[SQUEAKING] [RUSTLING] [CLICKING]

HELENA VAILLICROSA: OK. So now, we're just going to jump directly into how to use R for statistics. I'm going to make a very quick tutorial about how to apply linear models into the data, also general linear models and mixed models. I'm not going to cover the mathematical principles of that. I'm just going to show you how to apply those models with code.

So first of all, I'm just going to start over with all the objects here, just cleaning a little bit, so it's not distracting us. It's just going to charge again the database we've been working with based on nitrogen, and phosphorus, and potassium fertilization in yields. Oops, sorry.

OK. So this is how the structure of a linear model should look like. First of all, we create the name of the model, model number 1. We assign all of these into model number 1-- LM, which is linear model. And here, we're going to have our dependent variable, the y, which is the yield.

Just to remind you again what the database is about, here we have how much yield has been produced after the different treatments of fertilization. And we have the different experiments set in different blocks. And we are going to-- just to separate the dependent and independent variables, we put this sign in here. And then these are the different variables we want to explore, comma, and then data. We assign the data we want to explore, which is the NPK.

So let's just run this. And if we see what we have in the object, this is what we're going to see, which is not very informative. So what we need is to go to the summary of that model. That's going to provide us with the information that we would need to understand what's going on in the model.

So just click on Summary M1. And here, what we're going to see-- just make it bigger so you see properly. First of all, it's going to display the formula-- so what's the question that we've asked to the computer. Then we're going to have the distribution of the residuals.

And here, it goes the interesting part, which is the coefficients. This is the intercept, and then the different effects of the different variables-- nitrogen, phosphorus, and potassium. So when these are 1, so when there is fertilization, we see that there is an increase, because it's a positive value, into the baseline. With phosphorus, there's a decrease. And with potassium, there's a decrease as well.

And just to see if those different treatments are significant or not, based on 0.05 for your value, we have the different-- the p values here. And when whatever is significant, we have one asterisk. If it is very significant, we have three. And if it's marginal, we have this dot here. And if there's nothing, it means that it's not significant enough.

Also some interesting features is the adjusted R squared. Just as a reminder, the R square goes between 0 and 1. In this case, it will be 0.23, or like 23%. And also, the p-value of the whole model is displayed right here. I'll just put this back.

Something that's also very useful is to explore how the residuals look like to see if we have an appropriate distribution of the residuals, if it follows a normal distribution. So I create a histogram of the residuals of the model 1. So that stands for residuals comma. And previously, I mentioned that you can change the number of bars that are displayed in your histogram. In this case, I want 10. Here it is. I could shorten that number. And now, you will see that I have less bars and so forth.

Something else you could do is a box plot. I want to see what happens when I fertilize with nitrogen. So I can create a box plot just like that. Yield and nitrogen-- so I'm just going to focus on nitrogen right now. And again, the data that we want to explore, so just click on it, and we see that the ones that are not fertilized are lower and the ones that are fertilized have higher yield.

And we can save that summary that we created because it's where all the important data lives. So we have here this result right here. So we can consult the different parts that we just consulted before. So everything is stored right here. And we could even call individually the ones that-- let's say I just want to store the coefficients. So I'll go to sum, like the dollar. And I'll go to coefficients. And here, I'm just going to see the coefficients, which are the same as before.

OK. Now, it's the turn of the GLM, the generalized linear models. They are quite similar to linear models. But one of the advantages with GLMs is that we can provide a distribution of the data, in case the data wouldn't be normal. We could apply family links, like so to say if we want to have a gamma distribution, then we can add that to the model, so we don't have normal problems.

And so with the structure, as you can see, it's mostly the same as the linear models, just that we add that as a GLM, same y variable and function of these different x variables, just go for the same. If we wanted to apply the family, just-- oh, let's just go back to the Help thing, just see how GLM works.

You see here, it contemplates the possibility of add a family. In this case, by default, it's Gaussian. So whatever you see here initially is what is displayed by default. But if you wanted to change this Gaussian to something else, then you could do it. So if you scroll down, you'll see the information about the family. And if you click on Family for Details, here you see the different possibilities you have.

So if you want to apply a binomial distribution, gamma, and so forth, this is what you should include in your model. So I'm just going to copy that command, go right here. Let's see. That's the function I want. I have to put family first and go ahead. And now, it follows my gamma distribution. I'm not saying that's mathematically correct or not. I'm just focusing on the commands for you to be able to do it in your code.

OK. So now that we've reached this point, I'll challenge you with this exercise, asking the question, are there any differences in yield due to its block? What is the block? Just a quick reminder, that's the block. So you can answer the question both using linear model or generalized linear model, up to you.

All right, let's just answer the exercise. To do so, I'm just going to create a model. It's going to be called M2. And the question we had-- I'm going to use linear model to answer that. The question was, are there any differences ir yield due to its block? So let's just see yield as a variable. That's a y variable. And we want to see if there's any difference due to block. And then data, it's NPK. And that should work.

Let's see what we have in the model. Again, I remind you that Summary is where we want to go in order to see what's going on in the model. OK. Let's see.

Well, I'd say, not really. We don't have a lot of significance here. There's only one that's marginally significant. This block number 3 might be a little bit different than the other blocks. The p-value of the whole model is not very good. So I would say that we don't have differences due to the block.

OK. So before moving forward to the mixed models, I would like to introduce them a little bit, just to go through why are they important and how can we benefit from linear mixed models. So let's imagine that we are working in our experiment. In this case, we are applying the experiment that we have in the database, the NPK database.

There are distributions based on blocks. So what are those blocks? So let's just imagine that this is our working space. This is our lab. This is where we are performing our experiments. But just because the distribution of the lab is like this, we have one side where there are windows, so where the sun could get into our experiment. And there is the other side of the lab where there's no windows.

So I would anticipate at the beginning that maybe the plants that have more light-- so the ones that are closer to the window-- might grow more. But this is something that I'm not interested in exploring in my experiment. So to do that, that's where I introduce blocks. So I would record each of these plants that I'm growing into a different block. So in this case, I would have three blocks. And these are distributed based on how close they are from the windows.

So this applies also into other sort of environments. So if we move to this other design, we see that this could apply in when you go to the field. That represents a forest. If you have a forest in the top of the hill, that might be differently affected by the forest that's in the bottom of the hill. So this could be block number 1. And this could be block number 2.

And we could use blocks for everything that comes into our mind. Let's say that we are exploring a variable and we assume that if our water river is close to a city or it's far from a city, that could be important and that could affect our experiment. So here is where we think about using linear mixed models to take away the error or take away the variability that could come from things that are outside of our control.

So we don't have any interest in studying the differences in top and bottom of the hill. And we don't have any interest in studying the differences between the plants that are close to the window and the ones that are out of the window. So linear mixed models are a good way to get rid of that variability and just put the emphasis on the experiment that we are currently running.

After the quick explanation about the linear mixed models, let's just go ahead with the code again. I'm just going to get rid of this because it might be distracting. So linear models and generalized linear models are included in the R functions. But for mixed models, we have to install and charge extra packages.

In this case, I'm going to use the NLME package and also the lmerTest. So I already have them installed. But feel free to install them first, and then charge them in your session. In this case, instead of LM or GLM, we have the LME.

This is a linear mixed model. And our variable is still the yield-- so how much yield is produced after the fertilization of nitrogen, phosphorus, and potassium. But now, we are adding this random factor. This part of the model is what we call the fixed part, the fixed variables.

But the ones that we are going to include in this section are the random variables, which are the ones that we are not particularly interested in understanding, but we want to control so they don't add noise into our results. So in this case, the random one is the block, which, based on the explanation of the linear mixed models, are the different tables, the different blocks, the data.

And here, I added something new that could be used in every model, which is how the model faces the N/As. The N/As are empty values where R doesn't have information. It's like a hole in the database. So what happens when R faces a hole in the data? In this case, I'm just saying just omit that data and just keep going.

You could do an N/A omit or N/A fail, so if you want the code to fail if it encounters an N/A. In this case, I'm just omitting that, so just keep going and give me a result. I'm just going to run that and see how the residuals look like.

Again, you can change the number of bars right here. And I'm going to see what's the result of the model. In this case, this package doesn't offer us the symbols next to the variables to see if they are significant or not. But we can still read the p-value. But it also gives us the value of the intercept.

In this case, we have more significant results because of the controlling of the noise in the blocks. So controlling that part of the variability allows us to better focus on the effect of the fertilization. Here, we would see that fertilization with nitrogen is significantly affecting our results. It is increasing the yield by 5. And also, potassium is significantly affecting our yield production, but, in this case, is reducing the yield.

Also, what we can do is again to store this information into an object. Just I named it A. You consult whatever is inside the summary. In this case, I want to see the information about the fixed variables, which are these ones here. And then I can also save this information into a file, in case I want. Just type Save and the object you want to save, comma, and where do you want to store it.

Here, I'm just going to use the R format that I told you about. But if I wanted to store it, let's say in CSV, just write CSV, again the object you want to store, and the route where you want to put it in your computer.