In the previous video, we only used one independent variable, but there are many different variables that could be used to predict wine price.

We used average growing season temperature, but we also have data for other weather-related variables-- harvest rain and winter rain.

Additionally, the age of wine is suspected to be important, and many other variables could also be used, such as the population of France.

We can use each variable in a one variable regression model.

Average growing season temperature gives the best R squared of 0.44, followed by harvest rain within R squared of 0.32.

France's population and age, both give models within R squared around 0.2, and winter rain gives a pretty low R squared of 0.02, or just barely better than the baseline.

So if we only used one variable, average growing season temperature is the best choice.

But multiple linear regression allows you to use multiple variables at once to improve the model.

The multiple linear regression model is similar to the one variable regression model that has a coefficient beta for each independent variable.

We predict the dependent variable y using the independent variables x1, x2, through xk, where k here denotes the number of independent variables in our model.

Beta 0 is, again, the coefficient for our intercept term, and beta 1, beta 2, through beta k are the coefficients for the independent variables.

We use i to denote the data for a particular data point or observation.

The best model is selected in the same way as before.

To minimize the sum of squared errors, using the error terms, epsilon.

We can start by building a linear regression model that just uses the variable with the best R squared-- average growing season temperature.

We saw before that this gives us an R squared of 0.44.

Then we can add variables one at a time and look at the improvement in R squared.

Note that the improvement is not equal to the one variable R squared for each independent variable we add, since they're interactions between the independent variables.

Adding independent variables improves the R squared to almost double what it was with a single independent variable.

But there are diminishing returns.

The marginal improvement from adding an additional variable decreases as we add more and more variables.

So which model should we use?

Often not all variable should be used.

This is because each additional variable used requires more data, and using more variables creates a more complicated model.

Overly complicated models often cause what's known as overfitting.

This is when you have a higher R squared on data used to create the model, but bad performance on unseen data.

For example, suppose we want to use this model to make a prediction for the year 2013.

We expect an overfit model to perform poorly on this new data.

In the next video, we'll learn how to build a regression models in R and then we'll discuss how to select the variables that should be included in the final model.