Stats 401 Lab 9

401 GSI team

3/15/2018 and 3/16/2018

Outline

- Quick Reminder: If you are thinking about withdrawing from the course, the deadline is Monday! (**March 19th**)
- Quiz Information
- Review of F tests and ANOVA
- Lab Activity
- Lab Ticket

Quiz Information

- Next quiz will be on **March 29th** and **March 30th**
- It will cover:
  - hardest/most missed questions on the midterm
  - new material since the midterm

F Test/ANOVA Motivation

Why perform the F test? - Want to know if additional variables are statistically significant.

How is this different from looking at the regression output? - Regression output is testing each b_i with all other b_j fixed - F test/ANOVA lets us test multiple variables for significance at once

F Test Review

Recall from lecture:

- $H_0 : Y = \mathbb{X}\beta + \epsilon$
- $H_a : Y = \mathbb{X}_a\beta_a + \epsilon$

$$f = \frac{(RSS_0 - RSS_a)/d}{RSS_a/(n-q)}$$

- RSS_0 and RSS_a are the residual sum of squares for the null and alternative models
- d is the difference in the degrees of freedom between the two models
- $n - q$ is the degrees of freedom in the alternative model

Examining the F Test in R

We will be examining a subset of the National Education Longitudinal Study of 1988 which examined schoolchildren's performance on a math test score in 8th grade. "ses" is the socioeconomic status of parents and "paredu" is the parents highest level of education achieved (less than high school, high school, college, BA, MA, PhD).

```
library(faraway)
data(nels88)
head(nels88)
```

```
##      sex  race   ses paredu math
## 1 Female White -0.13     hs   48
## 2   Male White -0.39     hs   48
## 3   Male White -0.80     hs   53
## 4   Male White -0.72     hs   42
## 5 Female White -0.74     hs   43
## 6 Female White -0.58     hs   57
```

Examining the F Test in R

Suppose we want to know whether the parents' level of education and socioeconomic status affect a student's performance.

Write out the linear model (sample version) in matrix form and subscript form.

Examining the F Test in R

```
ses_edu_lm <- lm(math ~ ses + paredu, data = nels88)
summary(ses_edu_lm) # run this in R to show full output
```

```
##
## Call:
## lm(formula = math ~ ses + paredu, data = nels88)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.1894  -5.9894  -0.2677   6.0463  24.0086
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     58.5712     1.7813  32.882  < 2e-16 ***
## ses              2.7913     1.3096   2.132 0.034012 *
## pareducollege   -7.5210     2.1414  -3.512 0.000526 ***
## pareduhs       -12.1792     2.6420  -4.610  6.4e-06 ***
## paredulesshs   -13.3645     3.3328  -4.010  8.0e-05 ***
## pareduma        -0.8709     2.2455  -0.388 0.698449
## pareduphd       -2.0494     2.5202  -0.813 0.416885
```

Examining the F Test in R (Calc by Hand)

First, we need to find our null hypothesis:

```
ses_lm <- lm(math ~ ses, data = nels88)
summary(ses_lm) # run this in R to show full output
```

```
##
## Call:
## lm(formula = math ~ ses, data = nels88)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.7491  -6.0454  -0.4198   6.3113  22.7178
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.8225     0.5438   95.29   <2e-16 ***
## ses           7.1276     0.5599   12.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 8.744 on 258 degrees of freedom
```

Calculate the F Statistic by Hand

```
rss_0 <- sum(residuals(ses_lm)^2); rss_0
```

```
## [1] 19725.18
```

```
rss_a <- sum(residuals(ses_edu_lm)^2); rss_a
```

```
## [1] 18082.73
```

$$f = \frac{(RSS_0 - RSS_a)/d}{RSS_a/(n-q)}$$

$$f = \frac{(19725.18 - 18082.73)/(258 - 253)}{18082.73/253} \tag{1}$$

$$f = \frac{(1642.45)/5}{18082.73/253}$$

$$f = 4.595986$$

Check using R

First find the p-value from the F that we calculated before. Then check using ANOVA.

```
pf(4.595986, 5, 253, lower.tail = F)
```

```
## [1] 0.0004957872
```

```
anova(ses_edu_lm)
```

```
## Analysis of Variance Table
##
## Response: math
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## ses          1 12391.4 12391.4 173.371 < 2.2e-16 ***
## paredu       5  1642.4   328.5   4.596 0.0004958 ***
## Residuals 253 18082.7    71.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Lab Activity

Q1) Write out the H_0 and H_a for the F test shown in the summary info of the linear regression.

Q2) We just saw that level of parents' education does infact affect students' test performance. Suppose we now want to know whether the sex of the student affects their performance along with their parents' level of education and socioeconomic status. That is, H_0 is the probability model where a students performance depends on parents' level of education and socioeconomic status and H_a is the model that includes sex. -(2a) Write out H_0 and H_a for this test and fit the models using lm(). (Clearly describe all the paameters in your formulae) -(2b) Find the F-statistic using the formula ($f = \frac{(RSS_0 - RSS_a)/d}{RSS_a/(n-q)}$) and the corresponding p-value using pf(). Report your conclusion. -(2c) Confirm your result by finding the p-value using anova().

Lab Ticket

Use pt() and pf() to show that if T has a t-distribution with n degrees of freedom, then T^2 follows the F-distribution with 1 and n degrees of freedom. Hint: Start with a specific value for n, say $n = 5$ and later exted to multiple values using a for-loop.