Multiple Linear Regression: Categorical Predictors

Author: Nicholas G Reich, Jeff Goldsmith

*This material is part of the **statsTeachR** project*

Today's Lecture

- Global F tests: review and examples

Addressing multiple comparisons

You should be concerned about Family-Wise Error Rates!

Three general approaches

- Do nothing in a reasonable way
  - Don't trust scientifically implausible results
  - Don't over-emphasize isolated findings
- Correct for multiple comparisons
  - Often, use the Bonferroni correction and use $\alpha_i = \alpha/k$ for each test
  - Thanks to the Bonferroni inequality, this gives an overall $FWER \leq \alpha$
- Use a global test

Global tests: an overview/review

Compare a smaller "null" model to a larger "alternative" model

- Smaller model must be nested in the larger model
- That is, the smaller model must be a special case of the larger model
- For both models, the *RSS* gives a general idea about how well the model is fitting
- In particular, something like

  $$\frac{RSS_S - RSS_L}{RSS_L}$$

  compares the relative *RSS* of the models

Global F tests: a common categorical example

$$\text{“Null” Model: } dis_i = \beta_0 + \beta_1 nut_i$$

$$\text{“Null” Model + Educ: } dis_i = \beta_0 + \beta_1 nut_i + \beta_2 educ_{6,i} + \cdots + \beta_{15} educ_{14,i}$$

```
mlrNull <- lm(disease ~ nutrition, data = dat)
mlr1 <- lm(disease ~ nutrition + factor(education), data = dat)
summary(mlr1)$coef

##                      Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          34.66557    4.82285  7.1878 2.042e-10
## nutrition            -0.04542    0.01829 -2.4836 1.490e-02
## factor(education)6  -0.91672    7.55158 -0.1214 9.037e-01
## factor(education)7  18.52195    5.86892  3.1559 2.191e-03
## factor(education)8  13.01127    5.23270  2.4865 1.479e-02
## factor(education)9  16.90911    5.23535  3.2298 1.742e-03
## factor(education)10 22.07698    5.08983  4.3375 3.828e-05
## factor(education)11 21.89305    5.26040  4.1619 7.332e-05
## factor(education)12 24.86794    5.55041  4.4804 2.231e-05
## factor(education)13 19.72658    6.76774  2.9148 4.513e-03
## factor(education)14 20.74128    9.57768  2.1656 3.305e-02
```

Global F tests: a common categorical example

$$\text{“Null” Model: } dis_i = \beta_0 + \beta_1 nut_i$$

$$\text{“Null” Model + Educ: } dis_i = \beta_0 + \beta_1 nut_i + \beta_2 educ_{6,i} + \cdots + \beta_{15} educ_{14,i}$$

```
mlrNull <- lm(disease ~ nutrition, data = dat)
mlr1 <- lm(disease ~ nutrition + factor(education), data = dat)
anova(mlrNull, mlr1)

## Analysis of Variance Table
##
## Model 1: disease ~ nutrition
## Model 2: disease ~ nutrition + factor(education)
##   Res.Df  RSS Df Sum of Sq    F  Pr(>F)
## 1     97 9193
## 2     88 6022  9      3171 5.15 1.3e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Global F tests

A couple of important special cases for the F test

- The null model contains the intercept only
  - When people say ANOVA, this is often what they mean (although all F tests are based on an analysis of variance)
- The null model and the alternative model differ only by one term
  - Gives a way of testing for a single coefficient
  - Turns out to be equivalent to a two-sided t-test: $t^2_{df_L} \sim F_{1,df_L}$

Lung data: single coefficient test

The F test is equivalent to the t test when there's only one parameter of interest

```
mlrNull <- lm(disease ~ nutrition, data = dat)
mlr2 <- lm(disease ~ nutrition + airqual, data = dat)
anova(mlrNull, mlr2)
```

```
## Analysis of Variance Table
##
## Model 1: disease ~ nutrition
## Model 2: disease ~ nutrition + airqual
##   Res.Df  RSS Df Sum of Sq    F  Pr(>F)
## 1     97 9193
## 2     96 5970  1      3223 51.8 1.3e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mlr2)$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  37.6254    2.43946   15.42 9.946e-28
## nutrition    -0.0347    0.01692   -2.05 4.307e-02
## airqual       0.3611    0.05016    7.20 1.347e-10
```

Today's Big Ideas

F tests can control for multiple comparisons!

- hands-on example

Today's Big Ideas

- Global tests: examples and special circumstances