1 is unbiased by taking either of two routes. No matter which route you plan to take, start by applying the Expected value operator to both sides:

$$\begin{aligned} E(\hat{b}_1) &= E\left(b_1 + \frac{\sum x_i \cdot e_i}{\sum x_i^2}\right) \\ &= E(b_1) + E\left(\frac{\sum x_i \cdot e_i}{\sum x_i^2}\right) \\ &= b_1 + E\left(\frac{\sum x_i \cdot e_i}{\sum x_i^2}\right) \end{aligned}$$

Route 1 claims “ x_i is not a random variable.” Rather, it is a fixed constant value representing an individual attribute. Since x_i is a constant, it means that $E(x_i e_i) = x_i E(e_i)$. Furthermore, recall the assumption that $E(e_i) = 0$, so it is clear that

$$E(\hat{b}_1) = b_1 + 0 = b_1$$

Route 2 claims that even though x_i may be thought of as a random variable, we can assume that it is uncorrelated with e_i . If two variables are uncorrelated, it means they have no covariance, so $E(x_i e_i) = 0$.

Either route leads to the same answer.

$$E(\hat{b}_1) = b_1$$

Meaning that variance of e_i plays no role in the question of whether or not the OLS estimate \hat{b}_1 is unbiased.

2. Variance (and hence standard error of b_1) is estimated incorrectly (underestimated, in fact) by the OLS formulas.

This means the t-tests with the computer printout from any standard program are **WRONG**.

If there is no covariance between errors, it can be shown (for the “nonstochastic x_i case”, as discussed in route 1 above):

$$\begin{aligned} Var(\hat{b}_1) &= Var \left[\frac{\sum x_i \cdot e_i}{\sum x_i^2} \right] = \frac{Var[\sum x_i e_i]}{(\sum x_i^2)^2} = \frac{\sum Var(x_i e_i)}{(\sum x_i^2)^2} = \frac{\sum x_i^2 \cdot Var(e_i)}{(\sum x_i^2)^2} \\ &= \frac{\sum x_i^2 \cdot \sigma_i^2}{(\sum x_i^2)^2} \end{aligned}$$

And, note that the variance of each individual error term, σ_i^2 , can be written as the sum of a mean variance s^2 and an individualized variance s_i^2 , so $\sigma_i^2 = s^2 + s_i^2$. Plug this into the expression above:

$$\frac{\sum x_i^2 \cdot \sigma_i^2}{(\sum x_i^2)^2} = \frac{\sum x_i^2 (s^2 + s_i^2)}{(\sum x_i^2)^2} = \frac{s^2}{\sum x_i^2} + \frac{\sum x_i \cdot s_i^2}{(\sum x_i^2)^2}$$

The first term is "roughly" what OLS would calculate for the variance of \hat{b}_1 . The second term is the additional "true variance" in the OLS estimator. That variance is “really in there” but the OLS formula for the variance does not include it.

3. \hat{b}_1^{OLS} is *inefficient*, meaning we can find another linear estimator with lower variance. That alternative estimator is known as the WLS, or Weighted Least Squares estimator, \hat{b}_1^{WLS} .

2 Information Sandwiches and the White HCE approach

If you are willing to ignore the problem of inefficiency in \hat{b}^{OLS} , there is a widely used short-cut that can be used to deal with the problem of heteroskedasticity. This is known as a robust estimate of the variance of \hat{b} because it is not built on the assumption that all observations are drawn from a homogeneous distribution.

This robust approach allows us to get consistent standard errors, and hence the t-test is meaningful again. The most famous approach is White’s Heteroskedasticity Consistent Estimator (HCE) (also known as the Heteroskedasticity Consistent Covariance Matrix, or HCCM) approach. There have been several variants of the HCE, here’s one:

$$Var(\hat{b}_1) = \frac{\sum x_i^2 \cdot \hat{e}_i^2}{(\sum x_i^2)^2} \tag{1}$$

Basically, this substitutes the observed “residual” \hat{e}_i^2 for the unknown error variance.

Matrix digression on White's formula

This formula works for a model in which there is one independent variable. If there are several input variables, then the matrix form is called for. Recall the OLS estimator

$$\hat{b} = (X'X)^{-1}X'Y \quad (2)$$

Recall, if the assumption of homogeneous variance is met, then the “true variance” of the estimates of the b 's is

$$\text{true Var}(\hat{b}) = \sigma^2 \cdot (X'X)^{-1} \quad (3)$$

and we estimate those values by replacing the “true variance of the error term”, σ^2 , with the Mean Squared Error (MSE). In the OLS model, then, we estimate the variance of \hat{b} by

$$\text{estimated Var}(\hat{b}) = \text{MSE} \cdot (X'X)^{-1} \quad (4)$$

How do we arrive at this formula? Well, there's a straightforward application of the variance operator to \hat{b} and at the second-to-last step, we arrive at this step:

$$\text{true Var}(\hat{b}) = (X'X)^{-1}(X'\text{Var}(e)X)(X'X)^{-1} \quad (5)$$

With OLS assuming homoskedasticity, all of the error terms have the same variance. In matrix form, e is a column of N error terms. And $\text{Var}(e)$ is a matrix showing the Variance/Covariance of the individual observations. Consider:

$$\text{Var}(e) = E(e \cdot e'|X) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ \dots & & \dots & \dots & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \dots & & \dots & \dots & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \sigma^2 \cdot I \quad (6)$$

On the diagonal, all values are σ^2 . Off the diagonal, we assume there is no covariance (no autocorrelation). Putting that knowledge to use, then the expression in 5 is radically simplified.

$$\text{true Var}(\hat{b}) = (X'X)^{-1}(X' \cdot \sigma^2 \cdot IX)(X'X)^{-1} = \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1} \quad (7)$$

And since $(X'X)^{-1}(X'X) = I$, this reduces to the result given in 3.

The simple formula for $\text{Var}(\hat{b})$ in equation 4 is valid only if the matrix of error term variance is the sort given by 6. If, instead of homoskedastic errors, we have heteroskedasticity, then there's trouble. We are using the wrong formula to estimate the variance of \hat{b} .

White's idea

In the heteroskedasticity corrected covariance approach that White proposed, the assumptions that led to this simple result are undone. One instead begins with the idea the variances of the error terms are not all the same. If we look at a particular observation,

$$\text{Var}(e_i) = E[(e_i - E(e_i))^2] = E[e_i^2] = \sigma_i^2.$$

If we still insist there is no autocorrelation, but we allow variances to differ, we get this:

$$Var(e) = E[e \cdot e' | X] = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & \dots & 0 \\ 0 & 0 & 0 & \sigma_{N-1}^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_N^2 \end{bmatrix}$$

which looks slightly nicer if we factor out a “common variance” σ^2 and then assume all of the observations have error variances that are proportional to σ^2 .

$$Var(e) = \sigma^2 \begin{bmatrix} w_1 & 0 & 0 & 0 & 0 \\ 0 & w_2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & \dots & 0 \\ 0 & 0 & 0 & w_{N-1} & 0 \\ 0 & 0 & 0 & 0 & w_N \end{bmatrix} = \sigma^2 \Omega \quad (8)$$

In William Greene’s *Econometric Analysis, 5ed* (p. 218), he shows that the OLS \hat{b} has “true variance” given by

$$\begin{aligned} true\ Var(\hat{b}) &= (X'X)^{-1}(X'Var(e)X)(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1} \end{aligned} \quad (9)$$

The expression given in 9 does not reduce to something that is estimated by the OLS formula for the variance in 6.

Recall that in OLS, we replace the true variance σ^2 with an estimate, MSE, a value we calculate from the fitted regression. We just need something to “plug in” for $Var(e)$ or $\sigma^2\Omega$ and then we proceed as usual.

White’s idea, which was a major breakthrough (Econometrica, 1980), was to estimate the variances of the individual observations. The variance of e_1 , for example, is never observed, but the best estimate we have for it is the mean square for that one case:

$$\widehat{e}_1^2 = (y_1 - X_1\hat{b})(y_1 - X_1\hat{b})$$

Hence, the “middle part” of the expression 9 can be estimated. Instead of $Var(e)$, we use a matrix of estimates like this:

$$Var(\widehat{e}) = \begin{bmatrix} \widehat{e}_1^2 & & & & \\ & \widehat{e}_2^2 & & & \\ & & & & \\ & & & \widehat{e}_{N-1}^2 & \\ & & & & \widehat{e}_N^2 \end{bmatrix}$$

The “heteroskedasticity consistent covariance matrix of \hat{b} ” is going to use this matrix in place of $Var(e)$ in the formula to calculate estimated variance.

$$hccm\ Var(\hat{b}) = (X'X)^{-1}(X'Var(\widehat{e})X)(X'X)^{-1}$$

White proved that the estimator is consistent, i.e., for large samples, the value converges to the true $Var(\hat{b})$.

This is sometimes called an “information sandwich” estimator. The matrix $(X'X)^{-1}$ is the “information matrix”, a term drawn from Maximum Likelihood estimation. If you want to know more details on that, I’ve written handouts on Maximum Likelihood estimation. Note this equation gives us a “sandwich” of $X'Var(e)X$ between two pieces of information matrix.

3 Weighted Least Squares

If you are concerned about inefficiency of the OLS estimator, \hat{b}^{OLS} , there is an alternative estimator which is known to have lower variance. In fact, that’s how we know that OLS is inefficient, because we can demonstrate a lower variance alternative.

If we put less weight on the cases that have high variance, we might protect the estimation process from the uncertainty they impose. The WLS estimator assumes these weights, w_i , and then uses the estimating criterion

$$\text{minimize } SS(\hat{b}) = \sum_{i=1}^N w_i (y_i - \hat{y}_i)^2$$

The idea of WLS is to homogenize the variances. Recall that

$$Var\left(\frac{e_i}{\sigma_i}\right) = \frac{1}{\sigma_i^2} Var(e_i) = \frac{\sigma_i^2}{\sigma_i^2} = 1$$

Hence, if we transform the residuals in the regression model, we can homogenize the variances of the error terms. You can look at this from either of two perspectives. In both, you multiply by $w_i = \frac{1}{\sigma_i}$ to implement a “weighting” procedure.

WLS Approach 1: minimize a weighted sum of squares. In OLS, you would minimize

$$\sum (y_i - \hat{y}_i)^2$$

Instead, minimize this:

$$\begin{aligned} & \sum \left[\frac{(y_i - \hat{y}_i)}{\sigma_i} \right]^2 \\ &= \sum \left[\frac{y_i - \hat{b}_0^{WLS} - \hat{b}_1^{WLS} x_i}{\sigma_i} \right]^2 \\ &= \sum \left[\frac{y_i}{\sigma_i} - \frac{\hat{b}_0^{WLS}}{\sigma_i} - \frac{\hat{b}_1^{WLS} x_i}{\sigma_i} \right]^2 \end{aligned}$$

Note, this ASSUMES you have the “true” value of σ_i and can insert it into the calculations. If you use an estimate of σ_i from your sample, you probably don’t have an exactly correct value. Sometimes to differentiate a WLS based on the true values of σ_i from an analysis based on the estimates, they call the latter an FWLS, “Feasible Weighted Least Squares.”

WLS Approach 2: Divide each term in the original equation by the weighting factor σ_i .

$$\frac{y_i}{\sigma_i} = \frac{\hat{b}_0^{WLS}}{\sigma_i} + \frac{\hat{b}_1^{WLS}}{\sigma_i} x_i + \frac{e_i}{\sigma_i}$$

If you did an OLS estimation on this revised equation, it would be equivalent to doing the WLS approach 1.

It is easy to see with the formula for estimating b_1 that the WLS estimate obtained with this model has lower variance than the b_1 from the OLS model. (DO SO!)

4 Feasible Weights.

Where do you get the σ_i to plug into the WLS model? This depends on the theory you have, and what might be causing the heteroskedasticity. Often, this is treated as a matter of “special cases,” substantively justified specifications for the variance.

4.1 Look at the scatterplot, try to “eyeball” it. (I know, it sounds awful.)

Suppose it looks like variance is proportional to x_i , $\sigma_i^2 = k^2 \cdot x_i$. That means the right weight would be

$$\sigma_i = k \cdot \sqrt{x_i}$$

Variance is proportional to x_i^2 , $\sigma_i^2 = k^2 \cdot x_i^2$. That means the weight should be:

$$\sigma_i = k \cdot x_i$$

Note that the weight coefficient k does not matter. If it is unknown, just $\sqrt{x_i}$ or x_i , as the case may be. The coefficient k is for scaling, and no matter what value you put in for it, the parameter estimates are the same. In other words, k is unimportant. Try to prove this to yourself.

There has been some interesting discussion about whether it is better to use OLS, knowing that the assumption about the variance of e_i is wrong, or should one instead use WLS, knowing the estimate of σ_i is wrong. You see radically different opinions, with many “high brow” econometric theorists in favor of WLS, but some “applied researchers” are not. They would rather use OLS and then use a robust estimator of the standard error.

4.2 Random coefficient model

The kind of heteroskedasticity discussed in the previous section can arise if your theory was not correctly specified at the outset.

Suppose, for example, that there is a “random coefficient”

$$b_i = b_1 + u_i$$

We proceed as though this error inside the coefficient is very well behaved, with a mean of 0 and a variance σ_u^2 for all i .

Instead of the original theoretical model,

$$y_i = b_0 + b_1 x_i + e_i$$

the theory is now

$$y_i = b_0 + b_1 x_i$$

If you insert the equation for b_i , this reduces to

$$\begin{aligned} y_i &= b_0 + (b_1 + u_i)x_i \\ &= b_0 + b_1 x_i + u_i x_i \end{aligned}$$

The unmeasured term, the error term, is $u_i x_i$. Apply the variance operator to that, what do you get?

$$Var[u_i x_i] = \sigma_u^2 \cdot x_i^2$$

In this case, the parameter σ_u^2 plays the role of k in the previous subsection.

4.3 With grouped data, the variance is proportional to sample size.

Recall from the fundamentals of statistics that mean of y is

$$\bar{y} = \frac{\sum y_i}{N}$$

Recall also that the variance of the mean is the variance of y divided by N .

$$Var(\bar{y}) = \frac{Var(y_i)}{N} = \frac{\sigma_y^2}{N}$$

As you will recall, the so-called “standard error of the mean” is the square root of this quantity,

$$Std.Err.(\bar{y}) = \frac{\sigma_y}{\sqrt{N}}$$

Anyway, suppose your data represents groups, not individual people. You have the average on some variable for many different groups. If you observe groups of different sizes, say N_i , then the means observed will have different variances if the groups are different sizes. Rather than acting as though the variances of your observed means for the groups are homogeneous, you should instead find out how many cases were used in each unit to calculate the means, and then proceed as if the variance of the error term is inversely proportional to N_i .

4.4 Meta analysis

Suppose you have many different data sets and you fit a regression $y_i = b x_i + e_i$ in each one. You would observe different b 's across the analysis. Label the estimates b^1, b^2 , etc. If you then wanted to find out if there was a pattern in the b 's, say they are related to a variable Z , you might want to run a regression

$$b^j = c_0 + c_1 Z_j + u_j$$

We ran each regression model already, so we have estimated the standard error (or variance) of each b^j . So we know there is heteroskedasticity. And we can use the estimates from the individual regressions as weights in WLS.

5 Testing for heteroskedasticity

Many tests exist for specialized forms of heteroskedasticity. Here's a brief list of the ones with which I'm most familiar.

5.1 Categorical X's: Bartlett's test for grouped X's

Basically, this estimates the error variances for the subgroups and contrasts that against the variance for the combined dataset. It uses a χ^2 test to compare them.

5.2 Continuous X's: Goldfield Quandt test to determine if

$$\sigma_i^2 = \sigma^2 \cdot x_i^2$$

An F test results if you calculate the Error Sum of Squares for 2 pieces of data, usually we compare the "lower set" ESS_1 against the "upper set" ESS_2 after excluding some observations in the middle. Check your stats books, basically it is

$$F = \frac{ESS_2}{ESS_1}$$

and the degrees of freedom for both numerator and denominator are $(N - d - 4)/2$, where d is the number of excluded observations. The more observations you exclude, the smaller will be your degrees of freedom, meaning your F value must be larger.

5.3 Breusch-Pagan test/White test

Versions of this test were proposed in 1979 & 1980 by Breusch & Pagan and White. The idea is the same. If there is no heteroskedasticity, then the estimate of the coefficients from Ordinary Least Squares, b^{OLS} , should not be grossly different from a maximum likelihood estimator b^{MLE} . After a long series of gyrations, we arrive at the conclusion that the variance of the residuals should not be predictable with the use of input variables. The squared residuals can be used as estimates. The in the BP test with 2 input variables is:

$$\frac{\hat{e}_i^2}{\sigma^2} = \gamma_0 + \gamma_1 Z1_1 + \gamma_2 Z2_i$$

Here, $\sigma^2 = MSE$. If the error is Normal, the coefficients γ_0 , γ_1 , and γ_2 will all equal zero. The input variables Z can be the same as the original regression, but usually they are also going to include squared values of those variables.

BP contend that $\frac{1}{2}RSS$ (the regression sum of squares) should be distributed as χ^2 with degrees of freedom equal to the number of Z variables.

The original form of the BP test assumed Normality. White's version got rid of that assumption. In fact, he has a simpler result. Run the regression

$$\hat{e}_i^2 = \gamma_0 + \gamma_1 Z1_1 + \gamma_2 Z2_i$$

White claimed that, under the assumption of homoskedasticity,

$$N \cdot R^2 \sim \chi_p^2$$

where N is the sample size, R^2 is the coefficient of determination from the fitted model, and p is the number of variables used in the regression.

Many statistical programs will provide variants of these tests. It is vital to read the MANUAL. There are several different White's tests, and not all programs are completely clear which they use.

You can also calculate these things on your own. Just use your statistical program to output the residuals, and square those to estimate the error Variances. They become a "new column" of data you can analyze.