Multiple Linear Regression

When we talk about regression, we’re referring to regression to the mean. We are able to collect data and create a linear relationship between variables in certain situations. Simple linear regression is easy to explain – quantifying the relationship between just two variables-- the larger the house, the more expensive it will be.

Multiple linear regression gets a bit more complicated as we introduce additional explanatory variables. We can predict the price of a house looking at other factors besides its size, like its neighborhood, quality of the school district, proximity to a city, or other variables that determine the value of a residential property. The difficulty comes in once we try to understand all these moving parts together.

Let’s take the correlation between home size and home price. As the size increases, the price does as well – this would look like a straight diagonal line going up and to the right. When we include other variables like location and number of bedrooms, it becomes increasingly difficult to plot the relationships because we stack the dimensions. Personally, I find it difficult to read and understand 3 dimensional charts – adding more dimensions would make my head explode.

So how do we start to understand the results of a multiple linear regression if we can’t read simple charts? I’d like to talk about two different methods – analyzing residuals and holding all other variables constant.

The first method, residual analysis, is a pathway to determining whether a model is appropriate or not. Looking at residuals applies to simple linear regression as well and allows us to determine if the assumptions of regression are being met. For simple linear regression, we can look at the correlation between the two variables and decide whether regression fits, which allow us to conceptualize how the variables relate. For multiple linear regression, we need to form an appropriate model with statistically significant variables before understanding any one of the explanatory variables alone.

Once we have a model, we can understand the effects of each of the explanatory variables, holding all other variables constant. We look at the coefficients of the model to determine the effect of each variable. Positive coefficients suggest a positive relationship and negative coefficients suggest negative relationships. If we have a reasonable model that predicts home prices using home size, home neighborhood, and proximity to a city where the coefficients are positive for home size and negative for proximity to a city, we can say- with neighborhood and proximity to a city held constant, home prices increase with size. Alternatively, holding home size and location constant, home prices decrease as you get further from a city.