Multiple linear regression models allow us to assess the relationship between a continuous outcome Y and a set of predictors $X_1, \ldots, X_p$. One (now very familiar!) way of writing this mathematically is

$$E[Y|X_1, \ldots, X_p] = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p.$$

For the first few weeks of class, we've mainly focused on how to construct this model and select the predictors $X_1, \ldots, X_p$.

Hypothesis Testing and Linear Models

But we also want to be able to use this model to draw statistical conclusions about our predictors $X_1, \ldots, X_p$ and their relationship with Y! There are several main types of hypotheses that we may want to test:

- $H_0 : \beta_j = 0$ for a specific predictor X_j
- $H_0 : \beta_j = \beta_k$ for two predictors X_j and X_k
- $H_0 : \beta_i = \beta_j = \beta_k = 0$ for a collection of predictors $\{X_i, X_j, X_k\}$
- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

$\boxed{H_0 : \beta_j = 0}$

Here, we want to assess whether a statistically significant relationship between X_j and Y exists. In other words, we want to answer the question: is X_j a statistically significant predictor of our outcome, Y?

Formal Hypothesis:	$H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$
Test Statistic:	$t = \frac{\hat{\beta}_j - 0}{\text{s.e.}(\hat{\beta}_j)} \sim t_{n-(p+1)}$
Confidence Interval:	$\hat{\beta}_j \pm t_{n-(p+1),1-\frac{\alpha}{2}} \cdot \text{s.e.}(\hat{\beta}_j)$

$\boxed{H_0 : \beta_j = \beta_k}$

Here, we want to assess whether a linear combination of the βs is significantly different than zero.

Formal Hypothesis:	$\big(H_0 : \beta_j = \beta_k$ versus $H_1 : \beta_j \neq \beta_k\big)$ or $\big(H_0 : \beta_j - \beta_k = 0$ versus $H_1 : \beta_j - \beta_k \neq 0\big)$
Test Statistic:	$t = \frac{(\hat{\beta}_j - \hat{\beta}_k) - 0}{\text{s.e.}(\hat{\beta}_j - \hat{\beta}_k)} \sim t_{n-(p+1)}$
Confidence Interval:	$(\hat{\beta}_j - \hat{\beta}_k) \pm t_{n-(p+1),1-\frac{\alpha}{2}} \cdot \text{s.e.}(\hat{\beta}_j - \hat{\beta}_k)$

This type of hypothesis test is also particularly useful in the setting where X_k is an interaction term, and where $\beta_j + \beta_k$ is the slope for the association between X_j and Y within a particular level of an effect modifier!

Formal Hypothesis:	$H_0 : \beta_j + \beta_k = 0$ versus $H_1 : \beta_j + \beta_k \neq 0$
Test Statistic:	$t = \frac{(\hat{\beta}_j + \hat{\beta}_k) - 0}{\text{s.e.}(\hat{\beta}_j + \hat{\beta}_k)} \sim t_{n-(p+1)}$
Confidence Interval:	$(\hat{\beta}_j + \hat{\beta}_k) \pm t_{n-(p+1),1-\frac{\alpha}{2}} \cdot \text{s.e.}(\hat{\beta}_j + \hat{\beta}_k)$

$\boxed{H_0 : \beta_i = \beta_j = \beta_k = 0}$

Here, we want to determine whether a particular subset of covariates—here X_i, X_j, and X_k—contributes significantly to our model. So we want to test whether, after accounting for all other covariates, X_i, X_j and X_k collectively explain a significant proportion of the remaining variability in Y.

We can equivalently view the test of the null hypothesis $H_0 : \beta_i = \beta_j = \beta_k = 0$ as comparing the fit of the **reduced model** without the covariates X_i, X_j, and X_k,

$$E[Y|X] = \beta_0 + \beta_1 \cdot X_1 + \ldots + 0 \cdot X_i + \ldots + 0 \cdot X_j + \ldots + 0 \cdot X_k + \ldots + X_p,$$

to the fit of the **full model**

$$E[Y|X] = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p.$$

So we can summarize this test by:

Formal Hypothesis:	$H_0 : \beta_i = \beta_j = \beta_k = 0$ versus H_1 : at least one of β_i, β_j and β_k is not 0
or	
Formal Hypothesis:	$\big(H_0$: the reduced model is sufficient$\big)$ versus $\big(H_1$: the full model is preferred$\big)$
Test Statistic:	$F = \frac{(SSE_{reduced} - SSE_{full})/3}{SSE_{full}/(n-(p+1))} \sim F_{3,n-p-1}$

We can fairly easily generalize this to a subset of the covariates of size r and—more broadly speaking—can use this kind of F-test (also known as an ANOVA) to compare any two **nested** models.

$\boxed{H_0 : \beta_1 = \ldots \beta_k = 0}$

Here, we want to determine whether the mean model, $E[Y|X] = \beta_0$, is alone sufficient to explain the variability in our outcome Y.

Formal Hypothesis:	$H_0 : \beta_1 = \ldots = \beta_p = 0$ versus H_1 : at least one of $\beta_1, \ldots, \beta_p$ is not 0
Test Statistic:	$F = \frac{(SSE_{reduced} - SSE_{full})/p}{SSE_{full}/(n-(p+1))} = \frac{(SST - SSE_{full})/p}{SSE_{full}/(n-(p+1))} \sim F_{p,n-p-1}$