MATHS 7107
Data Taming
Practical 7

Multiple Linear Regression

1 Preliminaries

- Set up a project in RStudio
- Download the `population.csv` file to a `data` subdirectory of your project directory
- Now load the `tidyverse`
- Read in the population data. You can call it whatever you like, but in the commands below we have assumed the data will be called `population`.

1.1 Aim of today's prac

Our exercise this week will be selecting an appropriate model to predict annual population growth `pop_growth_2015_20` in the `population.csv` 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 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. This is **Backwards Stepwise Regression**

Today we will use method number 2, as was done in Module 7, page 17.

1.2 Eat your veggies!

Make sure you type all these commands **BY HAND!** (Don't just copy and paste.) You will learn a lot faster if you type the code!!

2 Visualise the data

The best place to start is by making some graphs comparing the variables we're interested in. Make sure you put the predictors on the horizontal axis!

So let's get started. Produce graphs comparing our outcome variable to the following predictors, and also describe the relationships that you see in the plots.

Questions:

1. The median age of all residents
2. The median age of male residents
3. The median age of female residents
4. Educational development
5. The continent in which the country is located
6. The GINI coefficient
7. The urbanisation rate

2.1 'Backwards' model selection

So now let's build the full linear model, with all the predictors in it. Then we will need to look at the p-values. Remember the p-value for the variables should be obtained using the `Anova()` function.

Questions:

8. Build the linear model using the `lm()` command.
9. Use `summary()` on your model.
10. What has happened to our categorical variable in the model? Which is the reference level?
11. Use `Anova()` on your model.
12. What are the differences between the `summary()` and `Anova()` outputs?

Once you've got your `Anova()` output, look at the p-values for the various coefficients. Remember that these p-values are telling us how significantly different the coefficients are **from zero**. For this prac, we'll stick to the conventional 95% significance level, ie.

we will declare the coefficient significantly different from zero if $p < 0.05$.

Questions:

13. Should any predictor variable be removed? If so, which one?
14. If there was a removable one, then remove it and repeat until all the variables in your model are significant at the 95% level.

3 Prediction under the model

Now that we have a significant model, let's use it for predicition. Suppose we wish to predict 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.

We will use the `predict()` function. But remember, that we need to give the function our new predictors as a tibble (dataframe). So let's make that first.

```
new_country = tibble(
   med_age_male = 38,
   med_age_female = 41,
   inequality = 28,
   per_urban = 90,
   continent = "Europe"
)
```

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 point prediction doesn't consider how accurate our prediction is, and it's important to know how we might be wrong;
- and we can't use a **confidence interval** because that's for the average country with these predictor values.

Questions:

15. Calculate the prediction and prediction interval for our `new_country` at the 90% level and at the 99% level. Which interval is bigger? Is that sensible?
16. Interpret the output in the context of the aim of today's prac.
17. Are there any negative numbers in the interval? If yes, are they reasonable?