MATHS 7107
Data Taming
Practical 6

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 solutions we have assumed the data will be called `population`.

2 Building a linear model

Looking at the population data, let's try to build a model predicting population growth, using the residents' mean number of years of schooling.

2.1 Aim of today's prac

In this prac, we'll try to answer the following questions about our linear model:

1. What is the slope and the intercept, and what do they mean in context?
2. Is there a significant relationship between mean years of schooling in a country, and its annual population growth rate between 2015 and 2020?
3. What is the expected population growth of a country in which the mean number of years' schooling is 5 years? What about for a country with mean years' schooling of 12 years?
4. How could we interpret a prediction interval for the annual population growth of a country with mean number of years' schooling of 5 years?
5. Are the assumptions of the model justified?

2.2 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!!

2.3 Start by looking at the data

Before we do anything else, let's see what the data looks like.

Questions:

1. Make a scatterplot of `pop_growth_2015_20` against `mean_years_school_2015`. Make sure you put the explanatory variable on the horizontal axis!
2. Add a (straight) line of best fit to your graph.

2.4 Building the model

Now, let's build the model and look at the output:

```
lm_pop <- lm(pop_growth_2015_20 ~ mean_years_school_2015,
             data = population)
summary(lm_pop)
```

```
##
## Call:
## lm(formula = pop_growth_2015_20 ~ mean_years_school_2015, data = population)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5615 -0.5624 -0.0651  0.4883  3.2092
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.3207     0.1706   19.46   <2e-16 ***
## mean_years_school_2015    -0.2415     0.0189  -12.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7757 on 173 degrees of freedom
##   (92 observations deleted due to missingness)
## Multiple R-squared:  0.4855, Adjusted R-squared:  0.4825
## F-statistic: 163.3 on 1 and 173 DF,  p-value: < 2.2e-16
```

Questions:

3. What is the value of the intercept β_0? Interpret this value in context.
4. What is the value of the slope β_1? Interpret this value in context.
5. What is the equation of the linear regression line?
6. Determine if this model is statistically significant. (By convention, this means significant at the 95% level.)

3 Prediction under the model

3.1 Point estimate

Now on to prediction, and we'll use the `predict()` function. This function needs the predictors in the form of a tibble (or dataframe), so first we'll create a tibble with our new data, then we can predict population growth for the two countries.

```
new_countries <- tibble(mean_years_school_2015 = c(5, 12))
predict(lm_pop, new_countries)
```

```
##         1         2
## 2.1133111 0.4228968
```

Questions:

7. For the country with 5 years average schooling, what is the expected annual population growth?
8. For the country with 12 years average schooling, what is the expected annual population growth?

3.2 Prediction interval

And finally, a **prediction interval**, for a country with a mean of five years of schooling. We'll choose the conventional 95% level.

```
predict(lm_pop, new_countries, interval = "prediction" , level = 0.95)
```

```
        fit        lwr      upr
1 2.1133111  0.5725153 3.654107
2 0.4228968 -1.1180399 1.963834
```

Question:

9. Interpret these prediction intervals in context.

3.3 Confidence interval

Here we are looking for a **confidence interval**, which is different to a **prediction interval**:

- prediction interval: gives an interval for a **particular element** to be sampled with the specified value of the predictor variable. In our example, this is an interval for a particular country where the population has a specific level of schooling.
- confidence interval: gives an interval for the **average element** in the population with the specified value of the predictor variable. In our example, this is an interval for the average country where the population has a specific level of schooling.

Here we will look for a confidence interval at the 99% level.

```
predict(lm_pop, new_countries, interval = "confidence" , level = 0.99)
```

```
        fit       lwr       upr
1 2.1133111 1.8839038 2.3427185
2 0.4228968 0.1918476 0.6539461
```

Notice that the predicted values are still the same — it is only the intervals that change.

Question:

10. Interpret these confidence intervals in context.

3.4 Assumption checking

Question:

11. What are the four assumptions of Linear Regression?

Questions:

12. Check if the assumption of linearity has been met, using:

```
plot(lm_pop, which = 1)
```

13. Check if the assumption of homoscedasticity (constant variance) has been met, using:

```
plot(lm_pop, which = 3)
```

14. Check if the assumption of normality has been met, using:

```
plot(lm_pop, which = 2)
```

15. Is the assumption of independence met? Try hard to think of something that would violate independence — there's always something that can mess up your data!

Question:

16. Test what happens if you run the following code:

```
plot(lm_pop)
```