STAT 308 – Chapter 9

Background Information

We have introduced the concept of multiple linear regression where we introduce linearly regressing a continuous response variable Y on multiple explanatory variables X_1, X_2, ..., X_p. This chapter will introduce how we perform hypothesis tests on these regression lines.

Test for Overall Regression

Recall, the method for testing for a significant linear relationship in simple linear regression.

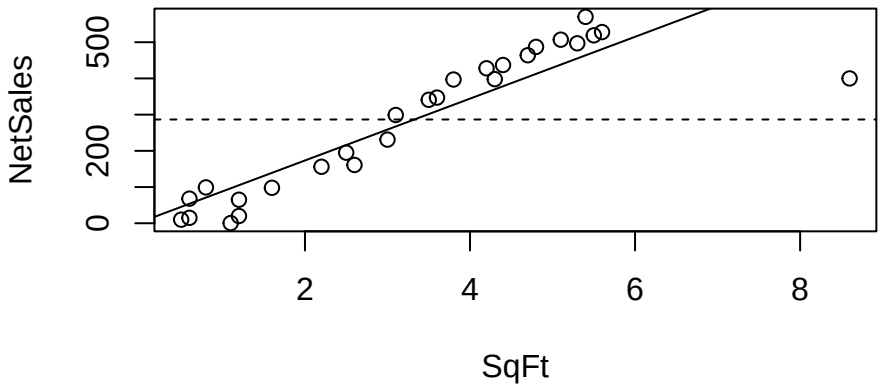
The null and alternative hypotheses are:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

.

In other words we are testing to see if the solid line in the below graph is significantly better at predicting Y than the horizontal dashed line.

```
allgreen <- read.csv("../data/AllGreen.csv")
plot(NetSales ~ SqFt,allgreen)
abline(lm(NetSales ~ SqFt,allgreen))
abline(h=mean(allgreen$NetSales),lty=2)
```



What is this equivalent in multiple dimensions?

What is the null and alternative hypothesis equivalent of testing if the non-horizontal linear plane is better at predicting than the horizontal linear plane?

Test Statistic and p-value

Recall the ANOVA table from simple linear regression:

	df	Sums of Squares	Mean Square	f Value	Pr(>f)
Model	1	SSM = SSY − SSE	MSM = $\frac{\text{SSM}}{1}$	$\frac{\text{MSM}}{\text{MSE}}$	$Pr(F_{1,n-2} > \frac{\text{MSM}}{\text{MSE}})$
Error	$n-2$	SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MSE = $\frac{\text{SSE}}{n-2}$		
Total	$n-1$	SSY = $\sum_{i=1}^{n}(y_i - \bar{y})^2$			

We said that the appropriate test statistic is $\frac{MSM}{MSE}$ and p-value is $Pr(F_{1,n-2} > \frac{\text{MSM}}{\text{MSE}})$. This is because, under the null the null hypothesis, the sum of squares for the model and the sum of squared errors are both independent estimates of the same variance, or equivalently, the linear model with X does not explain a significant amount of the variation in Y.

How does this compare with the full ANOVA table for multiple linear regression?

	df	Sums of Squares	Mean Square	f Value	Pr(>f)
Model	p	SSM = SSY − SSE	MSM = $\frac{\text{SSM}}{p}$	$\frac{\text{MSM}}{\text{MSE}}$	$Pr(F_{p,n-(p+1)} > \frac{\text{MSM}}{\text{MSE}})$
Error	$n-(p+1)$	SSE	MSE = $\frac{\text{SSE}}{n-(p+1)}$		
Total	$n-1$	SSY = $\sum_{i=1}^{n}(y_i - \bar{y})^2$			

This provides us test statistics and p-values for the test for overall significance of our linear regression model.

Example

Perform a test for a significant linear regression for All Green's net sales with both advertising dollars and square footage as predictors.

Test for Partial Significance

Now, suppose we know that advertising is a significant linear predictor for net sales for All Green. Does adding square footage to our linear model significantly improve the prediction of net sales?

If our regression line is $\hat{NetSales} = \beta_0 + \beta_1 Advertising + \beta_2 Sales$, then the null and alternative hypotheses are:

What about the test statistic and p-value?

Using summary(mod)

One way is to note that, generically speaking:

$$t = \frac{\hat{\beta}_k - \beta_k}{s_{\hat{\beta}_k}} \sim t_{df=n-(p+1)}$$

for $k = 0, 1, \ldots, p$. This naturally leads to a test statistic:

$$t = \frac{\hat{\beta}_k - 0}{s_{\hat{\beta}_k}} = \frac{\hat{\beta}_k}{s_{\hat{\beta}_k}}$$

. We can find this information using `summary(mod)`.

Using anova(mod)

The ANOVA table can actually be broken down into different sums of squares for each variable added to the model!

	df	Sums of Squares	Mean Square	f Value	Pr(>f)
X_1	1	$\text{SS}(X_1)$	$\text{MS}(X_1)$	$\frac{\text{MS}(X_1)}{\text{MSE}}$	$Pr(F_{1,n-(p+1)} > f)$
$X_2 \vert X_1$	1	$\text{SS}(X_2 \vert X_1)$	$\text{MS}(X_2 \vert X_1)$	$\frac{\text{MS}(X_2 \vert X_1)}{\text{MSE}}$	$Pr(F_{1,n-(p+1)} > f)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X_p \vert X_1, \ldots, X_{p-1}$	1	$\text{SS}(X_p \vert X_1, \ldots, X_{p-1})$	$\text{MS}(X_p \vert X_1, \ldots, X_{p-1})$	$\frac{\text{MS}(X_p \vert X_1, \ldots, X_{p-1})}{\text{MSE}}$	$Pr(F_{1,n-(p+1)} > f)$
Error	$n-(p+1)$	SSE	$\text{MSE} = \frac{\text{SSE}}{n-(p+1)}$		
Total	$n-1$	$\text{SSY} = \sum_{i=1}^{n}(y_i - \bar{y})^2$			

where

- $\text{SS}(X_1)$ are the sums of squares for the model with only X_1 as a predictor (i.e. $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$)
- $\text{SS}(X_2|X_1) = \text{SS}(X_1, X_2) - \text{SS}(X_1)$ (i.e. the additional model sums of squares when X_2 is added as a predictor to the model with X_1 already included)
- $\text{SS}(X_p|X_1, X_2, \ldots, X_{p-1}) = \text{SS}(X_1, X_2, \ldots, X_p) - \text{SS}(X_1, X_2, \ldots, X_{p-1})$ (i.e. the additional model sums of squares when X_p is added as a predictor to the model with $X_1, X_2, \ldots, X_{p-1}$ already included)
- SSE are the sum of squared errors for the model with all p predictor variables included.

Let's see the differences between the simple linear regression ANOVA table and the multiple linear regression ANOVA table.

Why are the results same for answering the does adding square footage to our linear model with advertising already included significantly improve the prediction of net sales for the `summary(mod)` method and the `anova(mod)` method? Recall that if a test statistic, t, follows a t-distribution with $n-(p+1)$ degrees of freedom, then

$$t^2 \sim F_{df1=1, df2=n-(p+1)}$$

Example

Suppose now I want to test whether or not including both square footage and amount of inventory significantly improves the linear model with advertising already included. How would I perform this?

Reduced Model: Model with some (but not all) explanatory variables excluded

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k, \sigma^2)$$

Full Model: Model with all explanatory variables included.

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \sigma^2)$$

where $p > k$

Just like we said $F = \frac{\text{SS}(X_p|X_1,\ldots,X_{p-1})/1}{SSE/(n-(p+1))}$ can be used as a test statistic, we can obtain a similar test statistic for testing if adding multiple explanatory variables can improve our model's predictability:

$$F = \frac{\text{SS}(X_{k+1},\ldots,X_p|X_1,X_2,\ldots,X_k)/(p-k)}{SSE/(n-(p+1))}$$

where $\text{SS}(X_{k+1},\ldots,X_p|X_1,X_2,\ldots,X_k)$ are the model sums of squares added when going from the reduced model (k predictors) to the full model (p predictors). This test statistic under H_0 follows an F-distribution with $p-k$ and $n-(p+1)$ degrees of freedom. Another way to write this statistic is

$$F = \frac{(\text{SSE}_{reduced} - \text{SSE}_{full})/(df_{reduced} - df_{full})}{\text{SSE}_{full}/df_{full}}$$

where

- $\text{SSE}_{reduced}$ and $df_{reduced}$ are the sum of squared errors and error degrees of freedom for the reduced model
- SSE_{full} and df_{full} are the sum of squared errors and error degrees of freedom for the reduced model

Not only can ANOVA break down the sums of squares by each variable for a given model, it can also compare the sums of squares by two competing models!

Res.df	SSE	df	Sum of Squares Added	F	p-value
$df_{reduced}$	$\text{SSE}_{reduced}$				
df_{full}	SSE_{full}	$df_{reduced} - df_{full}$	$\text{SSE}_{reduced} - \text{SSE}_{full}$	F	p-value