MATHS 7107 Data Taming Tutorial

Multiple Regression

Load the data

First, let's make sure the tidyverse package is loaded and we'll look at the population data again this week.

Outline

This week we're putting together a lot of what we have done so far.

Our exercise this week will be selecting an appropriate model to predict annual population growth in the population data. To limit things a little, we will not consider interaction terms and the only predictors we will consider are:

1. `med_age_all` - the median age of all residents.
2. `med_age_male` - the median age of male residents.
3. `med_age_female` - the median age of female residents.
4. `ed_index_2015` - the United Nations index of educational development, as at 2015.
5. `continent` - the continent the country is in.
6. `inequality` - the GINI coefficient measuring income inequality in the country.
7. `per_urban` - the percentage of the population living in urban centres.

Using these predictors (... or not), we want to find the best model to predict annual population growth. We can do this in a step-by-step process by choosing which predictors are significant. You can do this in one of two ways:

1. Start with an empty model, and try all possible predictors alone. Add the predictor with the smallest p-value. Using this model, try all possible remaining predictors, adding the predictor with the smallest p-value. Continue until there are no remaining significant predictors to possibly add.
2. Start with a 'full' model, with all possible predictors. Remove the predictor with the highest p-value. Fit the model without this predictor. Remove the predictor with the highest p-value. Continue until all predictors left in the model are significant. Today we will use method number 2.

Visualise the data

Produce the appropriate plots to compare each predictor variable with the response variable. Consider the relationship in each plot.

'Backwards' model selection

First, let's start by building the full linear model.

Then you will need to look at the p-values. Remember the p-value for the categorical variable continent will be given by an ANOVA (using the `anova()` function), while the other p- values will be given by the model's summary (using the `summary()` function):

Consider the p-values. Can any of the predictor variables be removed? If so, which one? Remove it.

Repeat this process until all variables in your model are significant.

Prediction under the model

Now on to prediction. Suppose we wish to project the population growth of a particular country with:

- median male age of 38,
- median female age of 41,
- GINI coefficient of 28,
- 90% of its population living in urban centres, and
- in Europe.

How could we do this in R?

First create a `tibble` with our new data.

Since we're interested in a particular country, rather than the average country of this character, we should make a prediction interval. This is because:

- A confidence interval is for an average country of this character.
- A point prediction doesn't consider how accurate our prediction is, and it's important to know how we might be wrong.

Using R, calculate the prediction interval based on the above information. Interpret this interval in context.

Are there any negative numbers in the interval? If yes, are they reasonable?