Code

Understanding perception of the effectiveness of the Criminal Justice System in England and Wales

Ana Morales

2020-01-31

1. Introduction

In the last session we used exploratory data analysis to make sense and understand the data that we have. This is a fundamental step in any data analysis process, it allows us to know the data, find errors, patterns, etc. It also helps us to understand what we can and can't do with the data we have.

We looked at bivariate associations between variables, which helps us to have an idea of the relationship between a set of two variables. This is powerful, but rather limited in terms of inference we can make regarding the association between variables.

It is very unusual, specially in the social sciences, that a response variable will only depend on a single explanatory variable. Usually, we have several variables that affect our outcome variable at the same time, for those cases we use multiple linear regression.

In this tutorial we will give you a practical introduction to multiple linear regression. It does not intend to replace a formal theoretical explanation of this data analysis technique, but rather we will only scratch the surface. Hopefully this will also spark some interest (and why not joy?) in you.

We will also discuses the underlying assumptions of Multiple linear regression and the interpretation of the results.

2. Linear models

Linear models looks at the **linear** association between a continuous dependent variable (also known as "outcome") Y and a set of explanatory variables (also known as "predictors") x_1, x_2, x_3, x_n. In this tutorial we will explore the association between our dependent variable Y the perception of the effectiveness of the Criminal Justice System [1] and a set of explanatory variables X_i age, whether victim of a crime, sex.

You will normally find an equation similar to this one:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \epsilon_i$$

We will get back to this equation later.

2.1. Linear models in R

Before we start, remember

- Open a script
- load your packages
- Set working directory
- Load your data

```
library(tidyverse)
library(haven)
```

We will use the modified file of the csew that we used in Practical 2. Since it's a `.RData` file, we use the function `load()`

```
load(file = "csew.RData")
```

Simple linear regression

A simple linear regression is a linear model with 1 outcome and a single explanatory variable.

We use the function `lm` which stands for "linear model" from the R base package. The model is specified as follows:

```
m0<- lm(effectx ~ age, data = csew)
summary(m0)
```

```
##
## Call:
## lm(formula = effectx ~ age, data = csew)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5138 -0.6656 -0.0103  0.6763  2.4681
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.4368191  0.0253230  -17.25   <2e-16 ***
## age          0.0086867  0.0004733   18.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9873 on 13437 degrees of freedom
##   (21932 observations deleted due to missingness)
## Multiple R-squared:  0.02445,    Adjusted R-squared:  0.02438
## F-statistic: 336.8 on 1 and 13437 DF,  p-value: < 2.2e-16
```

There is a lot of information in that output and we will be checking them all, but step by step. Remember that regression equation above?. This is a version of the same one for a simple linear regression

$$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

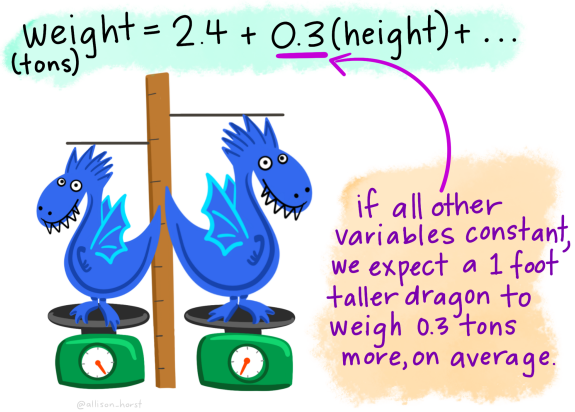
- Y_i: is effectx, our outcome variable
- x_{i1}: is age, our explanatory variable
- β_0: is the intercept term
- β_1: is the estimate value for the variable age (also known as "slope")

We can assess whether there is a significant association between the variables effectx and age by looking at the p-value column (`Pr(>|t|)`), the asterisks denote the levels of significance, `***` is equivalent to 0.001.

Interpretation

The association between effectx and age is positive, the value is 0.0086867. This is the predicted value of the perception of effectiveness of the CJS score *when age is 1* (remember that we only have data for respondents aged 16 years and older).

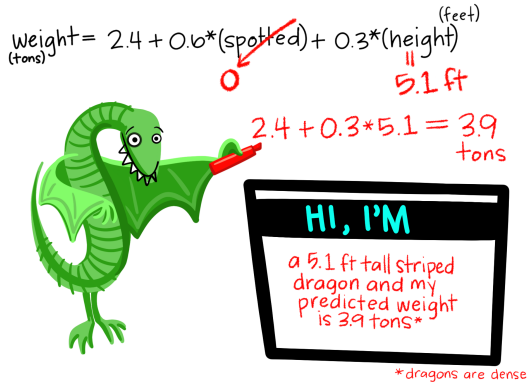
This indicates that for each year of age the perception of the effectiveness in the CJS increase by 0.009. This might not seem a lot, but we have to consider the scale of the effectx variable (range from -3 to 2.2 approx.).



Credits: Allison Horst (https://github.com/allisonhorst/stats-illustrations)

So, the predicted value for a 16 year old will be (using the equation) as indicated below. We will ignore the residual term for now.

$$Y_i = \beta_0 + \beta_1 x_{i1}$$
$$\hat{y} = -0.44 + 0.009 * age$$
$$\hat{y} = -0.44 + 0.009 * 16$$
$$\hat{y} = -0.296$$



Credits: Allison Horst (https://github.com/allisonhorst/stats-illustrations)

If you are not a fan of doing calculations by hand, you can use the function predict, we put the data and the variable in the data that we want to predict, in this case age, followed by the value (differences are due to rounding)

```
predict(m0, data.frame(age = c(16)))
```

```
##         1
## -0.2978326
```

Our model is not a very good one, the adjusted R-squared is 0.02438, which means that we are explaining only a 2.4% of the variation of effectx with a single explanatory variable.

Multiple linear regression

In R this is really simple, we use the same function for simple linear regression (SLR) and to include more predictors we use the mathematical sign +

```
m1<- lm(effectx ~ age + bcsvictimf, data = csew)

summary(m1)
```

```
##
## Call:
## lm(formula = effectx ~ age + bcsvictimf, data = csew)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5733 -0.6710 -0.0079  0.6756  2.5332
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.514866   0.026305  -19.57   <2e-16 ***
## age                         0.009448   0.000477   19.81   <2e-16 ***
## bcsvictimfVictim of crime  0.242432   0.023203   10.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9834 on 13436 degrees of freedom
##   (21932 observations deleted due to missingness)
## Multiple R-squared:  0.03232,    Adjusted R-squared:  0.03217
## F-statistic: 224.3 on 2 and 13436 DF,  p-value: < 2.2e-16
```

The interpretation is similar to the SLR, the only difference here is the inclusion of a categorical variable “bcsvictimf”. This variable has only two values, “yes” and “not” but the model estimates table will only print one of these values. The one that is printed in the resulting table depends what your reference category is; in this case, “not being a victim”[2] is the reference, so the value that is printed corresponds to who have been “victims of crime”.

The estimated value (“estimate” in the R output) will tell you how much or less (depending on the sign) being a victim a crime affects the perception of the effectiveness of the CJS.

Credits: Allison Horst (https://github.com/allisonhorst/stats-illustrations)

Now it is your turn to interpret m1:

Q1: How Does being a victim of crime in the previous 12 months associated with the perception of effectiveness of the Criminal Justice Service?

_____

_____

Q2: Is this association significant?

_____

_____

Q3: Did the model improve after controlling for another variable?

_____

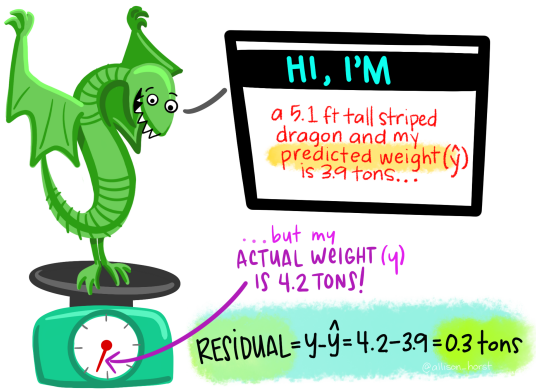
_____

Your turn

Now add other predictors to the model and describe your results

2.2. Residuals

Residuals are the difference between the observed values of y for each case in our data minus the predicted or expected value of $\hat{y}$, the one we obtained with our model. If the model is good, these differences will be minimal and we can say that "our model fits the data well". Unfortunately, most of the time, this is not the case.



Credits: Allison Horst (https://github.com/allisonhorst/stats-illustrations)

Residuals can tell us how much variation is left unexplained after we control for key variables. In other words, how much of the variation of the perception of the effectiveness is left unexplained by our model that includes being a victim of crime and age. Remember that R-squared is a measure of

how much variation we explain, our models don't perform really well here.

We need to find a model that fits the data "adequately", remember the saying:

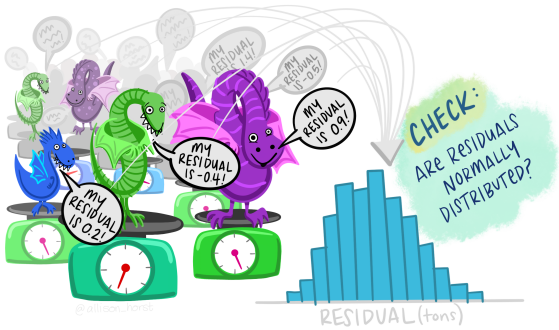
> "all models are wrong, but some are useful"
> (https://en.wikipedia.org/wiki/George_E._P._Box)
>
> —George Box

How do we know if our model is good enough for our research question? besides our literature review, we need to check the assumptions of the model.

2.3. Assumption checking

Normality of the residuals

A key assumption in linear regression is that our residuals are normally distributed. This means roughly that whatever is left unexplained by our model could be thought of as random variation or "white noise".

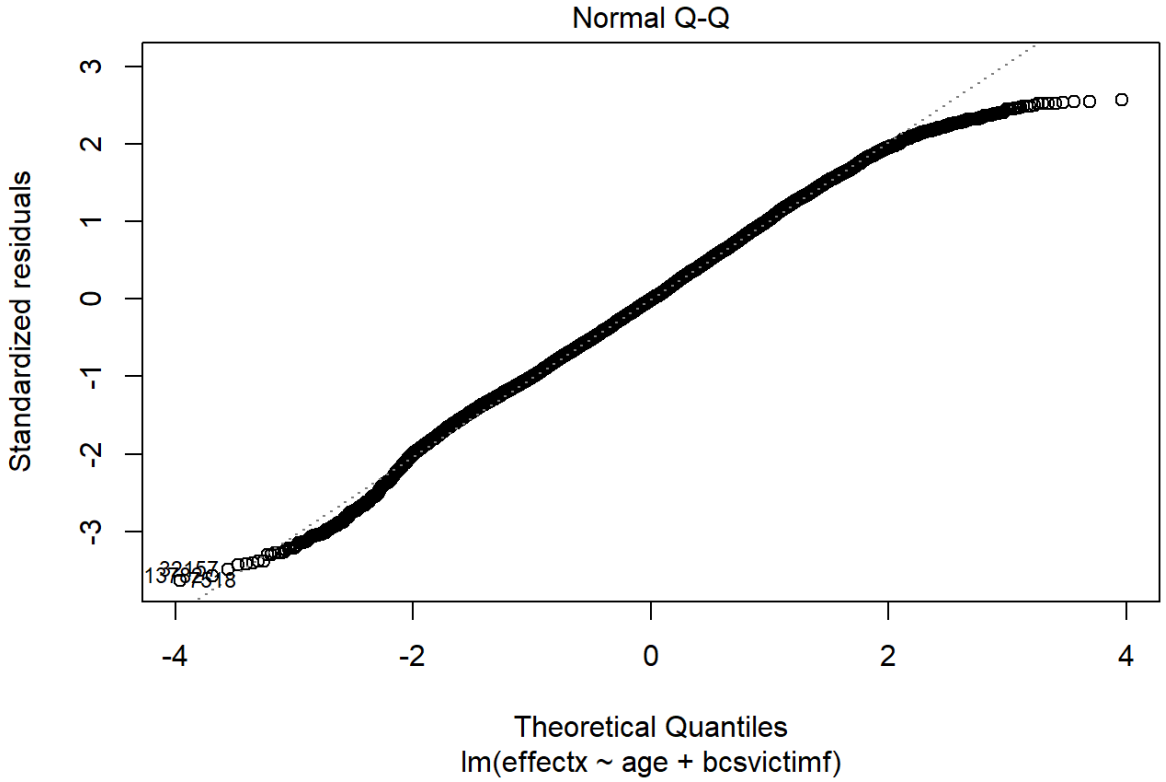


Credits: Allison Horst (https://github.com/allisonhorst/stats-illustrations)

There are formal statistical tests for this, but visual inspection of the residuals is usually the preferred method, since a non-normal distribution stands out very clearly.

To plot the residuals and check their distribution we can retrieve from our model results as follows:

```
plot(m1, 2)
```



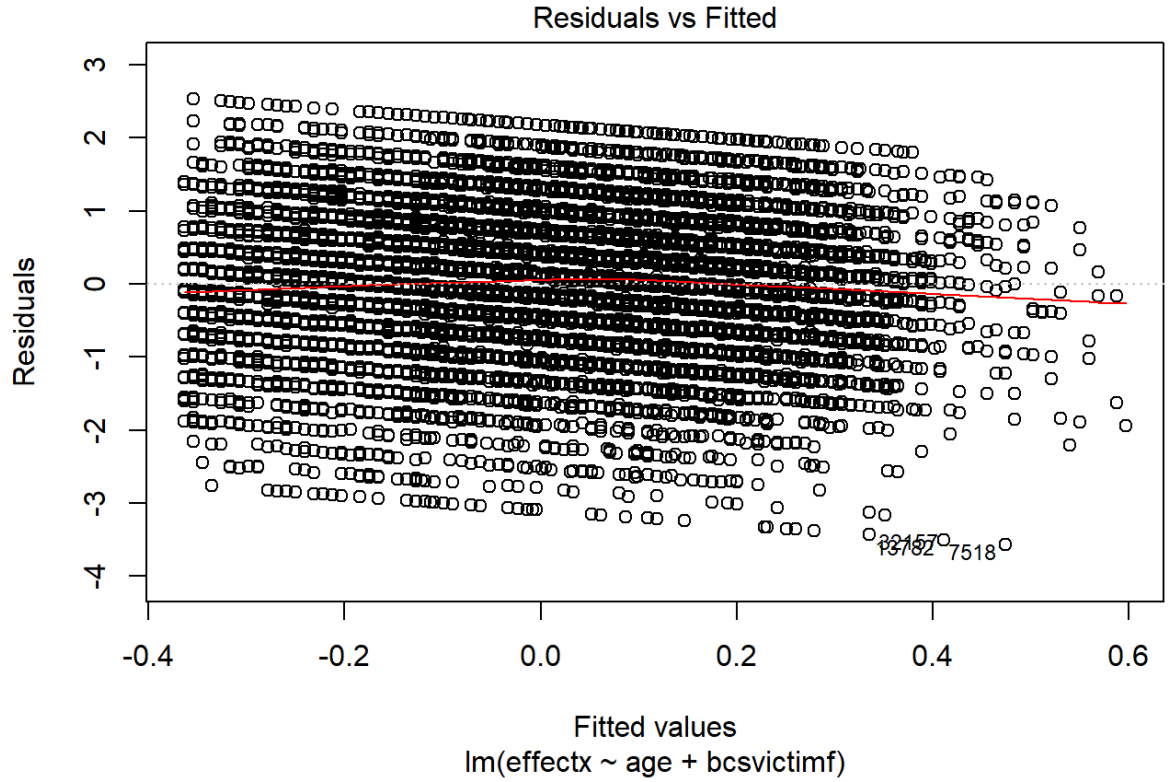
The result is a Q-Q plot or normal probability plot. It compares the residuals against the normal distribution. Put simply, the residuals (dots on the plot) should all lie on the diagonal line. Any obvious patterns would indicate that our model does not meet the normality assumption.

Homoscedasticity

Homoscedasticity means that residuals should distribute evenly in terms of spread across the predicted values of our model. In other words, if we plot residuals vs fitted values, it should look completely random, i.e. a random cloud of dots.

To check this, we can run the following code:

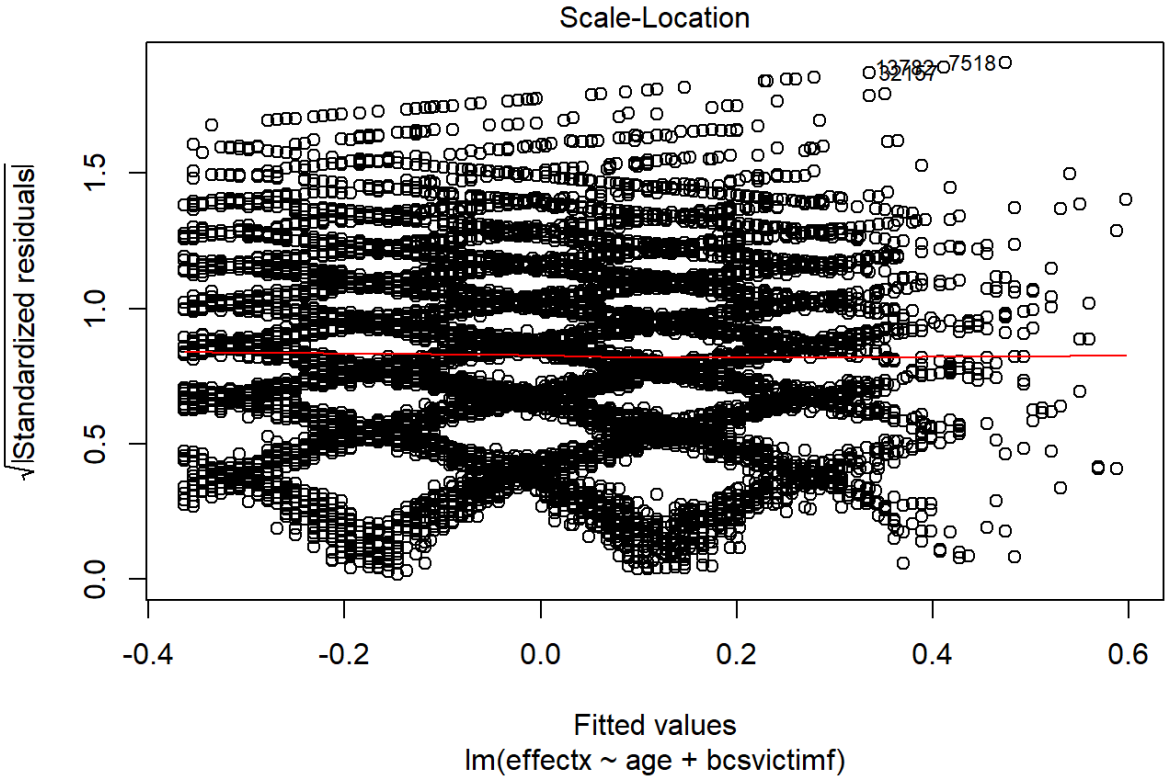
```
plot(m1, 1)
```



Another plot to check the assumption of homoscedasticity (equal variance) is the Scale-Location plot. We have good news if there is an horizontal line with equally random dots around.

Check this:

```
plot(m1, 3)
```

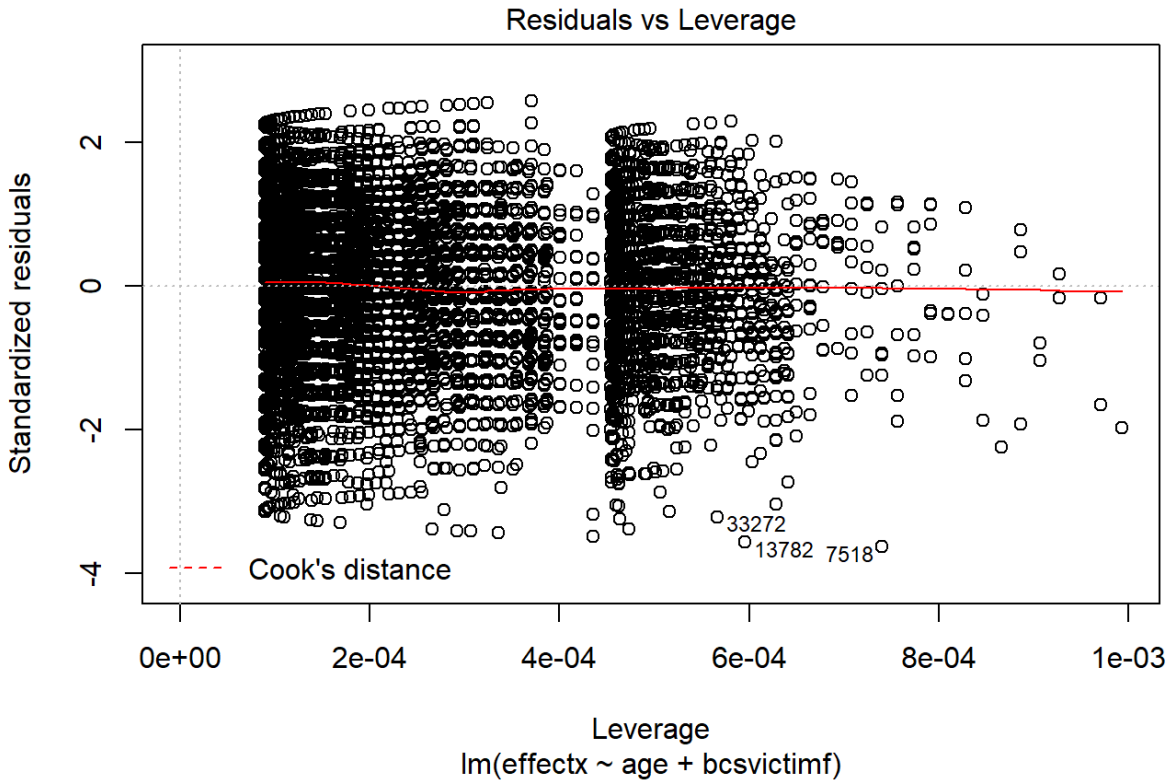


Influential cases: Residual vs Leverage

This plot will display cases that have a great influence on the regression (leverage points). These cases are individualised in the plot.

To check this, we can run the following code:

```
plot(m1, 5)
```



Q.4 Now it's judgment time. How our model perform in terms of:

- Variance explained?
- Regression assumptions?

Q.5 Write a short conclusion about the model, it's strength and limitations.

---

1. The variable `effectx` effectiveness of the Criminal Justice System is a derived variable made available for teaching purpose only (for more information check the user guide (http://doc.ukdataservice.ac.uk/doc/7911/mrdoc/pdf/7911_csew_2013-14_teaching_dataset_user_guide.pdf)). As a consequence, the conclusions reached from these analysis can not be taken as the real association between the variables analysed.↩
2. The variable "bcsvictimf" is a categorical variable of class `factor`, so there is no numerical value associated with these labels.↩