Two-Way ANOVA Model Selection Process

We will be using a “backward” selection process for deciding on the “best” ANOVA model. Meaning, we will start with a more complicated model and decide if it is “worth it.”

# Step 1 – Fit an Interaction Two-Way ANOVA

## Plot the model

|  |
| --- |
|  |

**What are we looking for?**

We are looking to see if relationship between the mean of and explanatory variable 1 changes over the values of explanatory variable 2. This is akin to looking for evidence of different slopes in a multiple linear regression.

For these data we are looking to see if the relationship between the mean movie rating and the genre of the movie changes over the different eras.

## Fit the model

*Notice I’m using a \* for an interaction model!*

aov(rating ~ era \* genre, data = movies) %>%   
 tidy()

# A tibble: 4 × 6  
 term df sumsq meansq statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 era 3 4.59 1.53 0.659 0.581  
2 genre 1 4.19 4.19 1.81 0.185  
3 era:genre 3 3.32 1.11 0.477 0.700  
4 Residuals 48 111. 2.32 NA NA

## Perform a Hypothesis Test

The era:genre line of the ANOVA table is testing if the mean movie rating and the genre of the movie changes over the different eras. It has the following hypotheses:

: The relationship between the mean movie rating and the genre of the movie does not change over the eras

: The relationship between the mean movie rating and the genre of the movie is different for at least one era

With a p-value of 0.7 (from an F-statistic of 0.477 with 3 and 48 degrees of freedom) at a significance level of 0.1, I fail to reject the null hypothesis. Thus, the data have unconvincing evidence that the mean movie rating and the genre of the movie is different for at least one era.

So, my next step is to fit an **additive** two-way ANOVA.

# Step 2: Fit an Additive Two-Way ANOVA

We are looking to see if the mean of differs across the values of explanatory variable 1 and / or explanatory variable 2. This is akin to fitting a parallel slopes model in a multiple linear regression, where the y-intercept is the estimated change in the mean of for each group and the slope is the estimated relationship between the explanatory variable and the mean of .

For these data we are looking to see if the genres have differences in the mean movie ratings **and** if the eras have differences in the mean movie rating.

## Fit the Model

*Notice I’m using a + for an interaction model!*

aov(rating ~ era + genre, data = movies) %>%   
 tidy()

# A tibble: 3 × 6  
 term df sumsq meansq statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 era 3 4.59 1.53 0.680 0.568  
2 genre 1 4.19 4.19 1.86 0.178  
3 Residuals 51 115. 2.25 NA NA

## Perform a Hypothesis Test

For an additive two-way ANOVA, we have **two** hypothesis tests, one test per explanatory variable.

### Hypothesis Test for genre

The genre line of the ANOVA table is testing if the mean movie rating is the same for all genres. It has the following hypotheses:

: The mean movie rating is the same for every genre, given the era of the movie

: The mean movie rating is different for at least one genre, given the era of the movie

*Notice the hypotheses are similar to a multiple linear regression with two explanatory variables, where the test for each variable is conditional on the other variable(s) included in the model.*

With a p-value of 0.178 (from an F-statistic of 1.86 with 1 and 51 degrees of freedom) at a significance level of 0.1, I fail to reject the null hypothesis. Thus, the data have unconvincing evidence that at least one genre has a different mean movie rating, given the era of the movie.

### Hypothesis Test for era

The era line of the ANOVA table is testing if the mean movie rating is the same for all eras. It has the following hypotheses:

: The mean movie rating is the same for every era, given the genre of the movie

: The mean movie rating is different for at least one era, given the genre of the movie

With a p-value of 0.568 (from an F-statistic of 0.68 with 3 and 51 degrees of freedom) at a significance level of 0.1, I fail to reject the null hypothesis. Thus, the data have unconvincing evidence that at least one era has a different mean movie rating, given the genre of the movie.

# Step 3: Fit a One-Way ANOVA

Technically, both variables had large p-values, but the p-value is conditional on the other variables in the model. So, the next step is to remove the variable with the **largest** p-value and fit a one-way ANOVA.

## Fit the Model

aov(rating ~ genre, data = movies) %>%   
 tidy()

# A tibble: 2 × 6  
 term df sumsq meansq statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 genre 1 4.68 4.68 2.13 0.150  
2 Residuals 54 119. 2.20 NA NA

## Perform a Hypothesis Test

The genre line of the ANOVA table is testing if the mean movie rating is the same for all genres. It has the following hypotheses:

: The mean movie rating is the same for every genre

: The mean movie rating is different for at least one genre

*Notice the hypotheses don’t include era this time, since that variable has been removed from the model!*

With a p-value of 0.15 (from an F-statistic of 2.13 with 1 and 54 degrees of freedom) at a significance level of 0.1, I fail to reject the null hypothesis. Thus, the data have unconvincing evidence that at least one genre has a different mean movie rating.

# Step 4: Fit a Mean-Only Model

Once we’ve failed to reject the null in a one-way ANOVA, the “best” model is a model with **one** mean for **all** observations.

summarize(movies,  
 mean\_rating = mean(rating)  
 )

# A tibble: 1 × 1  
 mean\_rating  
 <dbl>  
1 5.64