Summary of Different Ways to Run Multiple Regression

**Ways to choose your model**

As I explained in the videos, the problem is that having too few predictors leads to poor prediction, but having too many can cause overfitting, non-convergence, difficulty in interpretation and explaining results to other people. **So how do you choose the predictors?**

It’s always a good idea to begin by reviewing the relevant literature and expert knowledge, but that will only get you so far. It may tell you that you should include age, gender and perhaps a few other things, but it’s very likely that you’ll have other variables in your data set that are worth trying. One option is to enter and keep them all in your model, whatever the p values. This is a good idea if you don’t have too many and / or you can use a priori knowledge for them all. (but beware when this prior knowledge comes from studies using poor methods)

**A variant of this is backwards elimination, where you then drop the non-significant variables.** This works OK in some circumstances, but you need to check for correlation between variables. The best way to do that is by inspecting the odds ratios for the predictors that you are keeping – first with all variables in the model and second when you drop some. If the odds ratios change noticeably when you drop some, then you’ll need to add back at least one of the dropped ones.

Forward selection involves starting with an empty model and trying variables one at a time. Stepwise selection involves a mixture of forwards and backwards. **Both of these should be avoided**. Similarly, “all-possible-regressions", which literally tests all possible subsets of the set of potential independent variables, is to be avoided. It might sound like it, but it’s not in fact guaranteed to give you the best model for your data.

Machine learning approaches are increasingly popular, but are complex and out of our scope.

**Training and testing data sets**

It’s good practice to split your data set if possible into a training and a testing data set and apply the training-set model to the test-set data. If you get very different answers, then rethink the complexity of training-set model and repeat. This is standard practice with machine learning because of the risk of overfitting and "overtraining", in which the algorithm fits the data too closely so it's unable to distinguish between signal and noise in the data set and so performs badly on a new data set. However, it's also strongly recommended with statistical models where possible. Another related technique is called k-fold cross-validation, which is often used when you have limited data.

With small data sets, splitting into a training and a testing set is not advisable. The data sets we are using in these statistics for public health courses are considered small, which is why I haven’t suggested splitting them. If you're keen to go beyond this introductory course, then I suggest learning about cross-validation. If you're taking our Global Master's degree in Public Health, then it will be covered in the Advanced Statistics specialisation, which will also cover things like LASSO and elastic nets that are relevant to model selection.