Econometrics I TA Session

Wang Xin

June 30, 2022

Contents

1 Testing for Heteroskedasticity

As usual, we start with the linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \tag{1}$$

where other assumptions are maintained here. In particular, we assume that $E(u|x_1, x_2, \cdots, x_k) = 0$, so that OLS is unbiased and consistent.

We take the null hypothesis to be that homoskedasticity assumption is true:

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

If we cannot reject H_0 at a sufficiently small significance level, we usually conclude that heteroskedasticity is not a problem. However, remember that we never accept H_0; we simply fail to reject it.

Because we are assuming that the error term u has a zero conditional expectation, $V(u|\mathbf{x}) = E(u^2|\mathbf{x})$, and so the null hypothesis of homoskedasticity is equivalent to

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

This shows that, in order to test for violation of 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

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

where v is an error term with $E(v|x_1, x_2, \cdots, x_k) = 0$ and is independent of $x_1, x_2, \cdots, x_k$. The null hypothesis of homoskedasticity is

$$H_0 : \delta_1 = \delta_2 = \cdots = \delta_k = 0 \tag{2}$$

We never know the actual errors in the population model, but we do have estimates of them: the OLS residual, $\hat{u}_i$, is an estimate of the error u_i for observation i. Thus, we can estimate the equation

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

We have learned that the F statistics for the overall significance. Similarly, it can be used to test (2), which is

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

The LM statistic for heteroskedasticity is just the sample size times the R-squared

$$LM = n * R^2_{\hat{u}^2} \sim \mathcal{X}^2(k)$$

The LM version of the test is typically called the **Breusch-Pagan test for heteroskedasticity (BP test)**.

We summarize the steps for testing for heteroskedasticity using the BP test:

1. Estimate the model (1) by OLS, as usual. Obtain the squared OLS residuals, $\hat{u}_i^2$ (one for each observation).

2. Run the regression in (3). Keep the R-squared from this regression, $R^2_{\hat{u}^2}$.

3. Form either the F statistic or the LM statistic and compute the $p - value$ (using the $F_{k,n-k-1}$ distribution in the former case and the $\mathcal{X}^2_k$ distribution in the latter case). If the $p - value$ is sufficiently small, that is, below the chosen significance level, then we reject the null hypothesis of homoskedasticity.

2 Estimated Heteroskedasticity: Feasible GLS

In lectures, we saw some examples of where the heteroskedasticity is known up to a multiplicative form. In most cases, the exact form of heteroskedasticity is not obvious. In other words, it is difficult to find the function $\sigma^2\Omega$. Nevertheless, in many cases we can model the function $\sigma^2\Omega$ as $\sigma^2 h(\mathbf{x})$ and use the data to estimate the unknown parameters in this model. This results in an estimate of Ω, denoted as $\hat{h}$. Using $\hat{h}$ instead of $\Omega(h(\mathbf{x}))$ in the GLS transformation yields an estimator called the feasible GLS (FGLS) estimator or EGLS.

Assume that

$$V(u|\mathbf{x}) = \sigma^2 exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k) \tag{4}$$

that is $h(\mathbf{x}) = exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k)$.

Attention, when testing for heteroskedasticity using the Breusch-Pagan test, we assumed that heteroskedasticity was a linear function of the $\mathbf{x}$. Linear alternatives are fine when testing for heteroskedasticity, but they can be problematic when correcting for heteroskedasticity using weighted least squares. We have encountered the reason for this problem before: linear models do not ensure that predicted values are positive, and our estimated variances must be positive in order to perform WLS.

Since we do not know the true value of (4) which means that δ_j is unknown, we will transform this equation into a linear form that, with slight modification, can be estimated by OLS.

Under assumption (4), we can write

$$u^2 = \sigma^2 exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k) v$$

where v has a mean equal to unity, conditional on $\mathbf{x}$. If we assume that v is actually independent of $\mathbf{x}$, we can write

$$\log(u^2) = \lambda + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + e$$

where e has a zero mean and is independent of $\mathbf{x}$; the intercept in this equation is different from δ_0, but this is not important in implementing WLS.

As usual, we must replace the unobserved u with the OLS residuals. Therefore, we run the regression of

$$\log(\hat{u}^2) = \lambda + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + e \tag{5}$$

From the unbiased estimators of the δ_j by using OLS, the fitted values of $\log(\hat{u}^2)$, $\hat{g}$ can be calculated. Then, the estimates of h_i are simply $\hat{h} = exp(\hat{g})$

We summarize the steps for FGLS.

1. Run the regression of y on $x_1, x_2, \cdots, x_k$ and obtain the residuals, $\hat{u}$.

2. Creat $\log(\hat{u}^2)$ from the OLS residuals in Step 1.

3. Run the regression in equation (5) and obtain the fitted values, $\hat{g}$.

4. Exponentiate the fitted values from (5): $\hat{h} = exp(\hat{g})$.

5. Estimate the equation (1) by WLS, using weights $1/\sqrt{\hat{h}}$. In other words, we replace Ω with $\hat{h}$.

3 A Simple Example

We consider the following model to estimate a demand function for daily cigarette consumption. Since most people do not smoke, the dependent variable, *cigs*, is zero for most observations. A linear model is not ideal because it can result in negative predicted values. Nevertheless, we can still learn something about the determinants of cigarette smoking by using a linear model.

The equation estimated by ordinary least squares, with the usual OLS standard errors in parentheses, is

$$\begin{aligned}\widehat{cigs} &= \underset{(24.08)}{-3.64} + \underset{(0.728)}{0.880}\log(income) - \underset{(5.773)}{0.751}\log(cigpric) - \underset{(0.167)}{0.501}educ \\ &\quad + \underset{(0.160)}{0.771}age - \underset{(0.0017)}{0.0090}age^2 - \underset{(1.11)}{2.83}restaurn \\ n &= 807,\ R^2 = 0.0526,\end{aligned} \tag{6}$$

where *cigs* number of cigarettes smoked per day, *income* is annual income, *cigpric* is the per-pack price of cigarettes (in cents), *educ* is years of schooling, *age* is age measured in years, *restaurn* is a binary indicator equal to unity if the person resides in a state with restaurant smoking restrictions.

Neither income nor cigarette price is statistically significant. Each year of education reduces the average cigarettes smoked per day by one-half of a cigarette, and the effect is statistically significant. Cigarette smoking is also related to age, in a quadratic fashion. Smoking increases with age up until $age = 0.771/(2 \times 0.009) \approx 42.83$, and then smoking decreases with age. Both terms in the quadratic are statistically significant. The presence of a restriction on smoking in restaurants decreases cigarette smoking by almost three cigarettes per day, on average.

To test the Heteroskedasticity, the Breusch-Pagan regression of the squared OLS residuals on the independent variables in (6) produces $R^2_{\hat{u}^2} = 0.040$ [see equation (3)]. The LM statistic is $LM = 807 \times 0.040 = 32.28$, and this is the outcome of a $\mathcal{X}^2_6$ random variable. The $p - value$ is less than 0.000015, which is very strong evidence of heteroskedasticity.

Therefore, we estimate the equation using the feasible GLS procedure based on equation (5). The weighted least squares estimates are

$$\begin{aligned}\widehat{cigs} &= \underset{(17.80)}{5.64} + \underset{(0.44)}{1.30}\log(income) - \underset{(4.46)}{2.94}\log(cigpric) - \underset{(0.120)}{0.463}educ \\ &\quad + \underset{(0.097)}{0.482}age - \underset{(0.0009)}{0.0056}age^2 - \underset{(0.080)}{3.46}restaurn \\ n &= 807,\ R^2 = 0.1134,\end{aligned} \tag{7}$$

The income effect is now statistically significant and larger in magnitude. The price effect is also notably bigger, but it is still statistically insignificant. [One reason for this is that *cigpric*varies only across states in the sample, and so there is much less variation in log(*cigpric*) than in log(*income*), *educ*, and *age*.]

The estimates on the other variables have, naturally, changed somewhat, but the basic story is still the same. Cigarette smoking is negatively related to schooling, has a quadratic relationship with age, and is negatively affected by restaurant smoking restrictions.