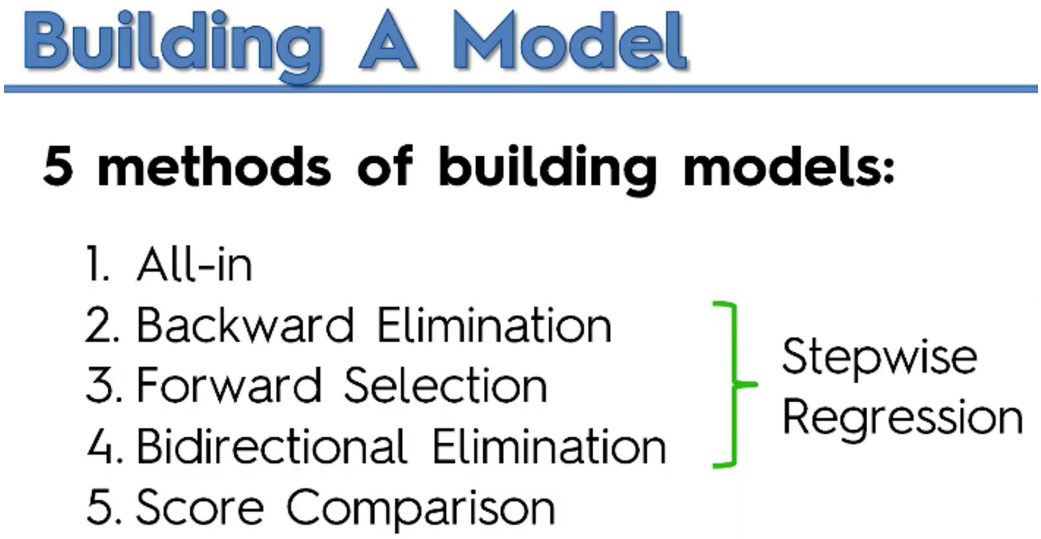
Step 5: Multiple Linear Regression Intuition

Sunday, April 10, 2022 1:58 PM

Building a model



1. "All-in" - cases"

"All-in" – cases:

- Prior knowledge; OR
- You have to; OR
- Preparing for Backward Elimination

2. Backward Elimination

Step 1: Select a significance level to stay in the model (e.g. SL = 0.05)
=> At the beginning, we need to select(decide) the significance level.

Step 2: Fit the full model with all possible predictors
=>We fit the full model of all possible predictors. So, we kind of do that "All-in" approach which we just talked about. And we put all the variables into our model.

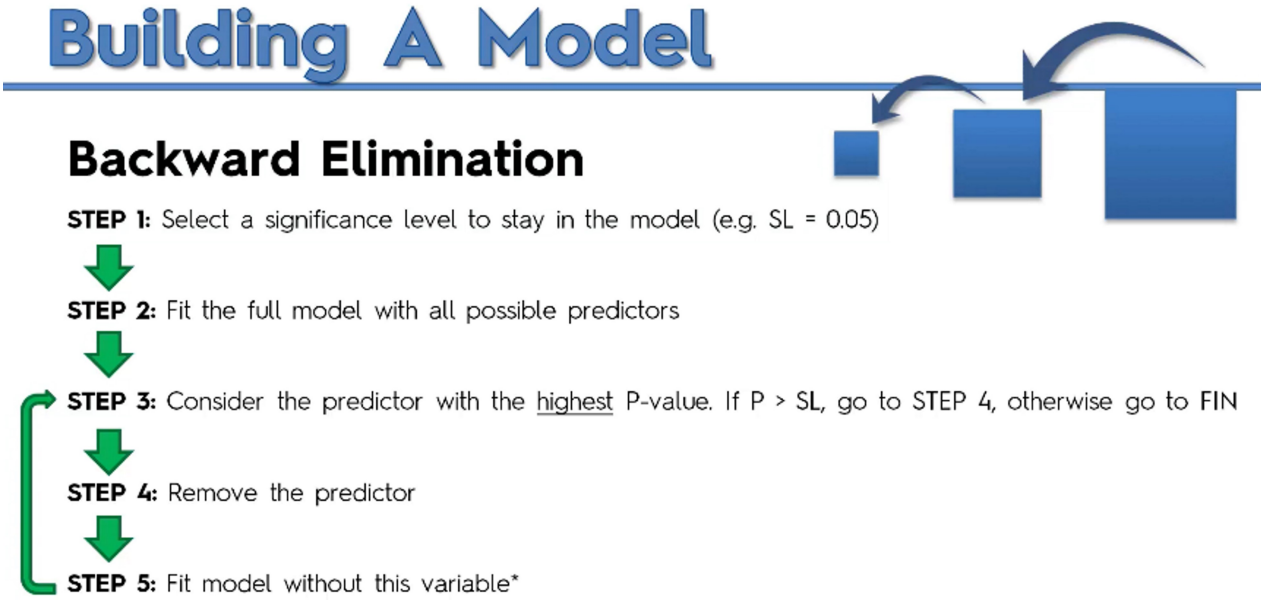
Step 3:(Now we are going start getting rid of them) Consider the predictor with the highest P-values. If P > SL, go to STEP 4, otherwise go to FIN(finish)
=>We consider the predictor with the highest P-value. If P > SL(Significance Level) then we will go to Step 4.

Step 4: Remove the predictor
=>We have to remove that predictor. Basically, we remove the variable that has the highest P-value.

Step 5: Fit the model without this variable*
=>From Step 4, we fit the model without this variable. There is a star here because to remind myself that if we just remove the variable; obviously, we can't just say that, "Okay! Now I have got the new model." We have to actually re-fit the model to re-create the model; rebuild it with the fewer number of variables.

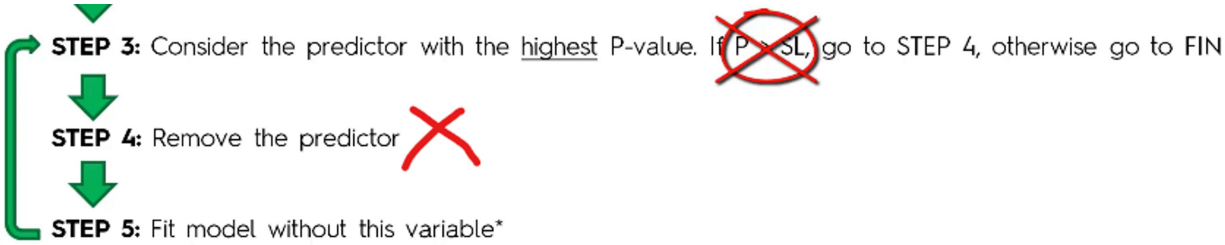
For instance: If we had 100 variables and we remove one of them, and we have 99 now. Well, we have to rebuild it so, the

coefficient are gonna be different, the constant is going to be different and we need to perform that step because once we remove a variable, it affects all the other variables in our whole regression.



So after Step 5, we go back to Step 3, once again we look for the variable with the highest P-value in our new model; we take it out("we remove it"); fit the model again with one less variable and so on. We keep doing that until we come to point where even the variable of highest P-value is less than the Significance level.

So, if that condition P > SL is not satisfied(correct), then we don't go to Step 4 anymore; we go to FIN. In this case, FIN is the 'finish' . Our model is ready. As soon as all of the variable that we have left in our model and their P-values are less than their Significance Level (P < SL) then your model is prepared.



So, that's how the Backward Elimination Method work

3. Forward Selection

Step 1: Select a significance level to enter the model (e.g. SL = 0.05)

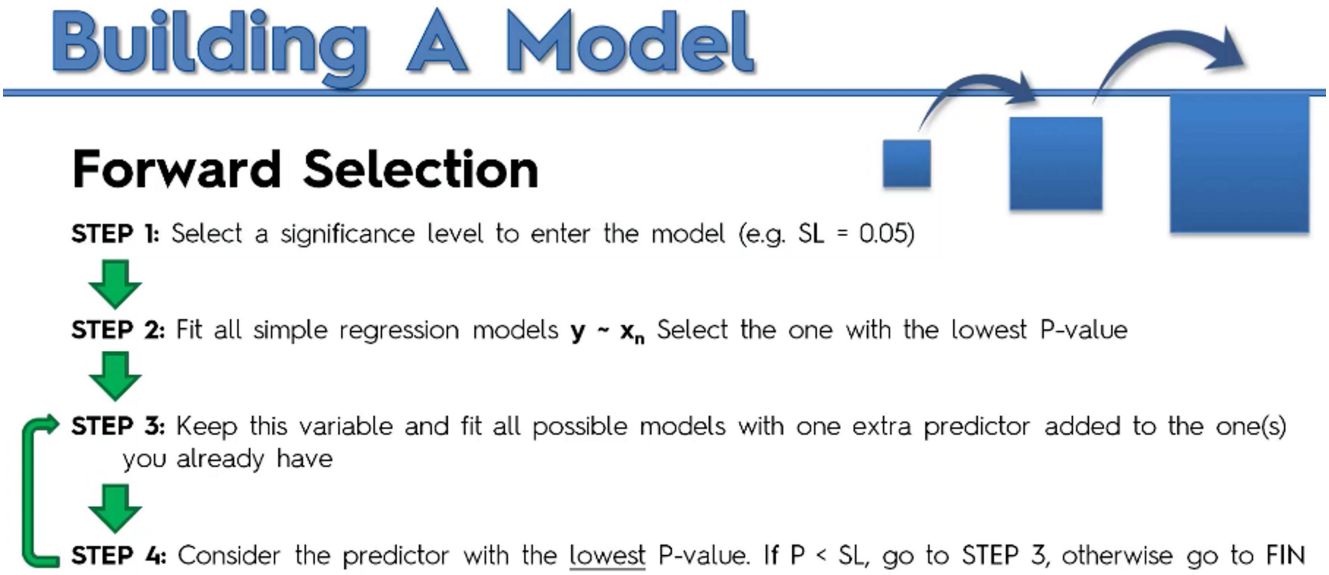
=>We start with Step 1 by selecting a significance level to enter the model and in this case, we are going to select 5%(SL = 0.05).

Step 2: Fit all simple regression models $y \sim x_n$ Select the one with the lowest P-value

=>We fit all possible Simple Regression models. We take the dependent variable and we create a regression model with every single independent variable that we have. Out of all those model, we select the one which has the lowest P-value for the independent variable.

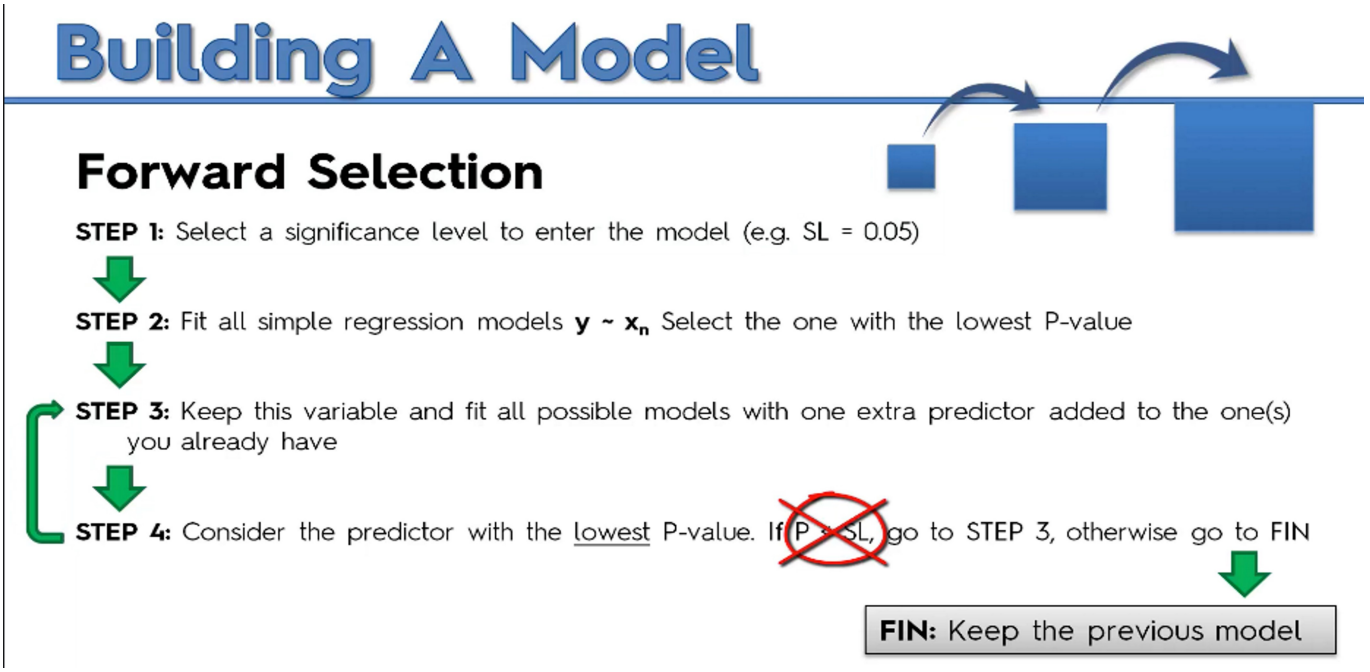
Step 3: Keep this variable and fit all possible models with one extra predictor added to the one(s) you already have

=>We keep this variable that we have just chosen and we fit all other possible models with one extra predictor added to the one we already have. We have selected a Simple Linear Regression with one variable.



Step 4: Consider the predictor with the lowest P-value. If P < SL, go to STEP 3, otherwise go to FIN

=>Out of all these possible 2 variable regressions we consider the one where the new variable that we added had the lowest P-value. So, if that P-value is less than our significance level meaning that variables is a good one; it's a significant variable. Then we move back to STEP 3. It means that, now we have a regression with two variables and now we will add a third variable. We will try all possible variables that we have left as our third variable and then out of all those models with three variables we will go to STEP 4 and we'll select again the one with the lowest P-value for that third variable that we added. And we will keep doing that.



So basically we'll be keep growing the regression model but not just randomly; we will be actually selecting out of the all of the possible combinations every single time and growing it with 1 variable at a time. And then we will only stop when the variable that we have added; it has a P-value that is greater than our Significance Level. So when this condition P < SL is not true then we don't go to STEP 3, we finish the regression because the variable we just added is not significant anymore. And we also know that we have selected the one with the lowest P-value. So there is no other variable that we can add that its P-value will be less than SL.

4. Bidirectional Elimination

Step 1: Select significance level to enter and to stay in the model

e.g.: SLENTER = 0.05, SLSTAY = 0.05

=>Select a significance level to enter and significance level to stay. So we are going to select in both cases 5% but its up to us(we can select as per our requirement)

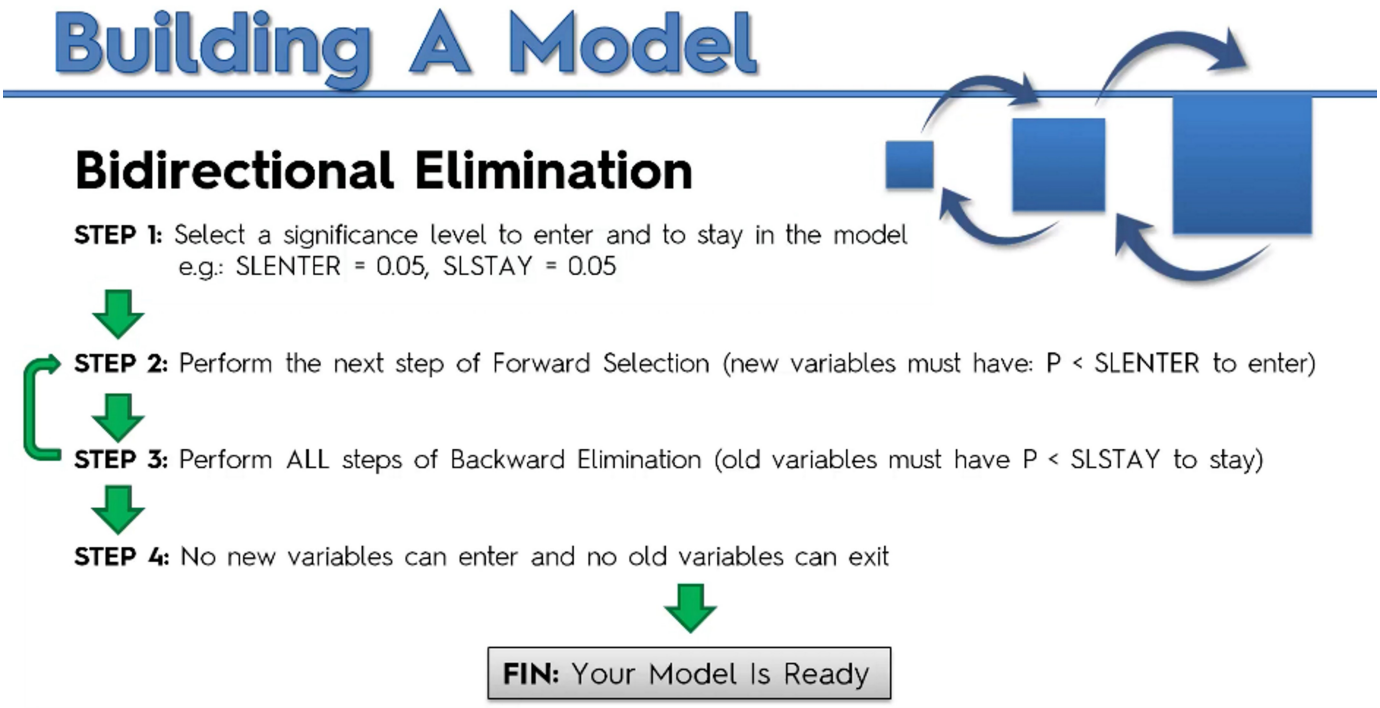
Step 2: Perform the next step of Forward Selection (new variable must have P < SLENTER to enter)

=>It means that the one we just discussed where new variables when they enter and in order to enter they have to be less than the SL to enter. So basically, add on a new variable based on the forward selection method.

Step 3: Perform ALL steps of Backward Elimination (old variable must have P < SLSTAY to stay)

=>So, now we have 2 variables. Start getting rid of them and see if you can get rid of any of them and then move back to STEP 2.

Step 4: No new variables can enter and no old variables can exit



4. All Possible Models (The most resource consuming approach)

Step 1: Select a criteria of goodness of fit (e.g. Akaike criterion)

Step 2: Construct All Possible Regression Models: 2^N-1 total combinations

=>You construct all possible Regression Models. So if you had N variables then it will be 2^N-1 total combinations of that variable and that's exactly how many models there can possibly be.

Step 3: Select the one of the best criterion

=>We select the one of these model with the best criterion that we're looking at.

Then our model is ready.

Building A Model

All Possible Models

STEP 1: Select a criterion of goodness of fit (e.g. Akaike criterion)

STEP 2: Construct All Possible Regression Models: 2^N-1 total combinations

STEP 3: Select the one with the best criterion

FIN: Your Model Is Ready

Example:
10 columns means
1,023 models

Note: The one we are going to be looking at in these tutorials in order to get our head around; how to build model step by step and get some practice is the Backward Elimination Process because it is the fastest one out of all of them and you will still get to see exactly how these step by step method works.