MATHS 7107 Data Taming Practical Solutions

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)
```

```
##
## 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
```

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

The intercept is 3.3207. This means that for a country in which residents have an average of zero years of schooling, we expect the population to grow by 3.3207% per year.

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

The slope is -0.215, meaning that if a country increases schooling by one year, we expect its annual population growth to decrease by 0.2415%.

What is the equation of the linear regression line?

$$y = 3.3207 - 0.215x$$

OR

$$population\ growth = 3.3207 - 0.215 \times years\ of\ schooling$$

Determine if this model is statistically significant

Since our p-value is <2e-16, our 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)
```

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

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

We expect the country with 5 years average schooling to have annual population growth of 2.113%.

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

We expect the country with 12 years average schooling to have annual population growth of 0.423%.

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")
```

```
##        fit       lwr      upr
## 1 2.113311 0.5725153 3.654107
```

Interpret this prediction interval in context We can be 95% confident that a country with 5 years average schooling will have annual population growth between 0.5725153% and 3.654107%.

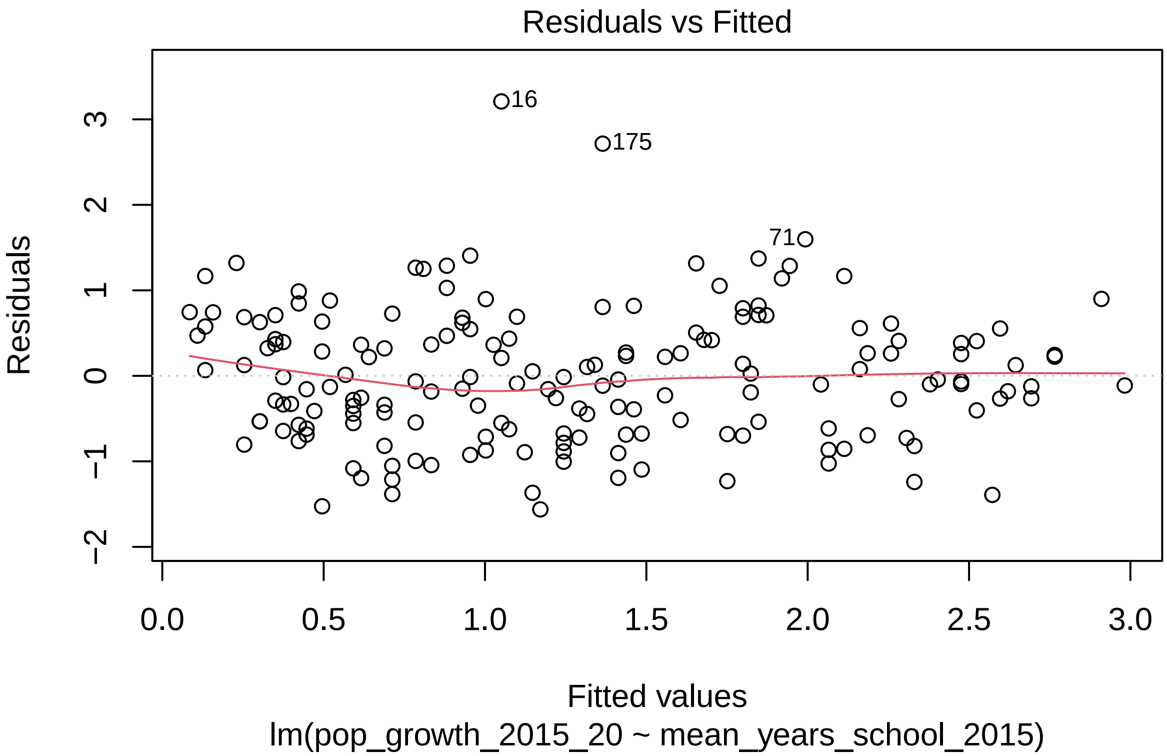
Assumption checking

What are the four assumptions of Linear Regression?

Our assumptions are linearity, homoscedasticity, normality and independence.

First, let's check linearity:

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

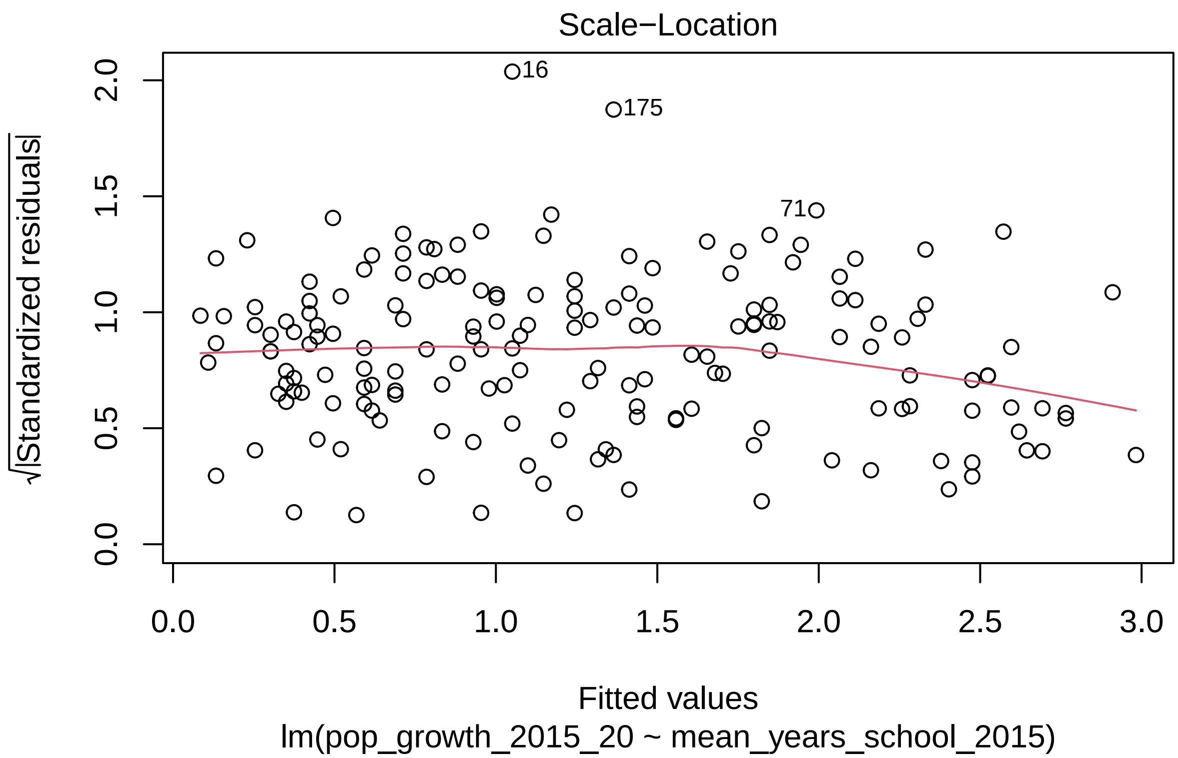


Is the assumption of linearity met?

This looks great; we see random scatter around zero and there is no strong trends as we move from left to right on the residuals vs fitted plot.

Now, let's check homoscedasticity:

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

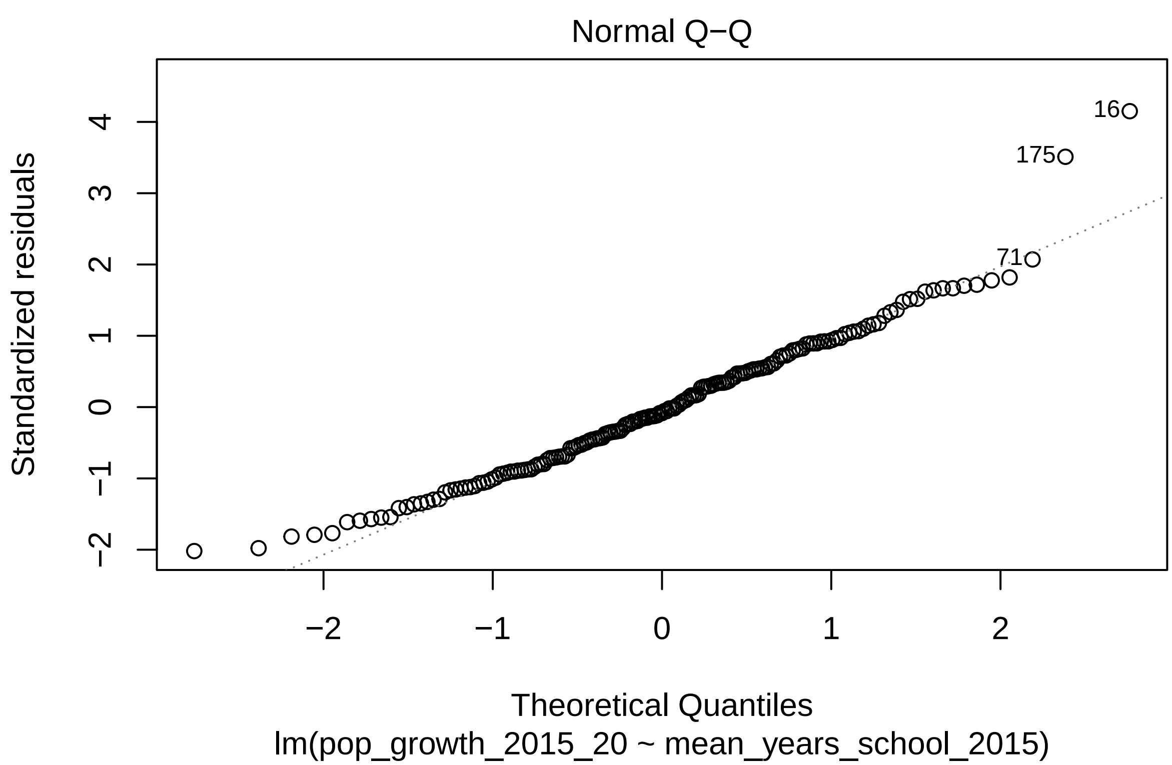


Is the assumption of homoscedasticity met?

This looks great; we see no change in vertical spread as we move from left to right on the scale-location plot.

Now, let's check normality:

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



Is the assumption of normality met?

This looks great; the points lie quite close to the normal quantile reference line.

Is the assumption of independence met?

Independence relies on the subjects being independent of each other; this is not necessarily true here, since, for example, if there's a mass migration from one country to another, the population growth of the two countries will be related.

Test what happens if you run the following code:

```
plot(lm_pop)
```