Live Session - Week 4: CLM Inference

Devesh Tiwari and Jeffrey Yau

September 17, 2016

Agenda

1. Any questions before we begin (5 - 10 minutes)
2. Quick review of Gauss Markov assumptions and Heteroskedasticity (15 minutes)
3. An additional assumption (10 minutes)
4. Estimation, Diagnosis, Hypthesis Testing - 3 Breakouts (60 minutes)

1. Any questions before we being (Total: 5 - 10 minutes)

2. Quick review of Gauss Markov Assumption and Heteroskedasticity (Breatkout Room: 10 minutes / Total: 15 Minutes)

In a *breakout room*, spend 10 minutes to 1. write down the Gauss Markov Assumption you've learned up to week 3

2. define the homoskedasticity assumption
3. discuss the consequence of its violation on
    - (1) the coefficient estimators
    - (2) the estimated variance of the coefficient estimators, and
    - (3) consequence on statistical inference (e.g. test of significance of coefficient)
4. discuss how to diagnose it using regression diagnostic

3. An Additional Assumption Needed for Hypothesis Testing

(Breakout Room: 5 minutes / Total: 10 Minutes)

An additional assumption: The population error u is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with mean 0 and variance σ^2

- This assumption implies MLR.4 and MLR.5 (in Wooldridge's text). Why?
- A succinct way to summarize the assumptions of CLM is

$$y|\mathbf{X} \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$$

- Why is this a succince way to summarize the assumptions of CLM?

Normal Sampling Distribution

$$\hat{\beta}_j \sim N(\beta_j, var(\hat{\beta}_j))$$

$$(\hat{\beta}_j - \beta_j)/sd(\hat{\beta}_j) \sim N(0,1)$$

4. Hypothesis Testing

Introduction

In this exercise, you will be examining some output of someone else's (say, your colleagues') data analysis output. This particular analysis examined *US county level data* with information about counties' average life expectancy, obesity rate, and the proportion of the population who are physically active. These data are obtained at the *Institute for Health Metrics and Evaluation at the University of Washington.* This data have been slightly modified for this exercise.

This analysis examines the question, "Do people who are classified as being obese have shorter life spans than people who are not obese?"

Breakout Session 1 - Quick Review of OLS Estimation and Interpretation (Breakout Room: 10 minutes, Total: 15 minutes)

0. Assuming that the data wrangling and cleaning part as well as the EDA are already done. In practice, these parts can be extremely time-consuming.

1. To what extent can these data answer the research question as it is currently posed? Based on the results below, how would you answer this question?
2. Run a bivariate regresssion with $lifeexpectall$ as the dependent variable and *obesityall* as the independent variable. Interpret β_0 and β_1 in Model 1. Do you think the result of $\beta 1$ is substantively important? Why or why not? (Let's not dwell too long in this question)
3. Add the variable, *physicalactivityall* to your regression. Note that there is an increase in this model's R^2. Does this mean that Model 2 is a better model than Model 1? Does the inclusion of this second variable increase the overall explanatory power of the model?
4. Why did the coefficient for obseity decrease? What does this tell you about the relationship between physical activity and life expectancy? Physical activity and obesity?
5. The standard error for β_1 is higher in Model 2 and it is in Model 1. Do you think that the inclusion of additional covariates always increases the standard error of your coefficient estimates?

Breakout Session 2

(Breakout Room: 10 minutes, Total: 15 minutes)

1. Conduct post-regression diagnositcs and discuss them with respect to each of the five CLM assumptions. Based on these findings, propose and if possible, implement changes to your model. Please be prepared to discuss what change you proposed, why you proposed it, and whether you were able to implement it.

Breakout Session 3

(Breakout Room: 15 minutes, Total: 25 minutes)

These data also contain county level data separated by gender. In other words, we can examine the relationship between life expectancy, obesity, and physical activity for men and women separately.

1. Run 2 regressions, one for men and one or women. Interpret the relationship between obesity and life expectancy for men and women separetly.
2. You may noteice that the estimated *effect of physical activity* on life expectancy seem to be different between men and women. Using this two regressions, test the hypothesis that the effect of physical activity on life expectancy is same women as it is men, assuming (for this exercise) that the estimated effect of physical activity, $\hat{\beta}_2$, for men is the true effect.
3. Is #2 the correct way to test the (linear restriction) hypothesis that the effect of physical activity on men's life expectancy is same as that for women? Please explain and make necessary change test this hypothesis. Discuss your result.