1. I want to predict how well 6 year-olds are going to do in their final school exams. Using the following variables am I likely under-fitting, fitting well or over-fitting? Postcode, gender, reading level, score in maths test, date of birth, family income.

Could be any of these.

1. If I have two models, one with an AIC score of 34,902 and the other with an AIC score of 33,559 which model should I use?

I would say if this was the only thing we had to go on we would go with the lower one, but we should compare other measures to confirm (BIC, aR2 etc)

1. I have two models, the first with: r-squared: 0.44, adjusted r-squared: 0.43. The second with: r-squared: 0.47, adjusted r-squared: 0.41. Which one should I use?

Either could be used, I would probably lean towards the first (r2 0.44, aR2 0.43) as it shows a “well fit” model, where-as the second, appears to be “over-fit” as the aR2 is quite a bit lower than the R2

1. I have a model with the following errors: RMSE error on test set: 10.3, RMSE error on training data: 10.4. Do you think this model is over-fitting?

Impossible to say, the RMSE’s across the test and train sets look good, so I would not use this information to answer this question

1. How does k-fold validation work?

It takes the data set and separates it into however many “folds” you wish, and then holds each of the folds as a test set in order to train with the remaining data enabling you to test your model across numerous different sets.

1. What is a validation set? When do you need one?

A validation set is a subset of your data that you do not use as part of your model analysis, so that you can check the viability of your model against data that it has not had access to.

1. Describe how backwards selection works.

Backward selection is where you add every possible correlation in your data set to your model, and one by one take out the least “relevant” predictors until you are satisfied with your model.

1. Describe how best subset selection works.

It searches for every possible combination of predictors at each size of model in order to find the best fitting model according to your chosen method of measure (r2, AIC, BIC etc)