POLS 6481. Research Design and Quantitative Methods II
Lecture 11 Leftover Slide on the Selection of Regressors
Readings: Wooldridge, *Introductory Econometrics 5e*, 4.5

–2. Reviewing the *F Statistic*

Unrestricted linear model with *k* regressors $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \cdots + \widehat{\beta_k} x_{ki}$

Restricted linear model with *k* – *q* regressors $\widehat{y_i}' = \widehat{\beta_0}' + \widehat{\beta_1}' x_{1i} + \cdots + \widehat{\beta_{(k-q)}}' x_{(k-q)i}$

Null hypothesis that *q* regressors are jointly insignificant $H_0 : \beta_{k-q+1} = \ldots = \beta_k = 0$

The *most* restricted model is the "***null model***" which has 0 explanatory variables;
it restricts all $\widehat{\beta_j} = 0$, for $j = 1,\ldots, k$; therefore (1) $\widehat{y_i} = \bar{y}$, (2) $q = k$, (3) $df_r = n-1$, and (4) $R_r^2 = 0$

Six steps

1. Regress y on $x_1, x_2, \ldots, x_{k-q}$ and save R_r^2 and degrees of freedom: $df_r = n - (k - q) - 1$
2. Regress y on $x_1, x_2, \ldots, x_k$ and save R_u^2 and degrees of freedom: $df_u = n - k - 1$
3. Calculate difference in R-squared $R_u^2 - R_r^2$
4. Calculate difference in degrees of freedom $q = df_u - df_r$
5. Calculate test statistic: $F = \dfrac{\left(R_u^2 - R_r^2\right)/q}{\left(1 - R_u^2\right)/(n-k-1)}$
6. Compare to critical value (F random variable with q , df_u degrees of freedom);
   to find critical value in R, type: *qf*(.05, #q, #n–k–1, *lower.tail=FALSE*)
   reject null hypothesis of joint *in*significance if $F > F^*$

POLS 6481. Research Design and Quantitative Methods II

Lecture 11 Leftover Slide on the Selection of Regressors

Readings: Wooldridge, *Introductory Econometrics 5e*, 7.3

–1. Introducing the *LM Statistic* (AKA the "*score statistic*", the "*n-R-squared statistic*")

Unrestricted linear model with *k* regressors $\hat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \cdots + \widehat{\beta_k} x_{ki}$

Restricted linear model with *k* – *q* regressors $\widehat{y_i}' = \widehat{\beta_0}' + \widehat{\beta_1}' x_{1i} + \cdots + \widehat{\beta_{(k-q)}}' x_{(k-q)i}$

Null hypothesis that *q* regressors are jointly insignificant $H_0 : \beta_{k-q+1} = \ldots = \beta_k = 0$

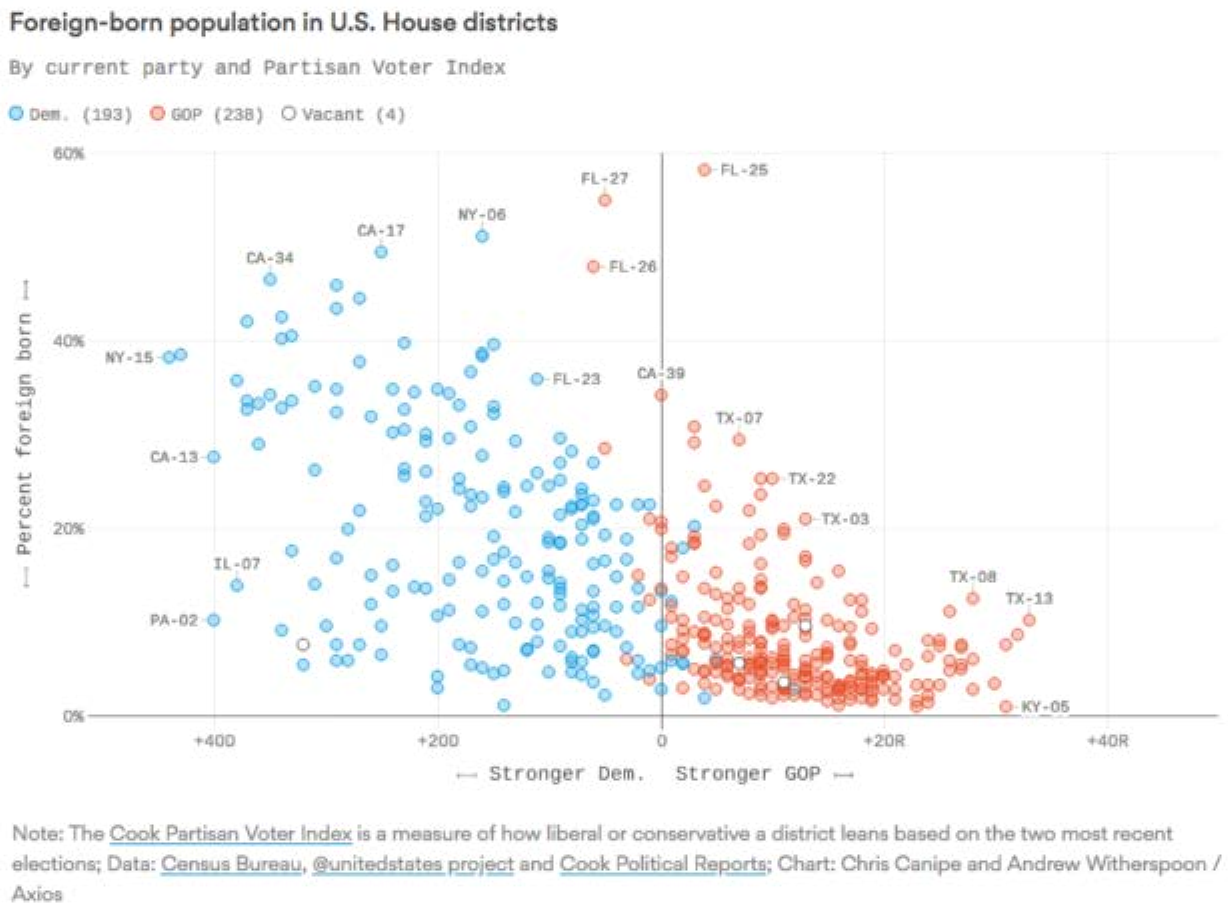
Five steps

1. Regress *y* on x_1, x_2, … , x_{k-q}
2. Generate predicted values $\widehat{y_i}'$ and residuals $\widehat{u_i}'$
3. Regress residuals $\widehat{u}'$ on x_{k-g+1}, … , x_k ; save the R^2 from this model, denoted $R^{2\prime}$
4. Compute $LM = n \cdot R^{2\prime}$
5. Compare to critical value (χ^2 random variable with *q* degrees of freedom);
   to find critical value in R, type: *qchisq*(*p*=.05, *df*=*#q*, *lower.tail=FALSE*)
   reject null hypothesis of joint insignificance if *LM* > critical value

POLS 6481. Research Design and Quantitative Methods II
Lecture 11. The Causes and Consequences of Heteroskedasticity

Readings: Wooldridge, *Introductory Econometrics 5e,* 8.1 + 8.3

0. An Illustration of Heteroskedasticity



Foreign-born population in U.S. House districts

By current party and Partisan Voter Index

Note: The Cook Partisan Voter Index is a measure of how liberal or conservative a district leans based on the two most recent elections; Data: Census Bureau, @unitedstates project and Cook Political Reports; Chart: Chris Canipe and Andrew Witherspoon / Axios

POLS 6481. Research Design and Quantitative Methods II

Lecture 11. The Causes and Consequences of Heteroskedasticity

Readings: Wooldridge, *Introductory Econometrics 5e,* 8.1 + 8.3

Outline

1. The Consequences of Heteroskedasticity
2. Diagnosing Heteroskedasticity
3. Example: *Education and Wages*

POLS 6481. Research Design and Quantitative Methods II
Lecture 11. The Causes and Consequences of Heteroskedasticity
Readings: Wooldridge, *Introductory Econometrics 5e,* 8.1 + 8.3

1. The Consequences of Heteroskedasticity

Recall from lecture 10, we use only MLR.1 – MLR.4 to ensure that the $\widehat{\beta}_j$ are unbiased; i.e.,

Theorem 3.1. $E(\widehat{\beta}_j) = \beta_j$ for $j = 1, \ldots, k$,

MLR.5 (homoskedasticity) only is used to ensure that $\hat{\sigma}^2$ is unbiased; i.e.,

Theorem 3.3. $E(\hat{\sigma}^2) = \sigma^2$

The definition of homoskedasticity is: $Var(u_i | x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2$

If this is violated, so the spread of the disturbances varies across values of *x*, then:

→ the estimate of $\hat{\sigma}^2$ is biased
→ the estimate of every $var(\widehat{\beta}_j)$ is biased
→ the standard error of every $\widehat{\beta}_j$ is biased
→ every *t* statistic is invalid
→ every *F* and/or *LM* statistic is invalid

Wooldridge claims that R^2 and adjusted R^2 are unaffected by heteroskedasticity (p. 268–9)

POLS 6481. Research Design and Quantitative Methods II
Lecture 11. The Causes and Consequences of Heteroskedasticity
Readings: Wooldridge, *Introductory Econometrics 5e,* 8.1 + 8.3

2. Diagnosing Heteroskedasticity

A. Theoretical

Think about potential sources of heteroskedasticity (context dependent)…

Individuals with higher incomes typically have greater discretion over spending, consequently we observe greater variation in retirement savings at higher income levels, and/or we might observe greater variation in cigarette smoking at higher income levels

Social statistics calculated from smaller samples typically have greater variability, consequently we observe greater deviations from the regression line in smaller states (because a few events can have a large impact on a proportion)

Countries with more *veto players* typically face more transaction costs in passing legislation, consequently we might greater less change over time in countries with fewer checks

POLS 6481. Research Design and Quantitative Methods II
Lecture 11. The Causes and Consequences of Heteroskedasticity
Readings: Wooldridge, *Introductory Econometrics 5e,* 8.1 + 8.3

2. Diagnosing Heteroskedasticity

B. Visual

plot dependent variable (y) against each predictor, with regression line included

plot residuals ($\hat{u}$) against each predictor, with horizontal line at 0

plot residuals ($\hat{u}$) against fitted values ($\hat{y}$), with horizontal line at 0

POLS 6481. Research Design and Quantitative Methods II
Lecture 11. The Causes and Consequences of Heteroskedasticity
Readings: Wooldridge, *Introductory Econometrics 5e,* 8.1 + 8.3

2. Diagnosing Heteroskedasticity

C. Breusch-Pagan Test

Theoretically, the null hypothesis is: $H_0 : var(u \mid \boldsymbol{x}) = E(u^2 \mid \boldsymbol{x}) = E(u^2) = \sigma^2$
where $\boldsymbol{x} = (x_1, x_2, \dots, x_k)$.

Practically, we might assume a linear functional form: $u^2 = \delta_0 + \delta_1 x_1 + \dots + \delta_k x_k + v$
where v is a disturbance orthogonal to $\boldsymbol{x}$

In this case, our null hypothesis becomes: $H_0 : \delta_1 = \dots = \delta_k = 0$

Five stop process:

1. regress y on $x_1, x_2, \dots, x_k$
2. generate fitted values ($\hat{y}$) and residuals ($\hat{u}$)
3. regress the squared residuals ($\hat{u}^2$) on $x_1, x_2, \dots, x_k$
4. save the R^2 from this regression (called ${R_{\hat{u}^2}}^2$)
5. perform the F test: $F = \dfrac{{R_{\hat{u}^2}}^2 / k}{\left(1 - {R_{\hat{u}^2}}^2\right) / n-k-1}$ (k, n–k–1 *d.f.*) or an *LM* test: $\chi^2 = n \cdot {R_{\hat{u}^2}}^2$ (k d.f.)

POLS 6481. Research Design and Quantitative Methods II

Lecture 11. The Causes and Consequences of Heteroskedasticity

Readings: Wooldridge, *Introductory Econometrics 5e,* 8.1 + 8.3

2. Diagnosing Heteroskedasticity

D. White Test

Alternatively, we might assume a quadratic and interactive function for the errors:

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 {x_1}^2 + \delta_5 {x_2}^2 + \delta_6 {x_3}^2 + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + v$$

where v is a disturbance orthogonal to $\boldsymbol{x}$ and all cross-products…

In this case, our null hypothesis becomes: $H_0 : \delta_1 = \ldots = \delta_9 = 0$

1. regress y on x_1, x_2, x_3
2. generate fitted values ($\hat{y}$) and residuals ($\hat{u}$)
3. regress the <u>squared</u> residuals ($\hat{u}^2$) on $x_1,\ x_2,\ x_3,\ {x_1}^2,\ {x_2}^2,\ {x_3}^3,\ x_1x_2,\ x_1x_3,\ x_2x_3$
4. save the R^2 from this regression (called${R_{\hat{u}^2}}^2$)
5. perform the F test: $F = \dfrac{{R_{\hat{u}^2}}^2 \big/ k}{\left(1-{R_{\hat{u}^2}}^2\right) \big/ n-k-1}$ (k, n–k–1 *d.f.*) or an *LM* test: $\chi^2 = n \cdot {R_{\hat{u}^2}}^2$ (k d.f.)

Problem: abundance of regressors in 3. With k = 3, there are 9 terms.

If k = 4, then there would be 14 terms.

POLS 6481. Research Design and Quantitative Methods II
Lecture 11. The Causes and Consequences of Heteroskedasticity
Readings: Wooldridge, *Introductory Econometrics 5e,* 8.1 + 8.3

2. Diagnosing Heteroskedasticity

D. White Test – alternative version

Substitute: $u^2 = \delta_0 + \delta_1\hat{y} + \delta_1\hat{y}^{\,2} + v$
for: $u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + v$

In this case, our null hypothesis becomes: $\mathrm{H}_0 : \delta_1 = \delta_2 = 0$

1. regress y on $x_1, x_2, \ldots, x_k$ for any positive integer k
2. generate fitted values ($\hat{y}$) and residuals ($\hat{u}$)
3. regress the squared residuals ($\hat{u}^2$) on $\hat{y}$ and $\hat{y}^2$
4. save the R^2 from this regression (called$R_{\hat{u}^2}{}^2$)
5. perform the F test: $F = \dfrac{R_{\hat{u}^2}{}^2 / k}{\left(1 - R_{\hat{u}^2}{}^2\right) / {n-k-1}}$ (2, n–3 *d.f.*) or an LM test: $\chi^2 = n \cdot R_{\hat{u}^2}{}^2$ (*2* d.f.)

POLS 6481. Research Design and Quantitative Methods II
Lecture 11. The Causes and Consequences of Heteroskedasticity
Readings: Wooldridge, *Introductory Econometrics 5e,* 8.1 + 8.3

3. Example: *Education and Wages*

R script is Lecture 11 cps08.R
Dataset is cps08.dta

Sub-sample of 2,989 respondents from March 2008 Current Population Study (conducted by U.S. Department of Labor); only respondents aged 29 and 30 are included in sub-sample

Simple regression predicts average hourly earnings from years of education (with or without controls for gender)

- Examine scatterplot with regression line
- Examine scatterplot of residuals against fitted values
- Estimate baseline model, which shows that each additional year of education increases wages by $1.76 per hour
- Carry out Breusch-Pagan test (and White test?)
- Compare ordinary standard errors to robust standard errors; compare *t* statistics