In the previous video, we created linear regression models in R. Using the summary function, we were able to see the coefficients, as well as some other information.

The output of the coefficient section of the summary function is shown here.

The independent variables are listed on the left.

The estimate column gives the coefficients for the intercept and for each of the independent variables in our model.

The remaining columns help us determine if a variable should be included in the model, or if its coefficient is significantly different from 0.

A coefficient of 0 means that the value of the independent variable does not change our prediction for the dependent variable.

If a coefficient is not significantly different from 0, then we should probably remove the variable from our model since it's not helping to predict the dependent variable.

The standard error gives a measure of how much the coefficient is likely to vary from the estimate value.

The t value is the estimate divided by the standard error.

It will be negative if the estimate is negative, and positive if the estimate is positive.

The larger the absolute value of the t statistic, the more likely the coefficient is to be significant, so we want variables with a large absolute value in this column.

The last column gives the probability that a coefficient is actually 0.

It will be large if the absolute value of the t statistic is small, and it will be small if the absolute value of the t statistic is large.

We want variables with small values in this column.

This is a lot of information, but the easiest way in R to determine if a variable is significant is to look at the stars at the end of each row.

Three stars is the highest level of significance, and corresponds to a probability less than 0.001.

The star coding scheme is explained at the bottom of the coefficient output.

Three stars corresponds to probability values between 0 and 0.001, or the smallest possible probabilities.

Two stars is also very significant, and corresponds to a probability between 0.001 and 0.01.

One star is also significant, and corresponds to a probability between 0.01 and 0.05.

A period, or dot, means that the variable is almost significant, and corresponds to a probability between 0.05 and 0.1.

When we ask you to list the significant variables in the model, we will usually not include these.

Nothing at the end a row means that the variable is not significant in the model.

Age and FrancePopulation are both insignificant in our model.

Let's switch to R, and see if we can improve our model.

In the previous video, we built a linear regression model called model3 that used all of our independent variables to predict the dependent variable, Price.

In R console, we can see the summary output for this model.

By looking at the coefficient section, we can see that both Age and FrancePopulation are insignificant in our model.

Because of this, we should consider removing these variables from our model.

Let's start by just removing FrancePopulation, which we intuitively don't expect to be predictive of wine price anyway.

Let's create a new model called model4, which again uses the lm function to predict price using the independent variables, AGST, HarvestRain, WinterRain, and Age.

Here, we're not using FrancePopulation.

Our data set, again, is wine, which will be the data used to create our model.

Let's look at the summary of model4.

We can see that our R-squared in this model is 0.8286, and our adjusted R-squared is 0.79.

If we scroll back up to our previous model, our R-squared was 0.8294, and our adjusted R-squared was 0.784.

So this model is just as strong, if not stronger, than before, because our adjusted R-squared actually increased by removing FrancePopulation.

If we look at each of our variables and the stars, we now see that something a little strange happened.

Before, age was not significant at all in our model.

But now, the variable Age has two stars, meaning that it's very significant in this new model.

Why did this happen?

This is due to something called multi-colinearity.

Age and FrancePopulation are what we call highly correlated.

Let's learn a bit more about correlation.

Correlation measures the linear relationship between two variables, and is a number between +1 and -1.

The highest a correlation can be is positivev 1, which corresponds to a perfect positive linear relationship between the two variables.

The smallest a correlation can be is negative 1, which corresponds to a perfect negative linear relationship between the two variables.

In the middle of these two extremes is a correlation of 0, which corresponds to no linear relationship between the two variables.

Let's look at some examples.

This plot graphs WinterRain on the x-axis, and wine price on the y-axis.

By visually inspecting this plot, we can see that it looks like there's a slight positive linear relationship between these two variables.

It turns out that the correlation between WinterRain and wine price is 0.14, which corresponds to a slightly positive linear relationship, as we saw visually.

This plot graphs HarvestRain on the x-axis, and AGST on the y-axis.

It's hard to visually see a positive or negative linear trend in this data.

It turns out that the correlation is equal to negative 0.06, which is very close to 0, and corresponds to very little linear relationship.

This plot shows the age of wine compared to the population of France.

It looks like there's a very strong negative linear relationship between these two variables.

It turns out that the correlation is equal to -0.99, which is very close to -1, the smallest a correlation can be.

Let's compute some correlations in R.

We can compute the correlation between a pair of variables by using the cor function.

Let's compute the correlation between WinterRain and price.

We start by typing the name of the function cor, then the name of the first variable-- wine$WinterRain, followed by a comma, and then the name of the second variable.

If we hit Enter, we see that the correlation here between WinterRain and price is 0.1366505.

We can also compute the correlation between two different variables-- let's say Age and FrancePopulation.

Again, we use the function cor, and then type the names of the two variables.

If we hit Enter, we see that the correlation between Age and FrancePopulation is about -0.99.

We can also compute the correlation between all variables in our data set by using the same function, cor, and typing the name of the data set, wine.

Here, our output shows us a lot of numbers, and the rows are labeled by the variable names, as well as the columns.

To find the correlation between two variables, we find one variable name on the row, and then go to the column labeled by the other variable name.

So for example, if we want to find the correlation between AGST and price, we can go down to the row labeled price, and then through that row across to the column, labeled AGST Here, we find that the correlation is about 0.6595.

So we have confirmed that the correlation between Age and FrancePopulation is very high, so we do have multi-colinearity in our model.

Note that multi-colinearity refers to the situation where two independent variables are highly correlated.

A high correlation between an independent variable and the dependent variable, like the correlation between AGST and price, is a good thing, since we're trying to predict the dependent variable using the independent variable.

Multi-colinearity only applies to the case where two independent variables are highly positive or negatively correlated.

Because of multi-colinearity, you always want to remove the insignificant variables one at a time.

Let's see what would've happened if we had removed both Age and FrancePopulation at the same time.

We'll call this model, model5, and use the lm function to predict price using only AGST, HarvestRain, and WinterRain.

Again, we'll use the data set, wine.

If we look at the summary of our model, model5, we can see that AGST, HarvestRain, and WinterRain are all fairly significant, but our multiple R-squared dropped to 0.75.

In our previous model-- model4-- our R-squared was about 0.83.

So if we had removed both Age and FrancePopulation at the same time, we would have lost a significant variable, Age, and the R-squared of our model would have been lower.

Why did we keep FrancePopulation, and remove Age instead?

Well, we expect Age to be significant.

Older wines are typically more expensive.

Since the population of France steadily increases with the year, this captures the same effect as Age, but is less interpretable in our model.

Multi-colinearity reminds us that coefficients are only interpretable in the presence of other variables being used.

High correlations can even cause coefficients to have the wrong sign.

We'll see this in the next lecture.

So we fixed the multi-colinearity problem between Age and FrancePopulation.

Do we have any other highly correlated independent variables?

If you look back at our correlation matrix, you can see that we don't.

Since all of our other remaining variables are also significant, we'll stick with model4 as our model for the rest of this lecture.