Advanced Topics in Applied Regression

Day 2: Heteroskedasticity and Its Solutions

Constantin Manuel Bosancianu

Doctoral School of Political Science
Central European University, Budapest
bosancianu@icloud.com

August 1, 2017

Detour into definitions

a, b_1, e = estimates (based on the sample).

$\alpha, \beta_1, \epsilon$ = parameters (in the population).

Naturally, we can rarely know the parameters in the population; all we can do is reduce our uncertainty in guessing them based on the sample.[1]

[1] We're basically talking about 2 models: a theoretical one, and an empirical one.

OLS estimator: unbiasedness

We're aiming for estimates to have a series of desirable properties. To begin with, in a finite sample:

- ✓ Unbiasedness: $E(b) = \beta$ (we're not *systematically* over- or under-estimating β);

It can be shown with a substitution and about 4 lines of math that:

$$E(b) = \beta + E\left(\frac{\sum_{i=1}^{n}(x_i - \bar{X})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}\right) \tag{1}$$

That fraction's top part is 0 if $E(\epsilon|X) = 0$.

OLS estimator: efficiency

Also in a finite sample:

- ✓ Efficiency: $Var(b)$ is smaller than that of any other linear unbiased estimator;

The proof here is a bit longer, but this variance of b depends on the variance of x and of the e_i.

When this condition is not met, we call the estimator *inefficient*.[2]

[2] The *Gauss-Markov* theorem guarantees that under 3 assumptions (homoskedasticity, linearity, and error independence), OLS is efficient.

OLS estimator: normality

As the sample grows to infinity (asymptotically), we can also expect to have:

- ✓ Normality: the sampling distribution of b will be centered on β, with known variance;

$$b \sim \mathcal{N}\left(\beta, \frac{\sigma_e^2}{(n-1)\sigma_x^2}\right) \tag{2}$$

This result[3] is important as it allows us to conduct significance tests, and then construct confidence intervals.

[3] Does not require that $e_i \sim \mathcal{N}$. Needs the same 3 assumptions as for efficiency.

OLS estimator: consistency

This one is not mentioned very often. As sample grows to infinity, we want to see b moving closer and closer to β.

$$\lim_{n \to \infty} P(|b - \beta| > \epsilon) = 0, \quad \forall\, \epsilon > 0 \tag{3}$$

`BLUE`: OLS is a "best linear unbiased estimator" if assumptions hold.

Homoskedasticity

Assumption of homoskedasticity

$$Y_i = a + b_1 X_{1i} + \cdots + b_k X_{ki} + e_i \quad (4)$$

This one targets the e_i, and specifically their variance—it must be constant, and not depend on any Xs.

$$Var(e_i | X_1, \ldots X_k) = \sigma_e^2 \quad (5)$$

Even when this assumption is violated, OLS estimates for b are still unbiased, consistent, and asymptotically normal.[4]

[4] The bias depends on whether $E(\epsilon | x) = 0$, not their variance.

Assumption of homoskedasticity

However, in the presence of violations of homoskedasticity, the estimator loses its efficiency: $Var(b)$ is not as small as it could be.

No amount of sample increase can solve this problem $\Rightarrow$ t-tests will be imprecise.

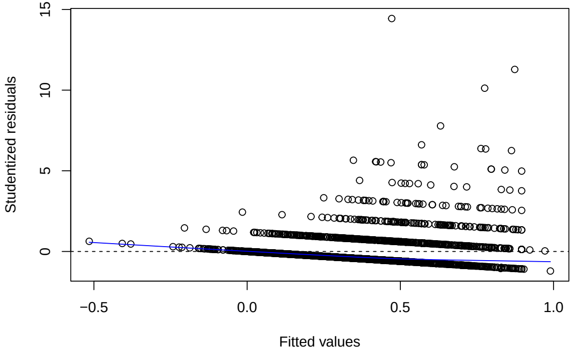
Heteroskedasticity: $Var(e_i) = h(X_1, \ldots, X_k)$.[5]

[5] $h()$ is a generic function of the predictors in the model, either linear or nonlinear.

Heteroskedasticity: diagnosis

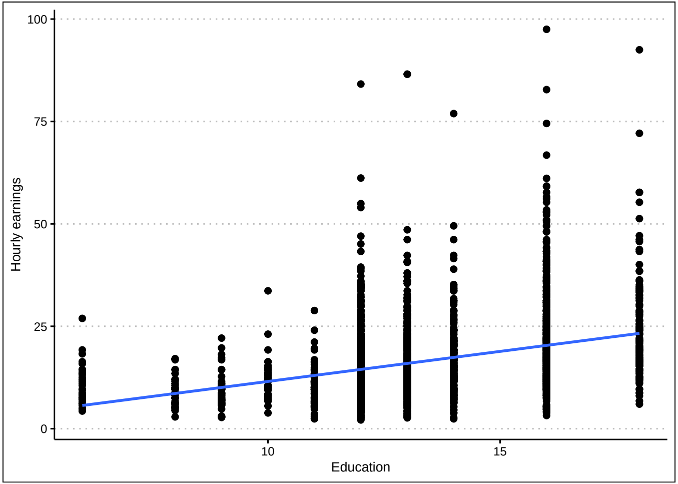
Ocular impact test

Does it hit you right between the eyes when you plot it?



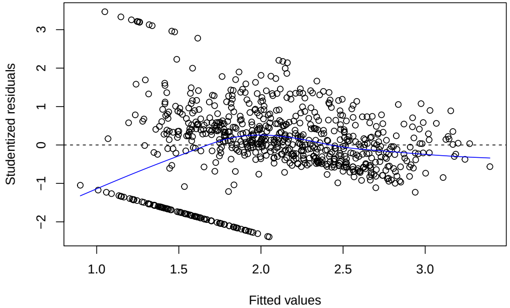
Predicting # of arrests in 1986

Ocular impact test



Predicting earnings for 29–30 year olds in US (2004)

Ocular impact test



Predicting students' GPAs in college

It can be effective, but only in the cases when there are glaring disparities between variances.

Statistical tests: Breusch–Pagan (I)

Take the standard form of the linear model:

$$Y_i = a + b_1 X_{1i} + \cdots + b_k X_{ki} + e_i \tag{6}$$

The null hypothesis of the test is that $Var(e_i|X_1, \ldots, X_k) = \sigma_e^2$.

What we want to check is that there is no association between e_i, and any function that can be produced with the Xs.

Statistical tests: Breusch–Pagan (II)

It's easiest to assume a linear form:

$$e_i^2 = \delta_0 + \delta_1 X_{1i} + \cdots + \delta_k X_{ki} + \upsilon \tag{7}$$

At this point, the last step is to just run an F test, or a Lagrange Multiplier test and check that $\delta_0 = \delta_1 = \cdots = \delta_k = 0$.[6]

This test is simply $LM = n \times R_*^2$, where R_*^2 is the model fit of the model in Equation 7, and n is the sample size.

[6]Technically, this second model would use ϵ_i, but these cannot be known, so they get replaced with e_i.

Statistical tests: Breusch–Pagan (III)

The value of the Breusch–Pagan test statistic (the LM from before) will have a χ^2 distribution with k degrees of freedom.

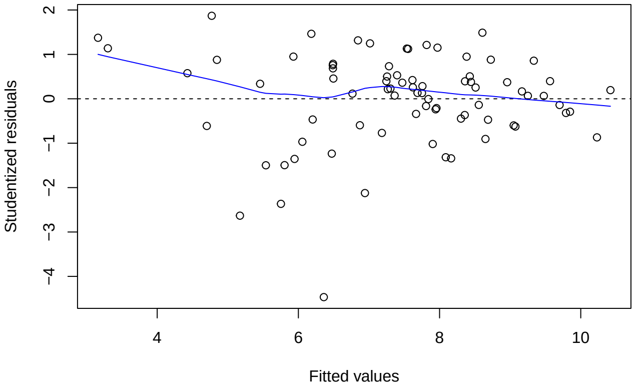
If you wanted to, you could also construct an F-test, using the same R^2_* value.

$$F = \frac{\frac{R^2_*}{k}}{(1 - R^2_*)(n - k - 1)} \tag{8}$$

This will have a $F_{k,n-k-1}$ distribution.[7]

[7] You won't need to know critical values for these distributions, as the software will do it automatically for you.

Statistical tests: Breusch–Pagan (IV)



Predicting GPA with IQ, gender, and self-concept

Statistical tests: Breusch–Pagan (V)

```
bptest(model1) # requires the "lm" object


studentized Breusch-Pagan test

data:  model1
BP = 7.953, df = 3, p-value = 0.04699
```

The most important fact to remember is H_0 (!): homoskedasticity.

In this case, we have to reject H_0 and accept that the data is heteroskedastic.

Statistical tests: White

Not as common as Breusch–Pagan, although it shares a lot of similarities.

It adds to Equation 7 all squares of the predictors, as well as all two-way interactions between predictors.

$$e_i^2 = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \delta_3 X_1^2 + \delta_4 X_2^2 + \delta_5 X_1 X_2 + \upsilon \quad (9)$$

It then proceeds either with an F-test or an LM test, as in the case of Breusch–Pagan.

A final word of caution

Tests are very convenient, and authoritative for a readership, but there is a catch.

If there is a problem with the functional form of the model, $E(Y|X)$, then it might well be that the test reveals heteroskedasticity, when in fact none exists if the "true" model were used.

This was the case with yesterday's example with Boston house prices.

Solution I: Heteroskedasticity-robust SEs

Robust SEs

We just concede that we don't know the form of $h(X_1, \ldots, X_k)$ that explains σ_e^2.

However, we know that the only worrying problem with heteroskedasticity are the SEs, not b.

Robust SEs fix just that.[8]

[8] Also known as heteroskedasticity-robust, heteroskedasticity-consistent, sandwich estimators, cluster-robust etc.

Robust SEs – simple regression

White (1980) shows how to obtain a valid estimator of $Var(b)$ even in conditions of heteroskedasticity.

$$V(b) = \frac{\sigma_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}[(x_i - \bar{X})^2\sigma_e^2]}{[\sum_{i=1}^{n}(x_i - \bar{X})^2]^2} \quad (10)$$

In cases of heteroskedasticity, there is no more constant σ_e^2, but a variable σ_i^2.

The top part no longer simplifies nicely with a variable σ_i^2.

Robust SEs – simple regression

However, the following modification *is* valid:

$$V(b) = \frac{\sum_{i=1}^{n}[(x_i - \bar{X})^2 e_i^2]}{[\sum_{i=1}^{n}(x_i - \bar{X})^2]^2} \tag{11}$$

The square root of this quantity will be the heteroskedasticity-consistent SE.[9]

[9] The bottom part of that fraction looks scary, but it's really the square of the total sum of squares: SST_x^2.

Robust SEs – multiple regression

$$V(b) = \frac{\sum_{i=1}^{n}(r_{ij}^2 e_i^2)}{SST_j(1 - R_j^2)} \tag{12}$$

- r_{ij}: the ith residual from a regression of X_j on all other predictors;
- SST_j: the total sum of squares of X_j;
- R_j^2: the R^2 from the regression of X_j on all other predictors.

Robust F-test and LM test

They can be computed, although for F-tests the formula does not have a simple form.

An LM test is simpler to conduct. Say that you have a full model

$$Y_i = a + b_1X_1 + b_2X_2 + b_3X_3 + e_i \tag{13}$$

and a restricted model

$$Y_i = a + b_1X_1 + u_i \tag{14}$$

Robust *LM* test (I)

The goal is to see whether $b_2 = b_3 = 0$.

First, take out the u_i from the restricted model.

Second, regress X_2 on X_1, and take out residuals $w1_i$. Also regress X_3 on X_1, and take out residuals $w2_i$.[10]

Third, compute $u_i w1_i$, and $u_i w2_i$ (and more products if we had more residuals from stage 2).

[10] If the restricted model would have had more predictors, we would have regressed X_2, and then X_3, on all these predictors.

Robust *LM* test (II)

Finally, run this regression (notice there is no intercept):

$$1 = \gamma_1 u_i w1_i + \gamma_2 u_i w2_i + \upsilon \tag{15}$$

Then $LM = n - SSR_1$, where n is the sample size and SSR_1 is the sum of squared residuals from the regression in Equation 15.

The LM test statistic will have a χ^2 distribution with q degrees of freedom, where q is the number of restrictions we imposed (here $q = 2$).

Solution II: Weighted Least Squares

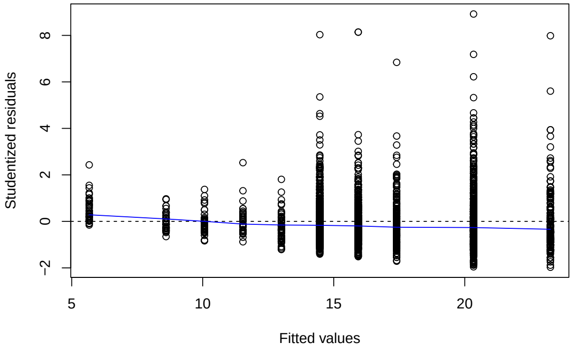
Weighted Least Squares

Suitable for when we can specify the functional form of the heteroskedasticity.

We can do this either by knowing the form beforehand (theory, studies), or by estimating it from the data itself.

If this is possible, WLS is more efficient than OLS in conditions of heteroskedasticity.

WLS I: simple version



Say that $Var(e_i|educ) = \sigma_e^2 h(educ)$, and that in this case $h(educ) = educ$.[11]

[11] I call this function h_i from now on, to simplify notation.

WLS I: simple version

Even though the variance of e_i is not constant, it turns out that the variance of $\frac{e_i}{\sqrt{h_i}}$ is constant and equal to σ_e^2.

This means we can use the $\frac{1}{\sqrt{h_i}}$ quantity as a weight, and re-specify the model of earnings.

$$Earn_i / \sqrt{educ_i} = a / \sqrt{educ_i} + b_1 \underbrace{educ_i / \sqrt{educ_i}}_{= \sqrt{educ_i}} + \underbrace{e_i / \sqrt{educ_i}}_{e_i^*} \tag{16}$$

The new errors, e_i^* are homoskedastic, with 0 conditional expectation.

WLS I: simple version

The new coefficients from this respecified model, $a^*, b_1^*, \ldots, b_k^*$ are GLS (generalized least squares) estimators.

They are more efficient than OLS estimators in this instance.

In practice, this procedure doesn't weight the variables themselves, but rather the e_i^2. The weights used are $\frac{1}{h_i}$. This means less weight is given to errors with higher variance.

Earnings specification

Model comparisons: OLS and GLS

	OLS	GLS	GLS through REML
(Intercept)	-3.134^{**}	-1.823^{*}	-1.823^{*}
	(0.959)	(0.840)	(0.840)
Education	1.467^{***}	1.370^{***}	1.370^{***}
	(0.070)	(0.063)	(0.063)
R^2	0.130	0.138	
Adj. R^2	0.130	0.138	
Num. obs.	2950	2950	2950
AIC			21032.547
BIC			21050.514
Log Likelihood			-10513.274

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

WLS II: slightly more complex

In case you don't have information about a clear function for the variance, you can always try estimating it.

$$Var(e_i|X_1, \ldots, X_k) = \sigma_e^2 exp(\delta_0 + \delta_1 X_1 + \cdots + \delta_k X_k) \tag{17}$$

The exponential function is preferred because it makes sure that the weights will always be positive.[12]

$$log(e^2) = \alpha + \delta_1 X_1 + \cdots + \delta_k X_k + \upsilon \tag{18}$$

[12] A linear specification would not insure this by default. $exp(a) = e^a$, where e is Euler's constant ≈ 2.71828.

FGLS

From Equation 18 all we need are the fitted values: $\hat{e}_i$.

Then we simply use $\frac{1}{exp(\hat{e}_i)}$ as the weights in the original regression.

Using the same data to estimate both weights and the model, means that FGLS estimates are not unbiased, but they are asymptotically consistent and more efficient than OLS.

WLS: final considerations

If estimates from OLS and WLS differ considerably, then there is a bigger problem than heteroskedasticity, e.g. perhaps model misspecification.

Getting the precise form of $h(x)$ right is not a big concern, as the estimates will still be consistent asymptotically.

With a wrong form of $h(x)$ there is no guarantee that WLS estimates are more efficient than OLS, but it's still better to use it than rely on OLS.

Thank you for the kind attention!

References

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817–838.