ECON 3740: INTRODUCTION TO ECONOMETRICS

INSTRUCTOR: CHAOYI CHEN
Department of Economics and Finance, University of Guelph

Lecture 21

Lecture outline

Last lecture, we studied the consequences with heteroskedasticity and suggested two tests for the heteroskedasticity. Today, we will

- Learn the White test
- Weighted Least Square
  - The heteroskedasticity is known - GLS
  - The heteroskedasticity is unknown - Feasible GLS
- Heteroskedasticity for OLS: Summary

Heteroskedasticity for OLS: Testing for Heteroskedasticity - Method III: White Test

- Similar to the BP test, White test depends on regressing $\widehat{\mu}^2$ on all explanatory variables, their squares, and interactions. Asumming we have three explanatory variables

$$\widehat{\mu}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^3 \\ + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + error$$

- Here, with $k = 3$, we have 9 regressors. Generally, if we have k explanatory variables, we have $\frac{k(k+3)}{2}$ regressors.
- The null hypothesis is

$$H_0 : \delta_1 = ... = \delta_9 = 0$$

  That is the White test detects more general deviations (not only linear form but quadratic form in $\mathbf{x}_i$) from homoskedasticity than the BP test.
- Hence, we can apply F or LM test to detect heteroskedasticity.

Heteroskedasticity for OLS: Testing for Heteroskedasticity - Method III: Another Form of the White Test

- Including all squares and interactions leads to a large number of estimated parameters. For example, $k = 6$ leads to 27 parameters to be estimated. As a result, it uses many degrees of freedom for models with just a moderate number of independent variables.
- Alternatively, we can use the following form for the White test:

$$\widehat{\mu}^2 = \delta_0 + \delta_1 \widehat{y} + \delta_2 \widehat{y}^2 + error,$$

  where $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_k x_k$ is the fitted value.
- It is important not to confuse $\widehat{y}$ and y in this equation. We use the fitted values because they are functions of the independent variables (and the estimated parameters). That is, $\delta_1 \widehat{y} + \delta_2 \widehat{y}^2$ is a special quadratic form of $\mathbf{x}_i$. Wrongly using y will produce an invalid test.

Heteroskedasticity for OLS: Testing for Heteroskedasticity - Method III: Another Form of the White Test Continue

- The null here is

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

  and the F or LM statistic can apply for this hypothesis test. (For example, $LM = nR^2_{\widehat{\mu}^2} \sim \chi^2_2$).

- An example: Recall the Log Housing Price problem. Suppose that we regress $\widehat{\mu}^2$ on $\widehat{log(price)}$ and $\widehat{log(price)}^2$ and obtain $R^2_{\widehat{\mu}^2} = 0.0392$. Therefore,

$$LM = 88 * 0.0392 \approx 3.45 \text{ with } p\text{-value} = 0.178$$

  Hence, we cannot reject the model is homoskedastic, which is consistent with the BP test.

Heteroskedasticity for OLS: Weighted Least Square Estimation and GLS

- Motivation:As we mentioned, the OLS will still be unbiased under heteroskedasticity but lose the efficiency. If we have correctly specified the form of the variance, we could propose a more efficient estimator than OLS. We consider two cases to discuss the more efficeint least square estimator.
  - The Heteroskedasticity is Known up to a multiplicative constant - Weighted Least Square (WLS)
  - The Heteroskedasticity Function is unknown and needs to be estimated- GLS Must Be Estimated

- Case 1: The Heteroskedasticity is Known up to a multiplicative constant
- Suppose $Var(\mu|\mathbf{x}) = \sigma^2 h(\mathbf{x})$, where $h(\mathbf{x}) > 0$ is known, but σ^2 is unkown.
- Hence, we can compute the value for $h(\mathbf{x}_i) = h_i$ for $i = 1, ..., n$.
- Consider the regression model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + mu_i.$$

Heteroskedasticity for OLS: Weighted Least Square Estimation and GLS Continue

- Now, if we know the h_i), we can tranform the model by dividing both sides by $\sqrt{h_i}$ and we have,

$$\frac{y_i}{\sqrt{h_i}} = \beta_0 \frac{1}{\sqrt{h_i}} + \beta_1 \frac{x_{i1}}{\sqrt{h_i}} + ... + \beta_k \frac{x_{ik}}{\sqrt{h_i}} + \frac{\mu_i}{\sqrt{h_i}},$$

- Why we transform in this way? We will have a "cleaned" new error term ($\frac{\mu_i}{\sqrt{h_i}}$ is homoskedastic).
- The reason is, note that,

$$E[\frac{\mu_i}{\sqrt{h_i}}|\mathbf{x}_i] = \frac{1}{\sqrt{h_i}} = 0 \text{ by MLR.4}$$

$$Var(\frac{\mu_i}{\sqrt{h_i}}|\mathbf{x}_i) = E[(\frac{\mu_i}{\sqrt{h_i}})^2|\mathbf{x}_i] - E[(\frac{\mu_i}{\sqrt{h_i}})|\mathbf{x}_i]^2$$

$$= E[(\frac{\mu_i}{\sqrt{h_i}})^2|\mathbf{x}_i] = \frac{1}{h_i}E[\mu_i^2|\mathbf{x}_i] = \frac{1}{h_i}\sigma^2 h_i = \sigma^2.$$

- The OLS of the transformed model is called *weightedleastsquare*.
- Intuition: Why is WLS more efficient than OLS in the original model? Observations with a large variance are less informative than observations with small variance and therefore should get less weight.

Heteroskedasticity for OLS: Weighted Least Square Estimation and GLS Continue

- Notice that the transformed regression model has no intercept.
- Hence, we should be cautious about the R^2 in practice.
- There are two ways to compute $R^2 = 1 - \frac{SSR}{TSS}$.
  - (1): TSS is calculated with dependent variable y_i ($TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$), and the SSR is based on $\widehat{\mu}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - ... - \widehat{\beta}_k x_{ik}$. In this case, R^2 for WLS is smaller than OLS although WLS is more efficient than OLS. This is because they share the same SST, but OLS minimizes SSR. This is the correct R^2 to be reported
  - (2): TSS is calculated with dependent variable $\frac{y_i}{\sqrt{h_i}}$ ($TSS = \sum_{i=1}^{n}(\frac{y_i}{\sqrt{h_i}} - \frac{\bar{y}_i}{\sqrt{h}})^2$), and the SSR is based on $\tilde{\mu}_i = \frac{y_i}{\sqrt{h_i}} - \beta_0 \frac{1}{\sqrt{h_i}} - \beta_1 \frac{x_{i1}}{\sqrt{h_i}} - ... - \beta_k \frac{x_{ik}}{\sqrt{h_i}}$. In this case, it is not correct because the transformed model does not include an intercept which is required for R^2 calculation.

Heteroskedasticity for OLS: Unknown Heteroskedasticity Function (Feasible GLS)

- Case 2: If unfortunately, we do not know the heteroskedastic form, we can model function h. This results in an estimate of each h_i denotes as $\widehat{h}_i$. USing $\widehat{h}_i$ istead of h_i to transform and estimate the model yields an estimator called feasible GLS (FGLS) estimator.
- There are many ways to model heteroskedasticity, but we will study one particular, fairly flexible approach. Assume that $Var(\mu|\mathbf{x}) = \sigma^2 e^{\delta_0+\delta_1 x_1+...+\delta_k x_k} = \sigma^2 h(\mathbf{x})$, where exponential function ensures positivity.
- Hence, under this assumption, we have

$$\mu^2 = \sigma^2 e^{\delta_0+\delta_1 x_1+...+\delta_k x_k} v$$

  where v is a multiplicative error independent of the explanatory variables with $E[v] = 1$.
- Take a log gives,

$$log(\mu^2) = \alpha_0 + \delta_1 x_1 + ... + \delta_k x_k + e$$

  where e has a zero mean and is independent of $\mathbf{x}$.

Heteroskedasticity for OLS: Unknown Heteroskedasticity Function (Feasible GLS) Continue

- Then, we regress $\log(\widehat{\mu}^2)$ on $x_1, ..., x_k$ and obtain the fitted values, call these $\widehat{g}_i$. Then, we can estimate the unknown function h_i simply as

$$\widehat{h}_i = e^{\widehat{g}_i}$$

.

- A generalized steps for a feasible GLS can be summarized as:
  - Step 1: Regress y ON $x_1, ..., x_k$ and obtain the residuals $\widehat{\mu}$.
  - Step 2: Compute $log(\widehat{\mu}^2)$. That is squaring the OLS residuals and taking log.
  - Step 3: Regress $\log(\widehat{\mu}^2)$ on $x_1, ..., x_k$ and obtain the fitted values $\widehat{g}$.
  - Step 4: Compute the estimates of h_i: $\widehat{h}_i = e^{\widehat{g}}$.
  - Step 5: (Get back to the case that heteroskedasticity function is known) Estimate the model by WLS by using $h = \widehat{h}$.

Heteroskedasticity for OLS: Unknown Heteroskedasticity Function (Feasible GLS) - An Example

- An example: Consider the followingh estimated demand equation for cigarettes by OLS,

$$\begin{array}{llll} \widehat{cigs} = & -3.64 & +0.88 log(income) & -0.75 log(cigpric) \\ & (24.08) & (0.728) & (5.773) \\ & -0.501 educ & +0.771 age & -0.009 age^2 \\ -2.83 restaurn & & & \\ & (0.167) & (0.16) & (0.0017) \\ (1.11) & & & \end{array}$$

where $n = 807$, $R^2 = 0.0526$, and

$$\begin{aligned} cigs &= \text{number of cigarettes smoked per day} \\ income &= \text{annual income} \\ cigpric &= \text{the per-pack price of cigarettes (in cents)} \\ restaurn &= \text{binary indicator for smoking restrictions in restaurants} \end{aligned}$$

- Note that the income effect is insignificant.
- The p value for the BP test (either F or LM) is .000, which suggests there is a strong evidence that the model is heteroskedastic.

Heteroskedasticity for OLS: Unknown Heteroskedasticity Function (Feasible GLS) - An Example Continue

- Now, if we estimated the model by FGLS, we have

$$\widehat{cigs} = \underset{(17.8)}{-5.64} + \underset{(0.44)}{1.3 log(income)} - \underset{(4.46)}{2.94 log(cigpric)}$$
$$- \underset{(0.12)}{0.463 educ} + \underset{(0.097)}{0.482 age} - \underset{(0.0009)}{0.0056 age^2}$$
$$- \underset{(0.8)}{3.46 restaurn}$$

where $n = 807$, $R^2 = 0.1134 > (0.0526)$

- Clearly, the income effect is now statistically significant and other coefficients are also more precisely estimated (without changing qualitative results).
- Interestingly, the turnaround point ($\frac{\partial}{\partial age} = \beta_{age} + 2 * \beta_{age^2} age = 0$) of the age in both model is about the same with 43:

$$\frac{0.771}{2 * 0.009} \approx \frac{0.482}{2 * 0.0056} \approx 43.$$

Heteroskedasticity for OLS: Summary

- In this topic, we have thoroughly discussed the heteroskedasticity.
- We began by reviewing the properties of OLS in the presence of heteroskedasticity. Heteroskedasticity does not cause bias in the OLS estimators, but the usual standard errors and test statistics are no longer valid. Hence, we showed how to compute heteroskedasticity-robust standard errors.
- Next, we introduced three common ways to detect for heteroskedasticity: the graphical method, the BP test, and the White test. For the latter two tests, they rely on regressing squared OLS residuals on either independent variables (BP test) or the fitted and squared fitted values (White test). Both F and LM test can be applied.
- Then, since the OLS is no long most efficient estimator with heteroskedasticity, we show GLS or FGLS estimation can be used to attain the efficiency. First, with the known heteroskedastic form, GLS can achieve the BLUE estimator properties. Secondly, if the heteroskedastic form is unknown, feasible GLS can be applied. (Note that FGLS is no longer unbiased, but it is consistent and asymptotically efficient.).