Introduction to ANOVA for Linear Regression

We talked about how r^2 represents the percent of the variability in the y-values that is explained by the model. We will make that precise here.

In linear regression, we have

$$\text{Data} = \text{Model} + \text{Residuals}$$

$$y_i = \hat{y}_i + (y_i - \hat{y}_i)$$

Each of the three terms in this equation has variability that can be measured by calculating a **sum of squares**.

Sum of Squares

- **Sum of Squares Total:** $SST = \sum(y_i - \bar{y})^2$.
  Measures how much the y-values deviate from $\bar{y}$.

- **Sum of Squares Model:** $SSM = \sum(\hat{y}_i - \bar{y})^2$.
  Measures how much the predicted y-values deviate from $\bar{y}$.

- **Sum of Squares Error:** $SSE = \sum(y_i - \hat{y}_i)^2$.
  Measures how much the residuals deviate from zero.

It is a linear algebra fact that

$$SST = SSM + SSE.$$

Meaning of r^2

We can now explain what we meant when we said that r^2 represents the percent of the variability of the y-values that is explained by the model. What that really means is:

$$r^2 = \frac{SSM}{SST}.$$

You can prove this by combining the variance formulas:

$$s_x^2 = \sum \frac{(x_i - \bar{x})^2}{n-1} \text{ and } s_y^2 = \sum \frac{(y_i - \bar{y})^2}{n-1},$$

with the linear regression model which predicts

$$\hat{y}_i = r\frac{s_y}{s_x}(x_i - \bar{x}) + \bar{y}.$$

Degrees of Freedom

Each of the sum of squares formulas above has a degrees of freedom. These come from the dimensions of the subspaces where the corresponding vectors reside.

- **Degrees of Freedom Total** SST has $DFT = n - 1$.
- **Degrees of Freedom Model** SSM has $DFM = 1$.
- **Degrees of Freedom Error** SSE has $DFE = n - 2$.

Another linear algebra fact is that:

$$DFT = DFM + DFE.$$

Mean Squares

Divide a sum of squares by the corresponding degrees of freedom to get what statisticians call a **mean square**.

- **Mean Square Total** $MST = \frac{SST}{DFT} = \frac{\sum(y_i - \bar{y})^2}{n-1}$.
  This is s_y^2 which is the best estimate for the variance of y.

- **Mean Square Model** $MSM = \frac{SSM}{DFM} = \frac{\sum(\hat{y}_i - \bar{y})^2}{1}$.

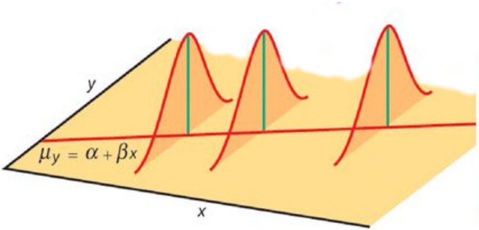
- **Mean Square Error** $MSE = \frac{SSE}{DFE} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$.
  This is the best estimate for the variance of the residuals in the population.

Residual Standard Error

We call the square root of the mean squared error the **residual standard error** and we denote it s. That is:

$$s = \sqrt{MSE}.$$

Linear regression models assume that the residuals are normally distributed with a standard deviation of σ. The residual standard error s is our best estimate for σ.



F-values

If there is no association between the x and y variables in a scatterplot, then the expression $F = \frac{MSM}{MSE}$ has an F-distribution with DFM degrees of freedom in the numerator and DFE degrees of freedom in the denominator. You can use this F-value to test whether there is a significant association between two variables in a scatterplot.

ANOVA Tables

All of the information above can be kept straight using an **ANOVA table**.

Source	Deg. of Freedom	Sum of Squares	Mean Square	F-value
Model	*DFM*	*SSM*	*MSM*	*F*
Error	*DFE*	*SSE*	*MSE*	
Total	*DFT*	*SST*	*MST*	

ANOVA Tables

Typically, you can fill in the entries of the ANOVA table by following these steps:

1. Find s_y^2. This is the MST.
2. Multiply MST by $n-1$ to find SST.
3. Find r^2 and use the fact that $r^2 = SSM/SST$ to find SSM.
4. Use the fact that $SSM + SSE = SST$ to find SSE.
5. Divide to find MSM, MSE, and F.

Summary

1. $r^2 = SSM/SST$.

2. $SST = SSM + SSE$ and $DFT = DFM + DFE$.

3. For each source, the mean square is the sum of squares divided by the degrees of freedom.

4. The sample variance of y is $s_y^2 = MST$.

5. The residual standard error is $s = \sqrt{MSE}$.

6. If there is no association between x & y, then $F = MSM/MSE$ has an F-distribution.