ECON 3740: INTRODUCTION TO ECONOMETRICS

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

Lecture 19

Lecture outline

Last lecture, we studied dummy variables for multiple categories. Today, we will

- Test for differences in regression functions across groups
- Summary: MLR further issues

- Heteroskedasticity for OLS
  - Define the terminology
  - Learn consequences with heteroskedasticity
  - Study a heteroskedasticity-robust inference

MLR, Further Issue: Allowing for Different Slopes, An Example

- Consider following fitted regression line

$$
\begin{array}{rccc}
\widehat{log(wage)} = & 0.389 & \color{red}{-0.227} female & +0.082 educ \\
 & (0.119) & (0.168) & (0.008) \\
 & \color{red}{-0.0056} female * educ & +0.029 exper & -0.00058 exper^2 \\
 & (0.0131) & (0.005) & (0.00011) \\
 & +0.032 tenure & -0.00059 tenure^2 & \\
 & (0.007) & (0.00024) &
\end{array}
$$

  where $n = 526$ and $R^2 = 0.441$

- Consider the null $H_0 : \beta_{female*educ} = 0$. $|t_{female*educ}| = |\frac{-0.0056}{0.0131}| = 0.43 < 1.96$. Hence, no evidence against hypothesis that the return to education is the same for men and women.
- Consider the null $H_0 : \beta_{female} = 0$. $|t_{female}| = |\frac{-0.227}{0.168}| = 1.35 < 1.96$. Does this mean that there is no significant evidence of lower pay for women at the same levels of $educ, exper, andtenure$? **No**, this is only the effect for $educ = 0$ because

$$
\frac{\partial log(wage)}{\partial female} = -0.227 - 0.0056 educ \tag{1}
$$

MLR, Further Issue: Testing for Differences in Regression Functions across Groups

- Let's consider a F test with the unrestricted model containing full set of interactions,

$$cumgpa = \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female * sat + \beta_2 hsperc + \delta_2 female * hsperc + \beta_3 tothrs + \delta_3 female * tothrs + \mu$$

and the restricted model with same regression for both group,

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + \mu$$

where

$cumgpa$= college cumulative GPA
sat=standardized aptitude test score
$hsperc$= high school rank percentile
$tothrs$ = total hours spent in college courses

- The null hypothesis is

$$H_0 : \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0,$$

- Under the null, the model is the same for male and female, which gets back to the restricted model.

MLR, Further Issue: Estimation of the Unrestricted Model

- The estimated unrestricted model is

$$\begin{aligned}\widehat{cumgpa} = & \underset{(0.21)}{1.48} \quad \underset{(0.411)}{-0.353 female} \quad \underset{(0.0002)}{+0.0011 sat} \\ & \underset{(0.00039)}{+0.00075 female * sat} \quad \underset{(0.0014)}{-0.0085 hsperc} \quad \underset{(0.00316)}{-0.00055 female * hsperc} \\ & \underset{(0.0009)}{+0.0023 tothrs} \quad \underset{(0.00163)}{-0.00012 female * tothrs}\end{aligned}$$

  where $n = 366$, $R^2 = 0.406$

- It can be shown (proof not required) that

$$SSR_{ur} = SSR_{male} + SSR_{female}$$

  where SSR_{male} is the SSR in the regression

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + \mu$$

  using **only** the data of male, and SSR_{female} is the SSR using **only** the data of female.

MLR, Further Issue: Testing Results

- Tested individually, the hypothesis that the interaction effects are zero cannot be rejected.
- Tested jointly, the F statistic is

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(85.515 - 78.355)/4}{78.355/(366-7-1)} \approx 8.18$$

  and by checking the F table, the null is rejected.
- An alternative way to compute SSR_{ur} is through the estimation results of both male only regression and female only regression.
  $SSR_{ur} = SSR_{male} + SSR_{female} = 58.752 + 19.603 = 78.355$,
  $n = n_{male} + n_{female} = 276 + 90 = 366$.
- This relationship is true only if all interaction terms are included in the unrestricted model.
- If the test is computed in this way, it is called the Chow-Test
- Caution: Chow-Test assumes a constant error variance across groups as assumed in the F test.

MLR, Further Issue Summary

- In this topic, we have introduced several further issues regarding the MLR
- First, we showed that, as we include the quadratic term of one independent variable into the model, the marginal effect of that specific independent variable on the dependent variable will not be a constant. It will be a linear function about that independent variable itself.
- Next, we studied that if we add the interation term (say, $x_1 x_2$) into the model. The marginal effect of x_1 will be a linear function of x_2 and vice versa.
- Then, we learned the the adjusted R^2 and its relationship with R^2. We show how to use R^2 to compare Nonnested Models.
- We also illustrated when we should add regressors into the model. Due to the fact of a tradeoff between decraese in the error variance and exacerbate in the multicollinearity, the choice in adding more regressors into the model should be cautious.

MLR, Further Issue Summary Continue

- Next, we explained the problem of predicting y when $log(y)$ is the dependent variable. We showed that, because of the *Jensens* inequality, $e^{\widehat{y}}$ will under estimate $E[y|\mathbf{x}]$.
- Later on, we move to focus on the dummy. We firstly provides the motivation to use a dummy - to represent the qualitative information.
- Then, we used an example to illustrate how to use a single dummy and how to avoid the dummy variable trap.
- Next, we extended the single dummy to use dummy variables for multiple categories. We studied how to identify the base category and the economic meaning in explaining the dummy coefficient.
- We also included the interaction among dummy variables into the model. We showed a "difference in difference" effect. Furthermore, we also learned that the model will allow for different slopes if we add the interaction terms between a dummy and a slope regressor into the model.
- Finally, we discussed the test for differences in regression functions across groups in today's lecture.

Heteroskedasticity for OLS: Definition

- Recall that if

$$Var(\mu_i|\mathbf{x}_i) = \sigma^2$$

is constant, that is, if the variance of the conditional distribution of μ_i given x_i does not depend on x_i , then μ_i is said to be homoskedastic.

- Otherwise, if

$$Var(\mu_i|\mathbf{x}_i) = \sigma^2(\mathbf{x}_i) = \sigma_i^2$$

that is, the variance of the conditional distribution of μ_i given $\mathbf{x}_i$ depends on $\mathbf{x}_i$, then μ_i is said to be heteroskedastic.

- Recall, following assumption MLR.4, $E[\mu_i|\mathbf{x}_i] = 0$, $Var(\mu_i|\mathbf{x}_i) = E[\mu_i^2|\mathbf{x}_i] - E[\mu_i|\mathbf{x}_i]^2 = E[\mu_i^2|\mathbf{x}_i]$.

Heteroskedasticity for OLS: Graphical illustration for *homoskedasticity*

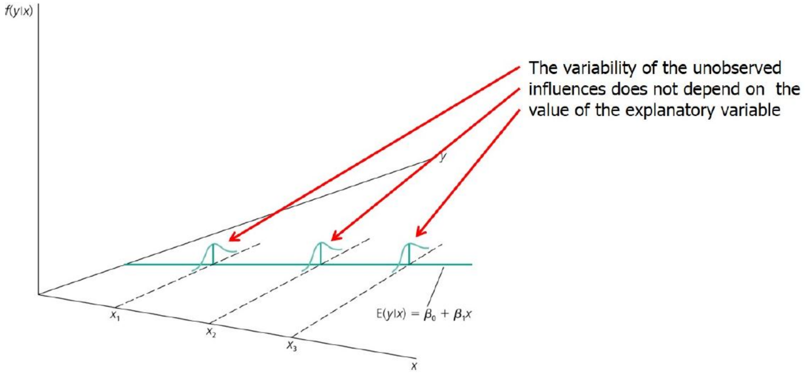

Figure: An Example for *Homoskedasticity*

Heteroskedasticity for OLS: Graphical illustration for *heteroskedasticity*

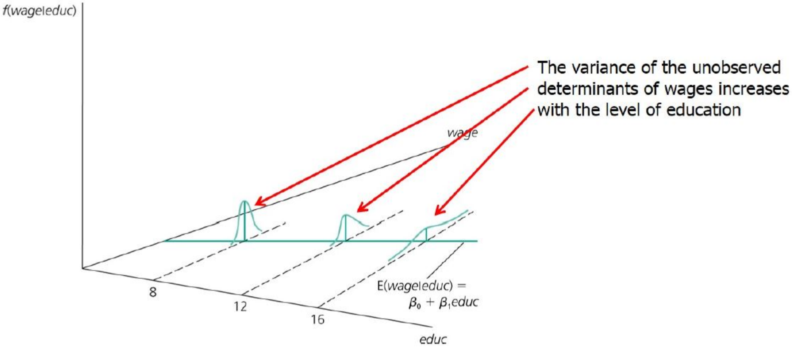

Figure: An Example for *Heteroskedasticity*

Heteroskedasticity for OLS: A Real-Data Example of Heteroskedasticity

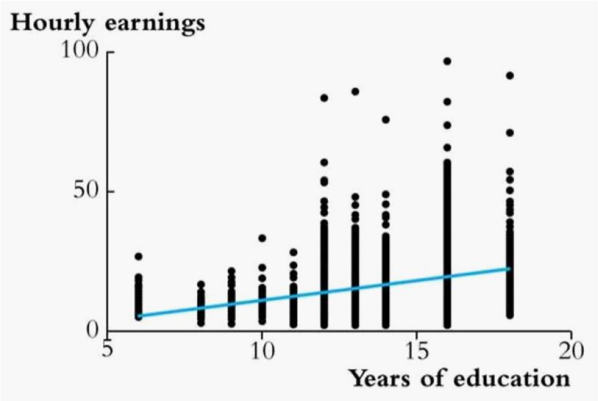

Figure: Average Hourly Earnings vs. Years of Education (data source: Current Population Survey)

Heteroskedasticity for OLS: Consequences

- OLS is still unbiased under heteroskedasticity because Assumption MLR.4 $E[\mu_i|\mathbf{x}_i] = 0$ does **not** involve conditional variance.
- Also, interpretation of R^2 and adjusted R^2 is **not** changed because

$$R^2 \approx 1 - \frac{\sigma_\mu^2}{\sigma_y^2}$$

  where σ_{mu}^2 is the unconditional variance of μ while heteroskedasticity is about the conditional variance of μ.
- However, heteroskedasticity invalidates variance formulas for OLS estimators.
- Hence, the usual F tests and t tests are **not** valid under heteroskedasticity because as mentioned before, normality assumption implies homoskedasticity.
- Under heteroskedasticity, OLS is **no longer** the best linear unbiased estimator (BLUE). There may be a more efficient linear estimator.

Heteroskedasticity for OLS: Heteroskedasticity-Robust Inference

- Formulas for OLS standard errors and related statistics have been developed that are robust to heteroskedasticity of unknown form.
- All formulas are only valid in large samples. (related to chapter five of the textbook, which will not be covered in this course)
- Formula for heteroskedasticity-robust OLS standard error is

$$\widehat{Var(\widehat{\beta}_j)} = \frac{\sum_{i=1}^{n} \widehat{r}_{ij}^2 \widehat{\mu}_i^2}{SSR_j^2} = SSR_j^{-1}[\sum_{i=1}^{n} \widehat{r}_{ij}^2 \widehat{\mu}_i^2]SSR_j^{-1}$$

  which is also called as Eicker/Huber/White standard errors or sandwich form standard errors. They involve the squared residuals from the regression, $\widehat{\mu}_i$ and from a regression of x_j on all other explanatory variables, $\widehat{r}_{ij}$.
- Using these formulas, the usual t-test is valid asymptotically ($n \longrightarrow \infty$)
- The usual F-statistic does not work under heteroskedasticity, but heteroskedasticity robust versions are available in most software (including *R* and *STATA*).