Model Comparisons

Data Analysis for Psychology in R 2

dapR2 Team

Department of Psychology
The University of Edinburgh

Week's Learning Objectives

- Understand how to use model comparisons to test questions about model variables.
- Understand the calculation of the incremental F-test
- Understand the difference between nested and non-nested models, and the appropriate statistics to use for comparison in each case.

Topics for today

- Discuss some motivating examples:
    - Categorical variables with 2+ levels
    - Interactions with categorical variables with 2+ levels
    - Controlling for covariates
- Statistical tools for selection/comparison
    - Incremental F-test
    - Nested vs. non-nested models
    - AIC & BIC

Some data

- We have previously looked at this example.
- A researcher was interested in whether the subjective well-being of patients differed dependent on the post-operation treatment schedule they were given, and the hospital in which they were staying.
- **Condition 1**: `Treatment` (Levels: TreatA, TreatB, TreatC).
- **Condition 2**: `Hospital` (Levels: Hosp1, Hosp2).
- Total sample n = 180 (30 patients in each of 6 groups).
    - Between person design.
- **Outcome**: Subjective well-being (`SWB`)
    - An average of multiple raters (the patient, a member of their family, and a friend).
    - SWB score ranged from 0 to 20.

The data

```r
hosp_tbl <- read_csv("hospital.csv", col_types = "dff")
hosp_tbl %>%
  slice(1:10)
```

```
## # A tibble: 10 x 3
##       SWB Treatment Hospital
##     <dbl> <fct>     <fct>
##  1   6.2 TreatA    Hosp1
##  2  15.9 TreatA    Hosp1
##  3   7.2 TreatA    Hosp1
##  4  11.3 TreatA    Hosp1
##  5  11.2 TreatA    Hosp1
##  6   9   TreatA    Hosp1
##  7  14.5 TreatA    Hosp1
##  8   7.3 TreatA    Hosp1
##  9  13.7 TreatA    Hosp1
## 10  12.6 TreatA    Hosp1
```

Example 1: Categorical Variables with 2+ levels

- What if the researcher wanted to ask a general question; Is there an overall effect of treatement?

- How might we do this with the skills we have learned already?

```
summary(lm(SWB ~ Treatment, data = hosp_tbl))
```

Results

```
##
## Call:
## lm(formula = SWB ~ Treatment, data = hosp_tbl)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.373 -1.987 -0.300  1.838  7.173
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.3267     0.3242  28.770  < 2e-16 ***
## TreatmentTreatB   1.9467     0.4585   4.246 3.51e-05 ***
## TreatmentTreatC  -0.2850     0.4585  -0.622    0.535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.511 on 177 degrees of freedom
## Multiple R-squared:  0.1369,    Adjusted R-squared:  0.1271
## F-statistic: 14.04 on 2 and 177 DF,  p-value: 2.196e-06
```

Example 2: Categorical Interactions with 2+ levels

- If we stay with the same example, what if we asked the question:

- Is there an interaction between hospital and treatement?

```
summary(lm(SWB ~ Treatment*Hospital, data = hosp_tbl))
```

Results

```
##
## Call:
## lm(formula = SWB ~ Treatment * Hospital, data = hosp_tbl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6000 -1.2533  0.1083  1.2650  5.7000
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   10.8000     0.3699  29.195  < 2e-16 ***
## TreatmentTreatB               -1.3700     0.5232  -2.619   0.0096 **
## TreatmentTreatC               -0.6967     0.5232  -1.332   0.1847
## HospitalHosp2                 -2.9467     0.5232  -5.632 7.02e-08 ***
## TreatmentTreatB:HospitalHosp2   6.6333     0.7399   8.966 4.74e-16 ***
## TreatmentTreatC:HospitalHosp2   0.8233     0.7399   1.113   0.2673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.026 on 174 degrees of freedom
## Multiple R-squared:  0.4476,    Adjusted R-squared:  0.4317
## F-statistic:  28.2 on 5 and 174 DF,  p-value: < 2.2e-16
```

Some more data

- How about this example based on data from the Midlife In United States (MIDUS2) study.
- Outcome: self-rated health
- Covariates: Age, sex
- Predictors: Big Five traits and Purpose in Life.

The data

```
midus <- read_csv("MIDUS2.csv")
midus2 <- midus %>%
  select(1:4, 31:42) %>%
  mutate(
    PIL = rowMeans(.[grep("PIL", names(.))],na.rm=T)
  ) %>%
  select(1:4, 12:17) %>%
  drop_na(.)
slice(midus2, 1:3)
```

```
## # A tibble: 3 x 10
##      ID   age sex    health     O     C     E     A     N   PIL
##   <dbl> <dbl> <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 10002    69 MALE        8  2.14   2.8   2.6   3.4  2     5.86
## 2 10019    51 MALE        8  3.14   3     3.4   3.6  1.5   5.71
## 3 10023    78 FEMALE      4  3.57   3.4   3.6   4    1.75  5.14
```

Example 3: Controlling for covariates

- Suppose our question was....
- Does personality signficantly predict self-rated health over and above the effects of age and sex?

```
summary(lm(health ~ age + sex + O + C + E + A + N, data = midus2))
```

```
##
## Call:
## lm(formula = health ~ age + sex + O + C + E + A + N, data = midus2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7723 -0.7921  0.2532  1.0097  3.9550
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.66172    0.45100  14.771  < 2e-16 ***
## age         -0.01310    0.00298  -4.396 1.17e-05 ***
## sexMALE     -0.09571    0.07955  -1.203    0.229
## O            0.09308    0.08306   1.121    0.263
## C            0.57147    0.08507   6.717 2.49e-11 ***
## E            0.56771    0.08061   7.043 2.70e-12 ***
## A           -0.40380    0.09025  -4.474 8.15e-06 ***
## N           -0.56493    0.06189  -9.128  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.521 on 1753 degrees of freedom
## Multiple R-squared:  0.1484,    Adjusted R-squared:  0.145
## F-statistic: 43.65 on 7 and 1753 DF,  p-value: < 2.2e-16
```

What do these questions have in common?

- A key feature of these questions is that they require us to evaluate whether multiple variables (think more than one beta coefficient) are significant.

- Another - potentially more useful - way to think are "are significant" is to say "do they improve my model"?

- Up to this point, we have only discussed ways to explore this:

    - In a very limited case (i.e. the single categorical predictor with 2+ levels)
    - Descriptively

- What we will look at next is how we can formally test such questions.

Time for a break

Grab a cup of tea/coffee....a few equations on the way.

Welcome Back!

Recall the F-test

- F-ratio is a ratio of the explained to unexplained variance:

$$F = \frac{MS_{Model}}{MS_{Residual}}$$

- Where the mean squares (MS) are the sums of squares divided by the degrees of freedom. So we can also write:

$$F = \frac{SS_{Model}/df_{Model}}{SS_{Residual}/df_{residual}}$$

F-ratio

- Bigger F-ratios indicate better models.
  - It means the model variance is big compared to the residual variance.
- The null hypothesis for the model says that the best guess of any individuals y value is the mean of y plus error.
  - Or, that the x variables carry no information collectively about y.
  - Or, a test that all β = 0
- F-ratio will be close to 1 when the null hypothesis is true
  - If there is equivalent residual to model variation, F=1
  - If there is more model than residual F > 1
- F-ratio is then evaluated against an F-distribution with df_{Model} and $df_{Residual}$ and a pre-defined α
- Testing the F-ratio evaluates statistical significance of the overall model

F-test as an incremental test

- One important way we can think about the F-test and the F-ratio is as an incremental test against an "empty" or null model.

- A null or empty model is a linear model with only the intercept.

    - In this model, our predicted value of the outcome for every case in our data set, is the mean of the outcome.
    - That is, with no predictors, we have no information that may help us predict the outcome.
    - So we will be "least wrong" by guessing the mean of the outcome.

- An empty model is the same as saying all $\beta = 0$.

- So in this way, the F-test we have already seen **is comparing two models**.

- We can extend this idea, and use the F-test to compare two models that contain different sets of predictors.

    - This is the **incremental F-test**

Incremental F-test

- The incremental F-test evaluates the statistical significance of the improvement in variance explained in an outcome with the addition of further predictor(s)

- It is based on the difference in F-values between two models.

    - We call the model with the additional predictor(s) model 1 or full model
    - We call the model without model 0 or restricted model

$$F_{(df_R - df_F), df_F} = \frac{(SSR_R - SSR_F)/(df_R - df_F)}{SSR_F/df_F}$$

$$\begin{aligned}
&\text{Where:}\\
&SSR_R = \text{residual sums of squares for the restricted model}\\
&SSR_F = \text{residual sums of squares for the full model}\\
&df_R = \text{residual degrees of freedom from the restricted model}\\
&df_F = \text{residual degrees of freedom from the full model}
\end{aligned}$$

Time for a break

Welcome Back!

Let's look at some examples

Incremental F-test in R

- In order to apply the F-test for model comparison in R, we use the `anova()` function.
- `anova()` takes as its arguments models that we wish to compare
  - Here we will show examples with 2 models, but we can use more.

Application to example 1

- Is there an overall effect of treatement?

```
ex1_r <- lm(SWB ~ 1, data = hosp_tbl)
ex1_f <- lm(SWB ~ Treatment, data = hosp_tbl)

anova(ex1_r, ex1_f)
```

```
## Analysis of Variance Table
##
## Model 1: SWB ~ 1
## Model 2: SWB ~ Treatment
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    179 1293.1
## 2    177 1116.1  2    177.02 14.037 2.196e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Application to example 1

```
summary(ex1_f)
```

```
##
## Call:
## lm(formula = SWB ~ Treatment, data = hosp_tbl)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.373 -1.987 -0.300  1.838  7.173
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.3267     0.3242  28.770  < 2e-16 ***
## TreatmentTreatB    1.9467     0.4585   4.246 3.51e-05 ***
## TreatmentTreatC   -0.2850     0.4585  -0.622    0.535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.511 on 177 degrees of freedom
## Multiple R-squared:  0.1369,    Adjusted R-squared:  0.1271
## F-statistic: 14.04 on 2 and 177 DF,  p-value: 2.196e-06
```

Application to example 2

- Is there an interaction between hospital and treatement?

```
ex2_r <- lm(SWB ~ Treatment + Hospital, data = hosp_tbl)
ex2_f <- lm(SWB ~ Treatment*Hospital, data = hosp_tbl)

anova(ex2_r, ex2_f)
```

```
## Analysis of Variance Table
##
## Model 1: SWB ~ Treatment + Hospital
## Model 2: SWB ~ Treatment * Hospital
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    176 1106.51
## 2    174  714.34  2    392.18 47.764 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Application to example 3

- Does personality signficantly predict self-rated health over and above the effects of age and sex?

```
ex3_r <- lm(health ~ age + sex, data = midus2)
ex3_f <- lm(health ~ age + sex + O + C + E + A + N, data = midus2)

anova(ex3_r, ex3_f)
```

```
## Analysis of Variance Table
##
## Model 1: health ~ age + sex
## Model 2: health ~ age + sex + O + C + E + A + N
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1758 4740.2
## 2   1753 4055.4  5    684.85 59.208 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Time for a break

Welcome Back!

Nested vs non-nested models and alternatives to the F-test

Nested vs non-nested models

- The F-ratio depends on the models being compared being nested
- Nested means that the predictors in one model are a subset of the predictors in the other
- We also require the models to be computed on the same data

Nested vs non-nested models

Nested

```
m0 <- lm(outcome ~ x1 + x2 , data = data)

m1 <- lm(outcome ~ x1 + x2 + x3, data = data)
```

- These models are nested.

- `x1` and `x2` appear in both models

Non-nested

```
m0 <- lm(outcome ~ x1 + x2 + x4, data = data)

m1 <- lm(outcome ~ x1 + x2 + x3, data = data)
```

- These models are non-nested

- There are unique variables in both models

    - `x4` in `m0`
    - `x3` in `m1`

Model comparison for non-nested models

- So what happens when we have non-nested models?

- There are two commonyl used alternatives

    - AIC
    - BIC

AIC

$$AIC = n \ln\left(\frac{SS_{residual}}{n}\right) + 2k$$

$$\begin{aligned}
&\text{Where:}\\
&SS_{residual} = \text{sum of squares residuals}\\
&n = \text{sample size}\\
&k = \text{number of explanatory variables}\\
&\text{ln} = \text{natural log function}
\end{aligned}$$

- Unlike the incremental F-test AIC does not require two models to be nested
- Smaller (more negative) values of AIC indicate better fitting models.
  - So we compare values and choose the model with the smaller AIC

AIC parsimony correction

$$AIC = n \ln\left(\frac{SS_{residual}}{n}\right) + 2k$$

- Main point to note is that the term $2k$ applies a penalty for having more predictors
- When you add more predictors, fit will improve (SSE will get smaller)
- The decrease is partially offset by the $+2k$
- This makes AIC a parsimony-corrected statistic
- Parsimony-corrected statistics help us avoid over-fitting

In R

```r
AIC(ex3_r, ex3_f)
```

```
##        df      AIC
## ex3_r  4 6749.246
## ex3_f  9 6484.457
```

Applied to non-nested models

```
ex3_nn1 <- lm(health ~ O + C + E + A + N, data=midus2)
ex3_nn2 <- lm(health ~ age + sex + PIL, data = midus2)
AIC(ex3_nn1, ex3_nn2)
```

```
##         df      AIC
## ex3_nn1  7 6501.524
## ex3_nn2  5 6564.953
```

BIC

$$BIC = n\ln\left(\frac{SS_{residual}}{n}\right) + k\ln(n)$$

Where:

$SS_{residual} =$ sum of squares residuals

$n =$ sample size

$k =$ number of explanatory variables

$\ln =$ natural log function

- Like AIC...
  - BIC doesn't require nested models
  - Smaller (more negative) BIC values mean better models
  - We can compare the BICs for two models and choose the one with the smaller BIC as the better model

In R

```
BIC(ex3_r, ex3_f)
```

```
##       df      BIC
## ex3_r  4 6771.141
## ex3_f  9 6533.719
```

```
BIC(ex3_nn1, ex3_nn2)
```

```
##         df      BIC
## ex3_nn1  7 6539.840
## ex3_nn2  5 6592.321
```

Parsimony corrections

$$AIC = n\ln\left(\frac{SS_{residual}}{n}\right) + 2k$$

$$BIC = n\ln\left(\frac{SS_{residual}}{n}\right) + k\ln(n)$$

- BIC has a ‘harsher’ parsimony penalty for typical sample sizes when applying linear models than AIC
- When $\ln(n) > 2$ BIC will have a more severe parsimony penalty (i.e. essentially all the time!)

Considerations for use of AIC and BIC

- The AIC and BIC for a single model are not meaningful
    - They only make sense for model comparisons
- AIC and BIC can be used for both nested and non-nested models.
- For AIC, there are no cut-offs to suggest how big a difference in two models is needed to conclude that one is substantively better than the other
- For BIC, a difference of 10 can be used as a rule of thumb to suggest that one model is substantively better than another

Summary of today

- We have set out the types of question that may require us to use model comparison methods.
- We have introduced the incremental F-test and linked it to the F-test from semester 1.
- We also introduced the concepts of nested and non-neste tests, and the use of AIC and BIC for model comparison of non-nested models.

Thanks for listening!