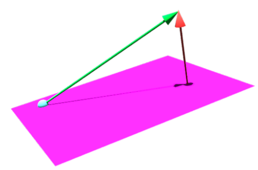
## Residuals: Setup

Working with the entire residual vector is always the best way of checking the goodness of fit of our model. Still some people keep on searching for a single number that can be used to measure goodness of fit.

Although such numbers are popular, it should be kept in mind that it is never possible to assess a complicated thing such as goodness of fit with a single score or estimate.

Nevertheless,since those things are popular,it is necessary to have some idea about how they work.

All those techniques of finding a ”magic” number to measure the goodness of fit are primarily based on a familiar picture:



We have the column space of X(pink),our data vector Y(green) and we are projecting this data vector orthogonally on the column space of X.The distance of this from the column space should reflect in some way how good the fit is. So the natural choice is to look at this error vector(red).

Now,we can look at this error vector in it's full glory.It consists of n numbers where n is the ambient dimension(3 in this case).We would naturally want to know how large this vector is.

There are various ways in which we can use it.The simplest is to look at the length and scale it appropriately to get an unbiased estimator of sigma square. Sigma hat is then the square length of this vector divided by degree of freedom.So we can use this as a way of measuring the goodness of fit of our model.

This type I estimate is not without problems of it's own.If we change the units of our measurements,the length invariably changes.Thus our estimate may be small for inches and large for mm.Certainly,the goodness of fit of our data cannot change depending on what units we are using.Still, this sigma hat estimate is used by some people to crudely measure how good the fit is.

The type II thing is kind of working on scaling the squared norm of the vector such that it does not depend on the units anymore and requires some linear algebra.We shall be following the type II method next.