Tutorial 2 - Linear models (Solutions).

Christopher A Oldnall

Welcome to the second tutorial of the Regression and Simulation methods module. This is the next script in developing your skills in R, whilst learning about Regression. Throughout this notebook we will be solidfying the process of normal linear modelling and ensuring you can complete the analysis for all parts we have discussed in the tutorial.

Exercise 0

Throughout remember we will need tidyverse, go ahead and do this as your first task.

Your Answer:

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

This week we will be working with a data set which is all about plastic pollution. Plastic pollution is a major and growing problem, negatively affecting oceans and wildlife health. Our World in Data has a lot of great data at various levels including globally, per country, and over time. Here we focus on data from 2010.

Additionally, National Geographic ran a data visualization communication contest on plastic waste as seen here.

The dataset for this notebook can be found as a csv file. We do not load this like we did previously, as it is not a built in R data set. Instead we must use the 'read_csv' command. Depending on where you store the data you may have to change the file pathway.

```
plastic_waste <- read_csv("plastic-waste.csv")
```

You may have not considered it, but a fun note is that '.csv' stands for 'comma seperated values'. Other common file extensions include '.xlsx' for Excel spreadsheets and '.tsv' for 'tab seperated values'. You will have to change the import command depending on your data type otherwise it may not load it in correctly. We'll practice this in future weeks.

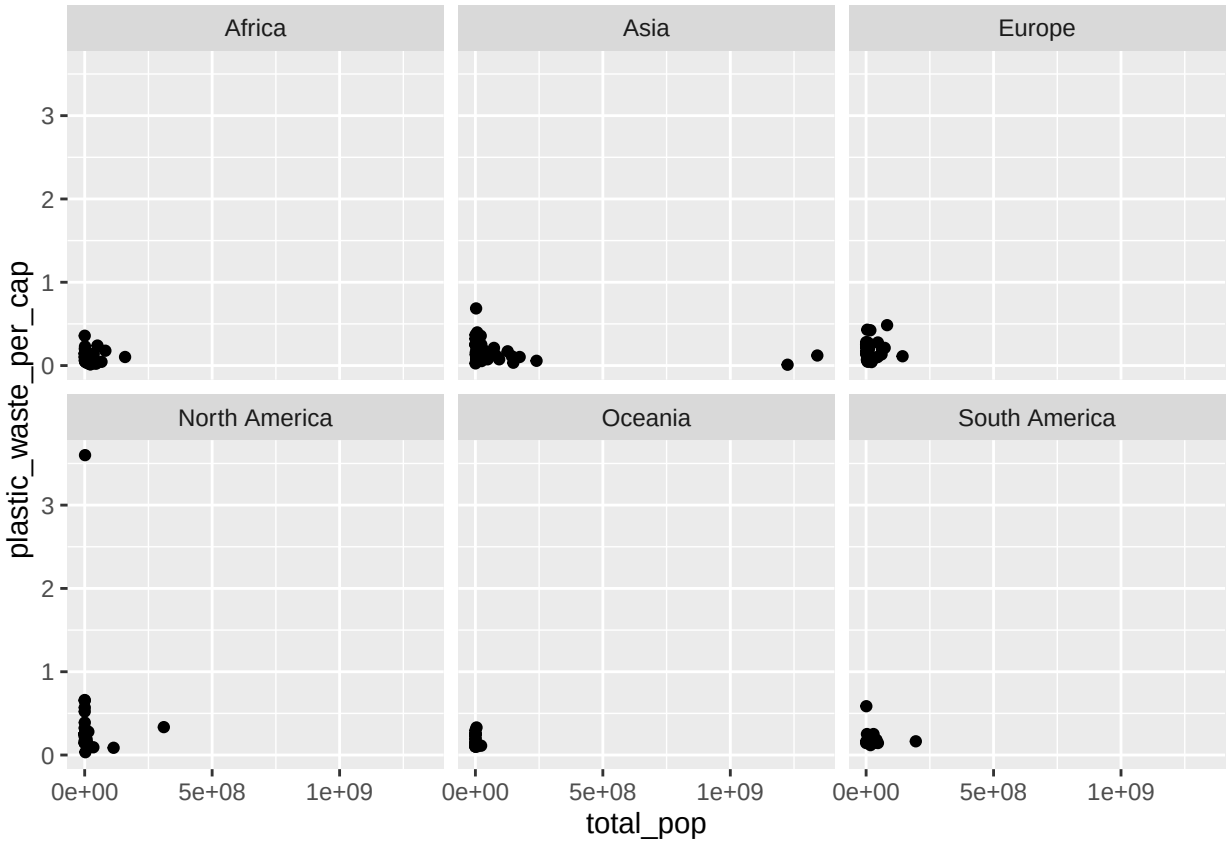
Exercise 1

Go ahead now and investigate this data. Write a paragraph to describe what it contains, how many observations there are, what the variables are and what types these variables are stored as. Additionally include a scatterplot of each countries total population (x) vs plastic waste per capita (y), grouped by continent. (Hint: you will need to use the command 'facet_wrap').

Your Answer: In this data set we see there are 240 records, with 10 different variables. 'code', 'entity' and 'continent' are strings/characters which represent different country codes, their names and which continent they sit on. Alongside this there is then a range of numerical variables including the 'year', 'gdp_per_cap', 'plastic_waste_per_cap', 'mismanaged_plastic_waste_per_cap', 'mismanaged_plastic_waste', 'coastal_pop', 'total_pop'. There are some missing data points. The data set is built to look at levels of plastic waste across different countries.

```
continent_scatterplot <- plastic_waste %>%
    ggplot(aes(x = total_pop, y = plastic_waste_per_cap)) +
    geom_point() +
    facet_wrap(~continent)

continent_scatterplot
```

```
## Warning: Removed 61 rows containing missing values (`geom_point()`).
```



Within much data, we will observe different trends depending on certain characteristics (also known as confounders). In this data, a potential confoudner could be continent. Therefore we peform **population stratification**. This is the process of dividing up our population in to groups to reduce any indirect effect induced by a potential confounder. To do this we can use the 'filter' command as follows:

```
North_America_Data <- plastic_waste %>%
    filter(continent == "North America")
```

Exercise 2

For the rest of this exercise we will work with the data for the 'Europe' continent only. Go ahead and filter for this, saving it as a new object. Once this has been done, again create a scatterplot, for plastic waste per capita against population, but consider that we might want to transform one of the scales. To this same plot use 'geom_smooth(method=lm)' to add a linear model line to the plot.

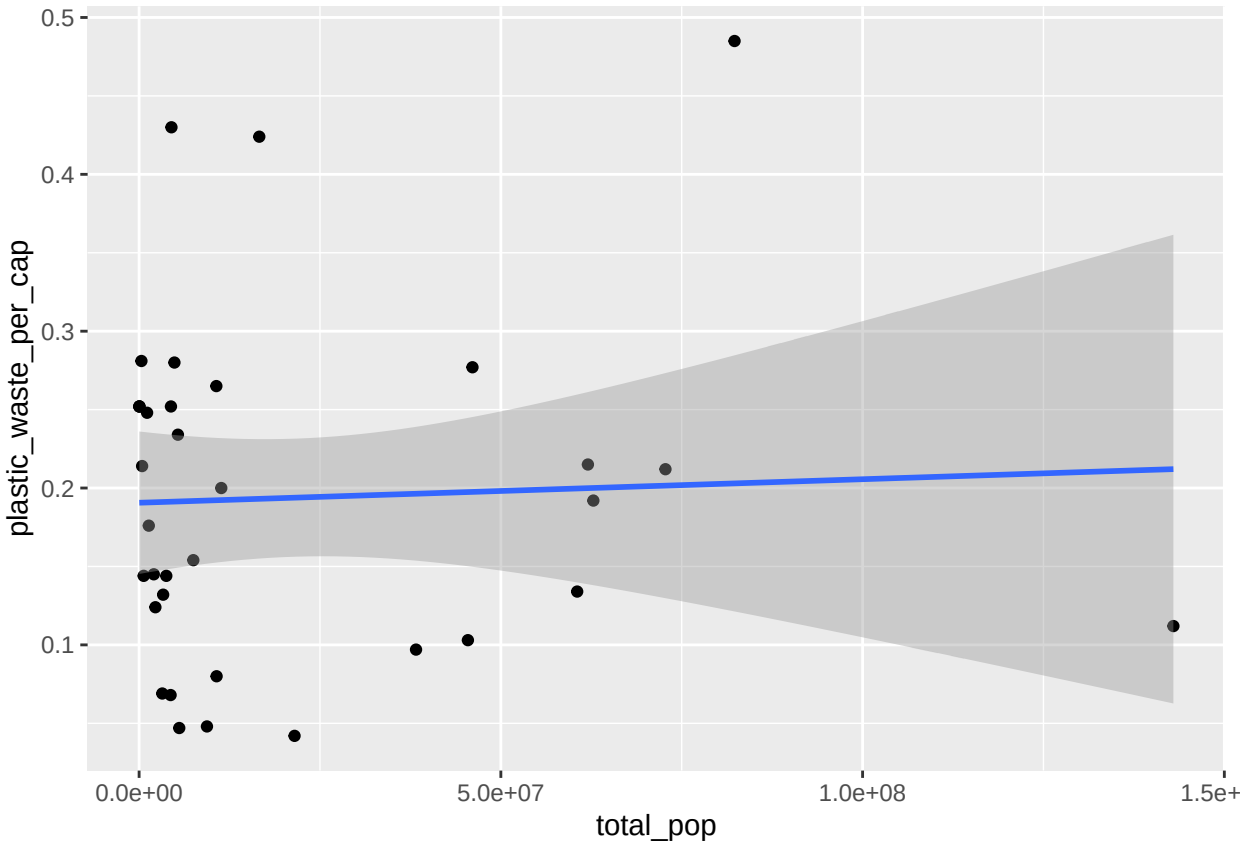
Your Answer:

```
Europe_Data <- plastic_waste %>%
    filter(continent == "Europe")

Europe_Scatterplot <- Europe_Data %>%
    ggplot(aes(x = total_pop, y = plastic_waste_per_cap)) +
    geom_point() +
    geom_smooth(method=lm)

Europe_Scatterplot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 18 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 18 rows containing missing values (`geom_point()`).
```



Having now visually inspected the graph, consider that there are multiple other variables in the data that might be contributing to the final outcome of plastic waste per capita.

Exercise 3

Formulate in words and equations the largest model you can which would make sense with the data you have. You can use '$ $' within markdown in a similar manner to LateX to write your model.

Your Answer: It would make sense for our model to rely on 'gdp_per_cap', 'coastal_pop' and 'total_pop'. The names/codes of countries are irrelvant. It could be argued that the mismanaged plastic waste variables could be relevant, however I would argue that they are a result of the plastic waste per capita and do not affect plastic waste per capita. Therefore our model is:

$$Y_{PWCP} = \beta_0 + \beta_1 X_{GDP} + \beta_2 X_{CP} + \beta_3 X_{TP} + \epsilon.$$

In R we can use the command 'lm' to construct our linear model and get an estimate of our parameters (through the least squares method). Note however this uses the '~' notation to write in what we want the model to be. We can read Y ~ X + Z as, 'explain Y by X and Z'.

Exercise 4

Put in to practice the linear model that you have formulated. Save the results in an object called 'Europe-Waste_LM'. Comment on what seems to be the strongest contributor to the model. [Hint: You will need to specify with lm, that you are giving it the formula when using tidyverse].

Your Answer:

```
EuropeWaste_LM <- plastic_waste %>%
    lm(formula = plastic_waste_per_cap ~ gdp_per_cap + coastal_pop + total_pop)

EuropeWaste_LM
```

```
##
## Call:
## lm(formula = plastic_waste_per_cap ~ gdp_per_cap + coastal_pop +
##     total_pop, data = .)
##
## Coefficients:
## (Intercept)  gdp_per_cap  coastal_pop    total_pop
##   1.488e-01    2.900e-06   -6.030e-10   -2.360e-12
```

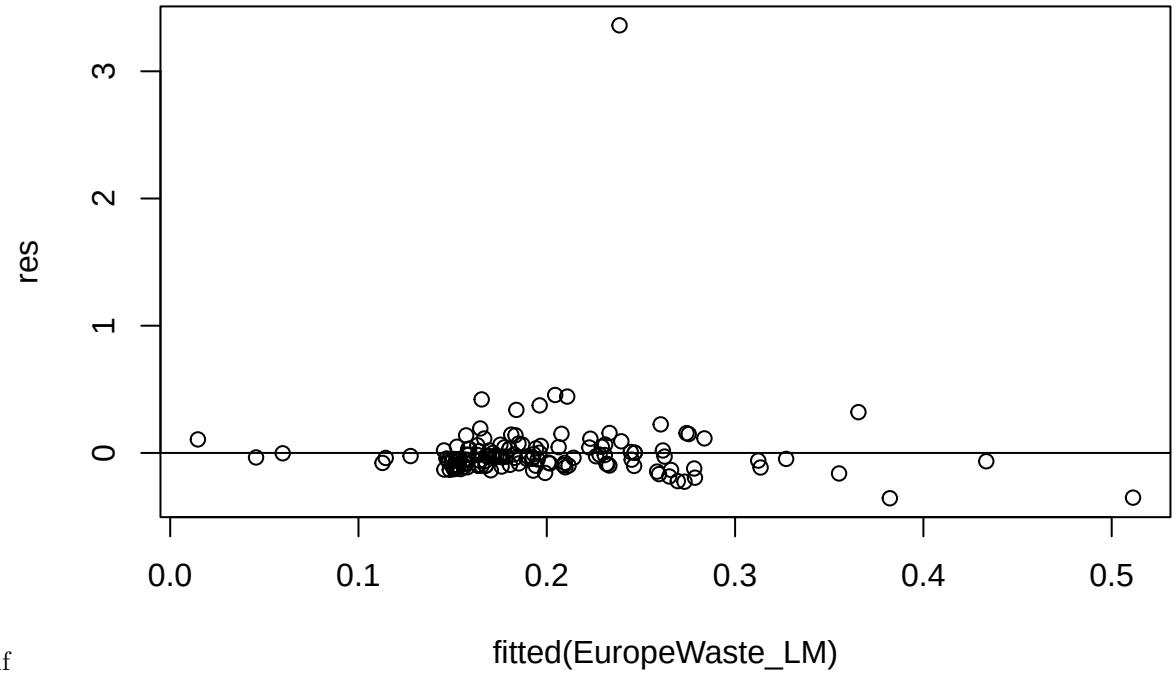
Now we have constructed our model, we need to check if it is indeed a good fit. As mentioned in the theory utilising residuals and constucting two of the most common ways to do this; a normal QQ-plot for the residuals and a residuals vs fitted values plot.

Exercise 5

Construct both a normal QQ-plot and a residulas vs fitted values plot, noting you already saved your model as an object. Interpret the relevant plots. [Hint: Your code may give you additional plots, we will discuss these in the next notebook]. Comment on your findings.

Your Answer:

```
res <- resid(EuropeWaste_LM)

fitted_res_plot <- plot(fitted(EuropeWaste_LM), res) %>%
    abline(0,0)
```

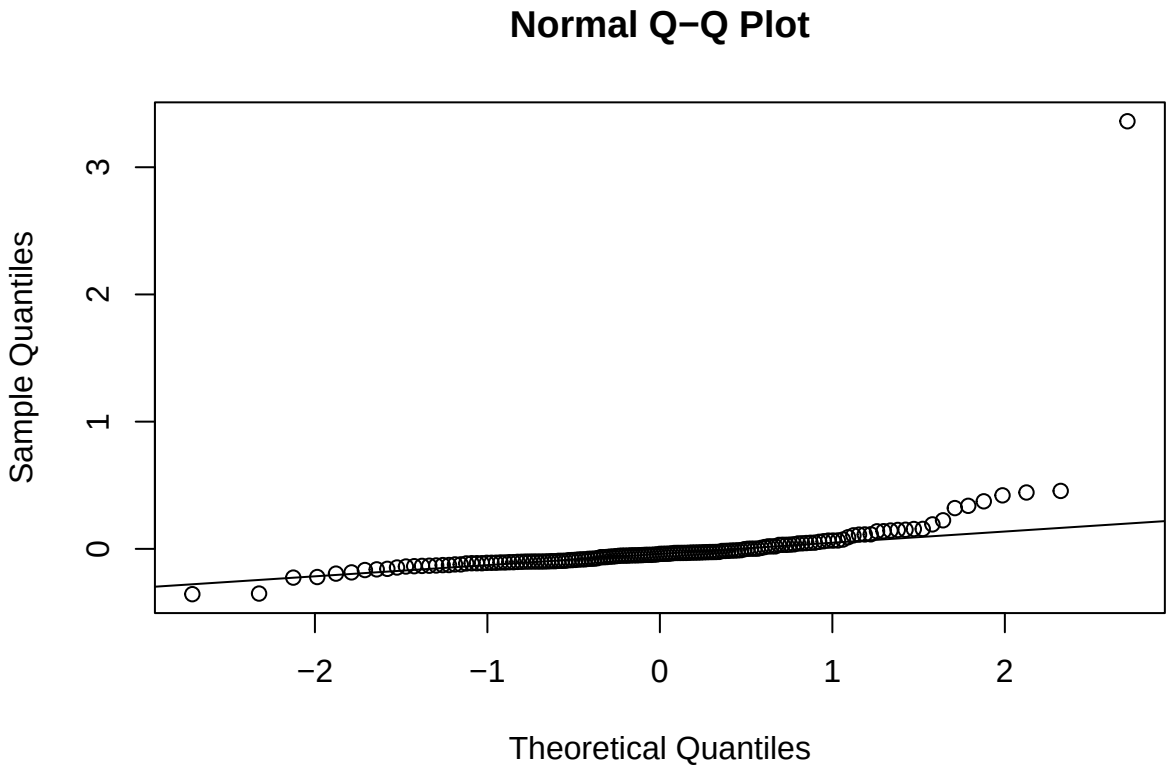


5-1.pdf

```
fitted_res_plot
```

```
## NULL
```

```
QQ_plot <- qqnorm(res)
qqline(res)
```



5-2.pdf

It would seem that the residuals are mostly random against the fitted values, with an even spread above and below the line. However, there is a large cluster which would indicate some trend. Furthermore whilst the residuals look close to the fitted line, towards the ends we see large deviations, potentially proving that the residuals are not normally distributed well.