

# Analysis of Variance

BIOE 498/598 PJ

Spring 2021

# The sum of squares

Our analysis is based on the *sum of squares*, or SS. In particular, the total SS is the combination of the SS explained by our model and the SS that is residual (or unexplained).

$$SS_{\text{total}} = SS_{\text{explained}} + SS_{\text{residual}}$$

# The sum of squares

Our analysis is based on the *sum of squares*, or SS. In particular, the total SS is the combination of the SS explained by our model and the SS that is residual (or unexplained).

$$SS_{\text{total}} = SS_{\text{explained}} + SS_{\text{residual}}$$

How do we calculate each one for a model  $\mathbf{y} = \mathbf{X}\beta$ ?

# The sum of squares

Our analysis is based on the *sum of squares*, or SS. In particular, the total SS is the combination of the SS explained by our model and the SS that is residual (or unexplained).

$$SS_{\text{total}} = SS_{\text{explained}} + SS_{\text{residual}}$$

How do we calculate each one for a model  $\mathbf{y} = \mathbf{X}\beta$ ?

$$SS_{\text{total}} = \sum_i (y_i - \text{mean}(\mathbf{y}))^2$$

# The sum of squares

Our analysis is based on the *sum of squares*, or SS. In particular, the total SS is the combination of the SS explained by our model and the SS that is residual (or unexplained).

$$SS_{\text{total}} = SS_{\text{explained}} + SS_{\text{residual}}$$

How do we calculate each one for a model  $\mathbf{y} = \mathbf{X}\beta$ ?

$$SS_{\text{total}} = \sum_i (y_i - \text{mean}(\mathbf{y}))^2$$

$$SS_{\text{residual}} = \sum_i (y_i - \text{predicted}(y_i))^2$$

# The sum of squares

Our analysis is based on the *sum of squares*, or  $SS$ . In particular, the total  $SS$  is the combination of the  $SS$  explained by our model and the  $SS$  that is residual (or unexplained).

$$SS_{\text{total}} = SS_{\text{explained}} + SS_{\text{residual}}$$

How do we calculate each one for a model  $\mathbf{y} = \mathbf{X}\beta$ ?

$$SS_{\text{total}} = \sum_i (y_i - \text{mean}(\mathbf{y}))^2$$

$$SS_{\text{residual}} = \sum_i (y_i - \text{predicted}(y_i))^2$$

$$SS_{\text{explained}} = SS_{\text{total}} - SS_{\text{residual}}$$

# Does our model do anything?

Let's analyze the data from the stuffed monkey throwing experiment.

```
##
## Call:
## lm(formula = distance ~ hand + hat + boots)
##
## Residuals:
##      1      2      3      4      5      6      7      8
## -0.375 -1.125  0.625  0.875  1.125  0.375 -1.375 -0.125
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.375      0.857   6.272  0.0033 **
## handright      2.250      0.857   2.626  0.0585 .
## hatyes        -1.500      0.857  -1.750  0.1549
## bootsyes       1.000      0.857   1.167  0.3081
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.212 on 4 degrees of freedom
## Multiple R-squared:  0.7389, Adjusted R-squared:  0.5431
## F-statistic: 3.773 on 3 and 4 DF,  p-value: 0.1161
```

## For our throwing data

```
ss <- function(x) sum(x^2)
sst <- ss(distance - mean(distance))
ssr <- ss(residuals(model))
ssx <- sst - ssr
c(sst, ssr, ssx)
```

```
## [1] 22.500  5.875 16.625
```



# Degrees of freedom

The amount of variation we expect to see depends on the number of independent parameters in the model. These are the *degrees of freedom*, and we need to normalize the SS by them.

For analyzing variation, the number of parameters does not include the intercept.

- ▶ For  $SS_{\text{total}}$ ,  $DF = (\# \text{ data points}) - 1$
- ▶ For  $SS_{\text{explained}}$ ,  $DF = \# \text{ of parameters}$
- ▶ For  $SS_{\text{residual}}$ ,  $DF = (\# \text{ data points}) - (\# \text{ parameters}) - 1$

## The $F$ -statistic

The value of our model is explained by the ratio between the explained variance and the residual (unexplained) variance *after adjusting for the DF*.

$$F = \frac{SS_{\text{explained}}/DF(SS_{\text{explained}})}{SS_{\text{residual}}/DF(SS_{\text{residual}})}$$

## The $F$ -statistic

The value of our model is explained by the ratio between the explained variance and the residual (unexplained) variance *after adjusting for the DF*.

$$F = \frac{SS_{\text{explained}}/DF(SS_{\text{explained}})}{SS_{\text{residual}}/DF(SS_{\text{residual}})}$$

For our throwing example

$$F = \frac{16.625/3}{5.875/(8 - 3 - 1)} = 3.773$$

# The $F$ -statistic

The value of our model is explained by the ratio between the explained variance and the residual (unexplained) variance *after adjusting for the DF*.

$$F = \frac{SS_{\text{explained}}/DF(SS_{\text{explained}})}{SS_{\text{residual}}/DF(SS_{\text{residual}})}$$

For our throwing example

$$F = \frac{16.625/3}{5.875/(8 - 3 - 1)} = 3.773$$

**How big should the  $F$ -statistic be?** The  $F$ -statistic follows the  $F$ -distribution. We can use this distribution to convert the  $F$ -statistic into a  $p$ -value.

```
summary(model)
```

```
##
## Call:
## lm(formula = distance ~ hand + hat + boots)
##
## Residuals:
##      1      2      3      4      5      6      7      8
## -0.375 -1.125  0.625  0.875  1.125  0.375 -1.375 -0.125
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.375      0.857   6.272  0.0033 **
## handright      2.250      0.857   2.626  0.0585 .
## hatyes        -1.500      0.857  -1.750  0.1549
## bootsyes       1.000      0.857   1.167  0.3081
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.212 on 4 degrees of freedom
## Multiple R-squared:  0.7389, Adjusted R-squared:  0.5431
## F-statistic: 3.773 on 3 and 4 DF,  p-value: 0.1161
```

## Testing single factors

We previously compared the entire model against the residuals to see if the model added value. We can apply the same procedure to a single variable.

This is called the *analysis of variance*, or ANOVA.

# ANOVA on handedness

Let's find the explained variance for a model with only handedness:

```
model_hand <- lm(distance ~ hand)
sst - ss(residuals(model_hand))
```

```
## [1] 10.125
```

Now let's compare this to the residuals of the entire model:

# ANOVA on handedness

Let's find the explained variance for a model with only handedness:

```
model_hand <- lm(distance ~ hand)
sst - ss(residuals(model_hand))
```

```
## [1] 10.125
```

Now let's compare this to the residuals of the entire model:

$$F = \frac{10.125/1}{5.875/(8 - 3 - 1)} = 6.894$$



## ANOVA on a linear model

We can repeat this procedure for every variable, or we can use R's built-in ANOVA command.

```
anova(model)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: distance
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## hand         1 10.125  10.1250    6.8936 0.05846 .
## hat          1  4.500   4.5000    3.0638 0.15495
## boots        1  2.000   2.0000    1.3617 0.30807
## Residuals    4  5.875   1.4688
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Conclusions

- ▶  $p$ -values on effect sizes tell us if the effect size is nonzero.
- ▶ A significant effect size does not mean the effect matters.
- ▶ ANOVA can tell us which variables explain a significant fraction of the variance in our data.
- ▶ Significance is relative to the unexplained variance in the model.