POLS 6481. Research Design and Quantitative Methods II

Lecture 10. Model Fit; Hypothesis Tests for Models

Readings: Wooldridge, *Introductory Econometrics 5e*, 2.3 + 4.5 + 6.3

Outline

1. Sums of Squared Errors
2. $R^2$ and Adjusted $R^2$
3. ANOVA for Group Means
4. The $F$ Distribution
5. ANOVA for Regression Models
6. ANOVA for Regression Specifications

POLS 6481. Research Design and Quantitative Methods II
Lecture 10. Model Fit; Hypothesis Tests for Models
Readings: Wooldridge, *Introductory Econometrics 5e*, 2.3

1. Sums of Squared Errors

| | | |
|---|---|---|
| Total Sum of 'Squares' | $\mathrm{SST} = \sum(y_i - \bar{y})^2$ | [2.33] |
| Residual Sum of 'Squares' | $\mathrm{SSR} = \sum(y_i - \hat{y}_i)^2 = \sum \hat{u}_i{}^2$ | [2.35] |
| Explained Sum of 'Squares' | $\mathrm{SSE} = \sum(\hat{y}_i - \bar{y})^2$ | [2.34] |

| | | |
|---|---|---|
| Relationship: | SST = SSR + SSE | [2.36] |

2. $R^2$ and Adjusted $R^2$

'Coefficient of Determination'

$$R^2 = 1 - \frac{SSR}{SST} \quad [2.38]$$

$$= 1 - \frac{SSR/_{n-1}}{SST/_{n-1}}$$

$$= 1 - \frac{var(\hat{u})}{var(y)}$$

Adjusted $R^2$

$$Adj.\ R^2 = 1 - \frac{\hat{\sigma}^2}{var(y)}$$

$$= 1 - \frac{SSR/_{n-k-1}}{SST/_{n-1}} \quad [6.21]$$

$$= 1 - \frac{SSR}{SST} \cdot \frac{n-1}{n-k-1}$$

$$= 1 - (1 - R^2) \cdot \left(\frac{n-1}{n-k-1}\right)$$

## 2. $R^2$ and Adjusted $R^2$

Adjusted $R^2$ imposes a "penalty" for adding explanatory variables to a model.

Interesting: " if we add a variable to a regression equation, then adjusted $R^2$ increases *if and only if* the $t$ statistic on the new variable is greater than 1 in absolute value. "

Wooldridge cautions: students typically put too much weight on $R^2$
$R^2$ are artificially high in time-series regressions
$R^2$ are artificially low when the dependent variable is under-dispersed

Readings: Wooldridge, *Introductory Econometrics 5e*, 2.3 + 6.3

2. $R^2$ and Adjusted $R^2$

**Wooldridge's advice on p. 205–207:**

Adding regressors reduces the error variance ...

... but adding regressors may exacerbate multicollinearity problems!

Variables that are uncorrelated with other regressors *should* be added because they reduce error variance without increasing multicollinearity ...

... but variables *might* be hard to identify (consider different levels of aggregation?)

3. ANOVA for Group Means

A. Compare $t$ test of difference-of-means to ANOVA $F$ test with two groups

$$t_{2-sample} = \frac{M_1 - M_2}{s_{M_1 - M_2}} \qquad s_{M_1 - M_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$t$ critical value found in $t$ table, with d.f. = $n_1 + n_2 - 2$

$$F = \frac{MSE_{between}}{MSE_{within}} \quad \text{where} \quad MSE_{between} = (n_1)(\bar{x}_1 - \bar{x})^2 + (n_2)(\bar{x}_2 - \bar{x})^2$$

$$\text{and} \quad MSE_{within} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

F "critical value" found in $F$ table; numerator d.f. = *number of groups* – 1 = 1
denominator d.f. = $n$ – *number of groups* = $n_1 + n_2 - 2$

Note that for two groups or two samples, $F = t^2$

3. ANOVA for Group Means

B. Expanding ANOVA to more than two groups/samples

For two groups:

$$MSE_{between} = (n_1)(\bar{x}_1 - \bar{x})^2 + (n_2)(\bar{x}_2 - \bar{x})^2$$

$$MSE_{within} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

"critical value" found in *F* table

$$F = \frac{MSE_{between}}{MSE_{within}}$$

numerator d.f. = *number of groups* – 1 = 1

denominator d.f. = *n* – *number of groups* = $n_1 + n_2 - 2$

3. ANOVA for Group Means

B. Expanding ANOVA to more than two groups/samples

For three groups: $$MSE_{between} = \frac{(n_1)(\bar{x}_1 - \bar{x})^2 + (n_2)(\bar{x}_2 - \bar{x})^2 + (n_3)(\bar{x}_3 - \bar{x})^2}{2}$$

$$MSE_{within} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{n_1 + n_2 + n_3 - 3}$$

"critical value" found in *F* table

$$F = \frac{MSE_{between}}{MSE_{within}}$$

numerator d.f. = *number of groups* – 1 = 2

denominator d.f. = *n* – *number of groups* = $n_1 + n_2 + n_3 - 3$

3. ANOVA for Group Means

B. Expanding ANOVA to more than two groups/samples

For four groups:

$$MSE_{between} = \frac{(n_1)(\bar{x}_1 - \bar{x})^2 + (n_2)(\bar{x}_2 - \bar{x})^2 + (n_3)(\bar{x}_3 - \bar{x})^2 + (n_4)(\bar{x}_4 - \bar{x})^2}{3}$$

$$MSE_{within} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + (n_4 - 1)s_4^2}{n_1 + n_2 + n_3 + n_4 - 4}$$

F "critical value" found in *F* table

$$F = \frac{MSE_{between}}{MSE_{within}}$$

numerator d.f. = *number of groups* – 1 = 3

denominator d.f. = *n* – *number of groups* = $n_1 + n_2 + n_3 + n_4 - 4$

3. ANOVA for Group Means

C. The Main Idea of ANOVA: Decomposition of Variance

Compute Sums of Squares via the 'Deviation Score Method'

1. Each observation deviates from the "grand mean," $M_{tot}$, by $x_i - M_{tot}$

$$SS_{tot} = \sum_{i=1}^{N} (x_i - M_{tot})^2$$ is the "total sum of squares"

2. Each group mean deviates from the grand mean by $M_g - M_{tot}$

$$SS_{bet} = \sum_{j=1}^{g} n_g (M_g - M_{tot})^2$$ is the "between groups sum of squares" where there are $n_g$ observations in group $g$

3. Each observation deviates from its "group mean," $M_g$, by $x_i - M_g$

$$SS_{with} = \sum_{j=1}^{g} \sum_{i=1}^{n_g} (x_i - M_g)^2$$ is the "within groups sum of squares"

3. ANOVA for Group Means

C. The Main Idea of ANOVA: Decomposition of Variance

From Sums of Squares to Mean Squares

1. The Between-Groups Mean-Squared Error is:

$MS_{bet} = SS_{bet} \div df_{bet}$, where $df_{bet}$ = number of groups – 1

2. The Within-Groups Mean Squared Error is:

$MS_{with} = SS_{with} \div df_{with}$, where $df_{with}$ = number of observations – number of groups

3. Total Mean Squared Error is not usually calculated… but it equals the variance of $x$:

$MS_{tot} = SS_{tot} \div df_{tot}$, where $df_{total}$ = number of observations – 1

4. The F-statistic is: $MS_{bet} \div MS_{with}$

3. ANOVA for Group Means

D. The ANOVA Table

| *Source* | *SS* | *df* | *MS* | *F* | *F-critical* |
|---|---|---|---|---|---|
| Between | ❶ | ❹ | ❼ | ❾ | ❿ |
| Within | ❷ | ❺ | ❽ | --- | |
| Total | ❸ | ❻ | --- | --- | |

❸ = ❶ + ❷

❻ = ❹ + ❺

❼ = ❶ ÷ ❹

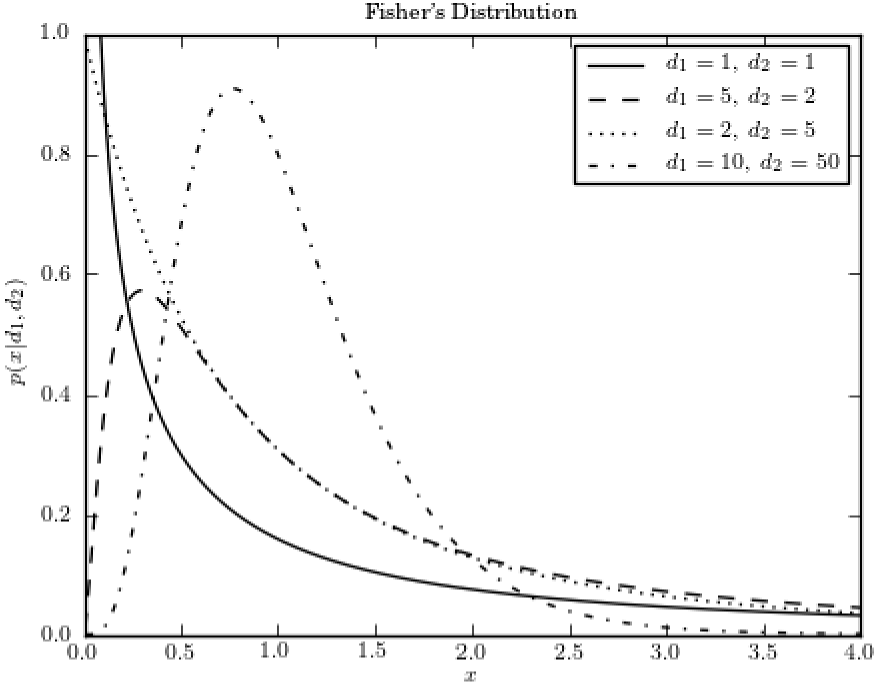
❽ = ❷ ÷ ❺

❾ = ❼ ÷ ❽

❿ has *d.f.* = ❹ , ❺

POLS 6481. Research Design and Quantitative Methods II
Lecture 10. Model Fit; Hypothesis Tests for Models
Readings: Wooldridge, *Introductory Econometrics 5e*, 4.5

## 4. The *F* Distribution




"We will not derive the *F* distribution because the mathematics is very involved. Basically it can be shown that the equation $F = \frac{(SST-SSR)/k}{SSR/(n-k-1)}$ is actually the ratio of two independent $\chi^2$ variables. The numerator $\chi^2$ random variable has *k* degrees of freedom, and the denominator $\chi^2$ random variable has *n*–*k*–1 degrees of freedom. This is the definition of an *F* distributed random variable."

## Table F The *F* Distribution

$\alpha = .05$

| $df_D$ \ $df_N$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 161.4 | 199.5 | 215.7 | 224.6 | 230.2 | 234.0 | 236.8 | 238.9 | 240.5 | 241.9 |
| **2** | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 | 19.40 |
| **3** | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 |
| **4** | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 |
| **5** | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 |
| **6** | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 |
| **7** | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 |
| **8** | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 |
| **9** | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 |
| **10** | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 |
| **11** | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 |
| **12** | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 |
| **13** | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 |
| **14** | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | 2.60 |
| **15** | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 |
| **16** | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | 2.49 |
| **17** | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 | 2.45 |
| **18** | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 |
| **19** | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 | 2.38 |
| **20** | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 | 2.35 |
| **21** | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 | 2.32 |
| **22** | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 | 2.30 |
| **23** | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 | 2.27 |
| **24** | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 | 2.25 |
| **25** | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 | 2.24 |
| **26** | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 | 2.22 |
| **27** | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.25 | 2.20 |
| **28** | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.24 | 2.19 |
| **29** | 4.18 | 3.33 | 2.93 | 2.70 | 2.55 | 2.43 | 2.35 | 2.28 | 2.22 | 2.18 |
| **30** | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 | 2.16 |
| **40** | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 | 2.08 |
| **60** | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 | 1.99 |
| **120** | 3.92 | 3.07 | 2.68 | 2.45 | 2.29 | 2.17 | 2.09 | 2.02 | 1.96 | 1.91 |
| **∞** | 3.84 | 3.00 | 2.60 | 2.37 | 2.21 | 2.10 | 2.01 | 1.94 | 1.88 | 1.83 |

POLS 6481. Research Design and Quantitative Methods II
Lecture 10. Model Fit; Hypothesis Tests for Models
Readings: Wooldridge, *Introductory Econometrics 5e*, 4.5

5. ANOVA for Regression Models

$F = \frac{(SST-SSR)/k}{SSR/(n-k-1)}$ where $SST = \sum(y_i - \bar{y})^2$ & $SSR = \sum(y_i - \hat{y}_i)^2 = \sum \hat{u}_i^{\,2}$

Recall: $R^2 = 1 - \frac{SSR}{SST}$

$$1 - R^2 = \frac{SSR}{SST}$$

$$\therefore (1 - R^2)\cdot SST = SSR$$

We can make the following substitutions: $\frac{(SST-SSR)/k}{SSR/(n-k-1)} = \frac{(SST-(1-R^2)SST)}{(1-R^2)SST}\cdot\frac{n-k-1}{k}$

$$= \frac{\left(1-(1-R^2)\right)SST}{(1-R^2)SST}\cdot\frac{n-k-1}{k}$$

$$= \frac{R^2}{1-R^2}\cdot\frac{n-k-1}{k}$$

$$= \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

5. ANOVA for Regression Models

$$F = \frac{(SST - SSR)/k}{SSR/(n-k-1)}$$

In R, type *anova*(*nullmodel*, *model*)

Compare $F$ to the $F$ critical value with $k$, $n$–$k$–1 degrees of freedom

In R, type: *qf*(.05, #*k*, #*n*–*k*–1, *lower.tail=F*)

6. ANOVA for Regression Specifications

Unrestricted model: $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \cdots + \widehat{\beta_k} x_{ki}$

$$= \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \cdots + \widehat{\beta_{(k-q)}} x_{(k-q)i} + \widehat{\beta_{(k-q+1)}} x_{(k-q+1)i} + \cdots + \widehat{\beta_k} x_{ki}$$

Unrestricted model Residual Sum of Squares: $\mathrm{SSR}_u = \sum (y_i - \widehat{y_i})^2 = \sum \widehat{u_i}^2$

Restricted model: $\widehat{y_i}' = \widehat{\beta_0}' + \widehat{\beta_1}' x_{1i} + \cdots + \widehat{\beta_{(k-q)}}' x_{(k-q)i}$

Restricted model Residual Sum of Squares: $\mathrm{SSR}_r = \sum \left(y_i - \widehat{y_i}'\right)^2 = \sum \widehat{u_i}^2$

Let $q \leq k$ denote the number of restrictions: $q = df_r - df_u$
$= (n-(k-q)-1) - (n-k-1)$

Then our test statistic for those restrictions is: $F = \dfrac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)}$ (note that $\mathrm{SSR}_u \leq \mathrm{SSR}_r$ )

What do we mean by restrictions? A *joint* null hypothesis, such as $H_0: \beta_{k-q+1} = \ldots = \beta_k = 0$

(Another name that you might encounter is “exclusion restrictions,” meaning that we might ‘safely’ exclude these variables from our regression model if the null hypothesis is retained.)

## 6. ANOVA for Regression Specifications (The $R^2$ Version)

Recall again that $R^2 = 1 - \frac{SSR}{SST}$, which implies $\frac{SSR}{SST} = 1 - R^2$; $\therefore$ $SSR = (1 - R^2) \cdot SST$

By extension, $SSR_r = (1 - R_r^2) \cdot SST$ and $SSR_u = (1 - R_u^2) \cdot SST$

We can make the following substitutions:

$$\frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} = \frac{(SSR_r - SSR_u)}{SSR_u} \cdot \frac{n-k-1}{q}$$

$$= \frac{SST \cdot (1 - R_r^2) - SST \cdot (1 - R_u^2)}{SST \cdot (1 - R_u^2)} \cdot \frac{n-k-1}{q}$$

$$= \frac{SST \cdot \left((1 - R_r^2) - (1 - R_u^2)\right)}{SST \cdot (1 - R_u^2)} \cdot \frac{n-k-1}{q}$$

$$= \frac{R_u^2 - R_r^2}{1 - R_u^2} \cdot \frac{n-k-1}{q}$$

$$= \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2)/(n-k-1)}$$

POLS 6481. Research Design and Quantitative Methods II
Lecture 10. Model Fit; Hypothesis Tests for Models
Readings: Wooldridge, *Introductory Econometrics 5e*, 4.5

6. ANOVA for Regression Specifications

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} = \frac{\left(R_u{}^2 - R_r{}^2\right)/q}{\left(1 - R_u{}^2\right)/(n-k-1)}$$

In R, type *anova*(*restricted model*, *unrestricted model*)

Compare *F* to the *F* critical value with *q*, *n*–*k*–1 degrees of freedom

In R, type: *qf*(.05, #*q*, #*n*–*k*–1, *lower.tail=F*)