Multiple regression analysis: Heteroskedasticity

Prof. Sebastián Vallejo Vera
University of Western Ontario

Heteroskedasticity

Recall the homoskedasticity assumption MLR.5 about the variance of the error term:

Assumption MLR.5: Homoskedasticity

The error u has the same variance given any values of the explanatory variables. In other words:

$$Var(u|x_1, x_2, ..., x_k) = \sigma^2$$

When that assumption does not hold, we say that we have *heteroskedasticity*.

Heteroskedasticity

The presence of heteroskedasticity does not cause bias in the OLS estimators of the β_j, but the estimators of the variances, $Var(\hat{\beta}_j)$, are biased without the homoskedasticity assumption.

And, because the OLS standard errors are based directly on these variances, they are no longer valid for constructing confidence intervals and calculating t and F statistics. In other words, the statistics we use to test hypotheses under the Gauss-Markov assumptions (MLR.1 through MLR.5) are not valid in the presence of heteroskedasticity.

Moreover, in the presence of heteroskedasticity, OLS is no longer BLUE.

Heteroskedasticity

The question is: how do we detect heteroskedasticity?

Frequently, we have theoretical reasons to suspect heteroskedasticity like in the example about wages as a function of education. At other times, however, we may be agnostic about the theoretical presence or not of heteroskedasticity.

Thankfully, it is possible to test empirically for heteroskedasticity. One common test is the **Breusch-Pagan test for heteroskedasticity**.

Heteroskedasticity

The idea behind the Breusch-Pagan test for heteroskedasticity is quite simple. Suppose we have the following linear model:

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

where the assumptions MLR.1 through MLR.4 hold.

We take the null hypothesis to be that assumption MLR.5 is true:

$$H_0 : Var(u|x_1, x_2, ..., x_k) = \sigma^2$$

Because we assume that $E(u|\boldsymbol{x}) = 0$ in MLR.4, then $Var(u|\boldsymbol{x}) = E(u^2|\boldsymbol{x})$ and we can rewrite H_0 as follows:

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

Heteroskedasticity

To evaluate the homoskedasticity assumption, we want to test whether u^2 is related (in expected value) to one or more of the explanatory variables. If H_0 is false, the expected value of u^2, given the independent variables, can be virtually any function of the x_j. A simple approach is to assume a linear function:

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + v$$

where v is an error term with mean zero given the x_j.

The null hypothesis of homoskedasticity is:

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

Heteroskedasticity

Recall, however, that the errors in the population are unknown. We can use instead the OLS residual, $\hat{u}_i$, as an estimate of the error u_i for observation i, and estimate the following equation:

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + error \tag{1}$$

and compute the F statistics for the joint significance of $x_1, x_2, ..., x_k$ using the R^2 from equation (2), labeled $R^2_{\hat{u}^2}$, using this formula for the F statistics:

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

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

Heteroskedasticity

Most statistical softwares will perform the Breusch-Pagan test for heteroskedasticity, but it is also simply done by following a few easy steps:

1. Estimate the following regression model by OLS:

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

and obtain the squared residuals, $\hat{u}_i^2$.

2. Estimate next the following regression model, also by OLS:

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + error \tag{3}$$

and keep the R^2 from this regression, $R^2_{\hat{u}^2}$.

3. Form the F statistic and compute the $p - value$ (using the $F_{k,n-k-1}$ distribution). If the $p - value$ is sufficiently small, that is, below the chosen significance level, then we reject the null hypothesis of homoskedasticity.

Heteroskedasticity

If the Breusch-Pagan test results in a small enough $p - value$, some corrective measure should be taken.

One possibility is to use the *heteroskedasticity-robust standard errors* provided by most statistical softwares and their associated robust t and F test statistics (with the corrected standard errors). This is the most commonly used solution but there are others, including *weigthed least squares* (WLS) when the form of the hetreroskedasticity is known or *feasible generalized least squares (FGLS)* when it is not.

Heteroskedasticity

The idea behind the heteroskedasticity-robust standard errors is to allow said standard errors to be a function of the model's independent variables. Specifically, under MLR.1 through MLR.4, it can be shown that a valid estimator of $Var(\hat{\beta}_j)$ is:

$$\widehat{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{RSS_j^2} \tag{4}$$

where $\hat{r}_{ij}^2$ denotes the i^{th} residual from regressing x_j on all other independent variables, and RSS_j is the sum of squared residuals from this regression.

The square root of $\widehat{Var}(\hat{\beta}_j)$ is the heteroskedasticity-robust standard error for $\hat{\beta}_j$.

These are readily computed in most statistical software (so no need to compute them manually).

Heteroskedasticity

So, if we can easily compute heteroskedasticity-robust standard error, why would we ever compute OLS standard errors–which would be subject to bias?

If (and only if) the assumption of homoskedasticity is valid, the OLS standard errors are preferred, since they will have an exact t-distribution at any sample size.

The application of robust standard errors is justified as the sample size becomes large.

Heteroskedasticity

If we are working with a sample of modest size, and the assumption of homoskedasticity is tenable, we should rely on OLS standard errors. But since robust standard errors are very easily calculated in most statistical packages, it is a simple task to estimate both sets of standard errors for a particular equation, and consider whether inference based on the OLS standard errors is fragile.

In large data sets, it has become increasingly common practice to report the robust standard errors.

In-class exercise, part 1

Using the 'fertil2' dataset from 'wooldridge' on women living in the Republic of Botswana in 1988, estimate the regression equation of the number of children on education (*educ*), age of the mother (*age*) and its square, electricity (*electric*), husband's education (*heduc*), and whether the women has a radio (*radio*) and/or a TV (*tv*) at home.

- interpret the effect of *age* on *children* and find the turning point;
- replace the quadratic functional form of *age* for a logarithmic form instead, that is, replace age and its square with just $log(age)$. What do you conclude?
- test the regression model for heteroskedasticity by using the three steps presented above and comparing it with the function provided in R for the Breusch-Pagan test.