MATHS 7107 Data Taming Practical

Linear Regression

Load the data

First, let's make sure the tidyverse is loaded and read in the population data.

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.

Then we'll answer a few questions:

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 it's 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?

First, let's build the model, and have a look at the output:

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

What is the value of the intercept β_0? Interpret this value in context

What is the value of the slope β_1? Interpret this value in context

What is the equation of the linear regression line?

Determine if this model is statistically significant

Prediction under the model

Point estimate

Now on to prediction. 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)
```

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

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

Prediction interval

And finally, a prediction interval, for a country with a mean of five years of schooling.

```
new_country <- tibble(mean_years_school_2015 = 5)
predict(lm_pop, new_country, interval = "prediction")
```

Interpret this prediction interval in context

Assumption checking

What are the four assumptions of Linear Regression?

First, let's check linearity:

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

Is the assumption of linearity met?

Now, let's check homoscedasticity:

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

Is the assumption of homoscedasticity met?

Now, let's check normality:

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

Is the assumption of normality met?

Is the assumption of independence met?

Test what happens if you run the following code:

```
plot(lm_pop)
```