Overview of Models

So let's take a look at the models that we use in this course. We'll study the general linear model, an umbrella term for several different analyses, and also logistic regression. In each case, there's a response variable and predictor variables. When the response variable is continuous and you can assume a normal distribution of errors, you can use a general linear model to model the relationship between predictor variables and the response variable.

$$Y = \beta_0 + \beta_1 X_1 ... + \beta_k X_k + \varepsilon$$

The formula is Y, the response, equal to a linear function of the predictors considered, the Xs, and the unknown parameters, the Betas, which we estimate from the data. In this case, we're attempting to predict the response, with k predictors. ε accounts for the unexplained variation in our model.

When will you use ANOVA, or analysis of variance? If you have a response variable that's continuous and all of your predictor variables are categorical, your best approach in terms of a statistical method is going to be ANOVA. With ANOVA, you're looking at how changing the level of your predictors can affect Y, your response variable. For example, what if you wanted to know how sale price relates to the heating quality of a home. This predictor, heating quality, has four levels: excellent, good, average, and fair. Our ANOVA model can explain how the sale price changes from one level of heating quality to another.

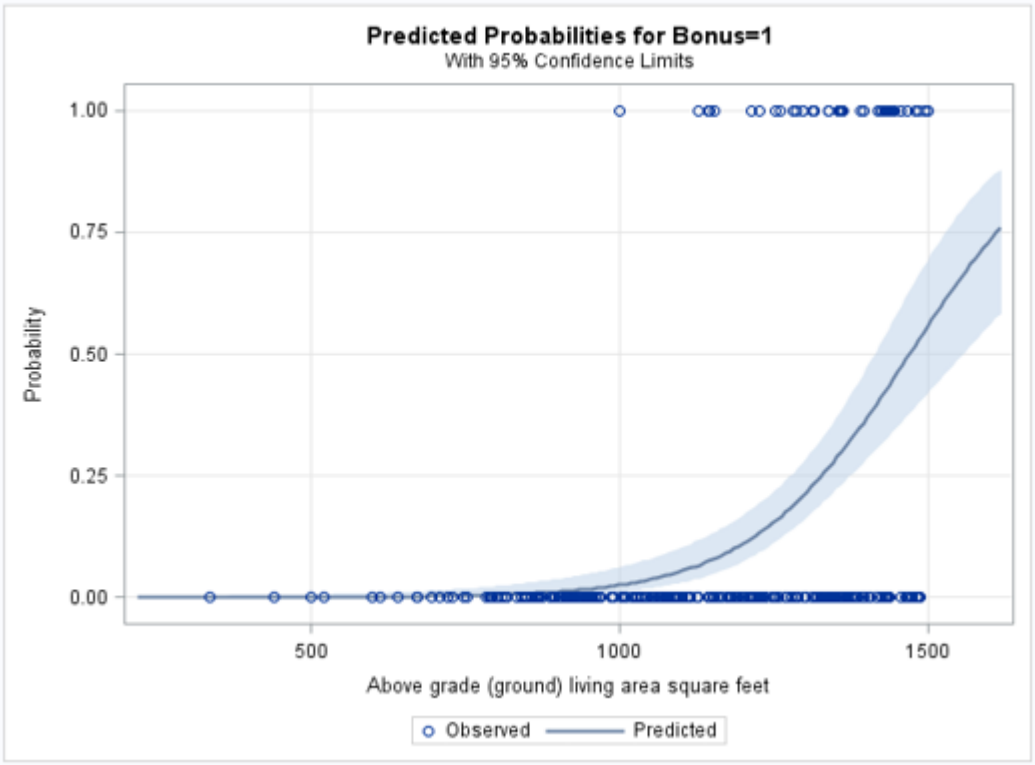
If your response is continuous and all of your predictors are continuous, then you're going to use ordinary least squares regression. In the simplest case, imagine you want to predict the sale price of a home using the size of the home in square feet. Ideally, the sampled data exhibit some kind of linear relationship. The regression model indicates what the expected response, or sale price, would be for each value of square feet.

What if you have a categorical response variable that's binary? Well, then regardless of your predictors, logistic regression is going to the be the optimum choice in terms of modeling. In logistic regression, you model the probability of an event given a set of predictors. Here's the formula.

$$\log it(Y) = \beta_0 + \beta_1 X_1 ... + \beta_k X_k$$

The logit of Y is the logit transformation, or the log odds transformation, of the probability of the event. β_0 is the intercept of the equation, and β_1 is the slope parameter for X_1.

So imagine that homes in Iowa selling for more than $175,000 are eligible for a tax incentive, and you want to estimate the probability of this event given the square footage of the home. Here, a value of 0 indicates that the home is not eligible for a tax incentive, and a value of 1 indicates that it is.



The logistic regression model predicts the probability of an incentive-eligible home with a sigmoidal curve. You can see that the estimated probability of being eligible increases as the size of the home increases.

Close