Practical 314

Stefano De Sabbata

2020-12-17

Comparing data

Stefano De Sabbata

Introduction

The first part of this practical guides you through the ANOVA (analysis of variance) and regression analysis seen in the lecture, the last part showcases a multiple regression analysis. Create a new R project for this practical session and create a new RMarkdown document to replicate the analysis in this document and a separate RMarkdown document to work on the exercises.

```
library(tidyverse)
library(magrittr)
library(knitr)
```

As many of the functions used in the analyses below are part of the oldest libraries developed for R, they have not been developed to be easily compatible with the Tidyverse and the `%>%` operator. Fortunately, the `magrittr` library (loaded above) does not only define the `%>%` operator seen so far, but also the exposition pipe operator `%$%`, which exposes the columns of the data.frame on the left of the operator to the expression on the right of the operator. That is, `%$%` allows to refer to the column of the data.frame directly in the subsequent expression. As such, the lines below expose the column `Petal.Length` of the data.frame `iris` and to pass it on to the `mean` function using different approaches, but they are all equivalent in their outcome.

```
# Classic R approach
mean(iris$Petal.Length)
```

```
## [1] 3.758
```

```
# Using %>% pipe
iris$Petal.Length %>%
  mean()
```

```
## [1] 3.758
```

```
# Using %>% pipe and %$% exposition pipe
iris %$% Petal.Length %>%
  mean()
```

```
## [1] 3.758
```

ANOVA

The ANOVA (analysis of variance) tests whether the values of a variable (e.g., length of the petal) are on average different for different groups (e.g., different species of iris). ANOVA has been developed as a generalised version of the t-test, which has the same objective but allows to test only two groups.

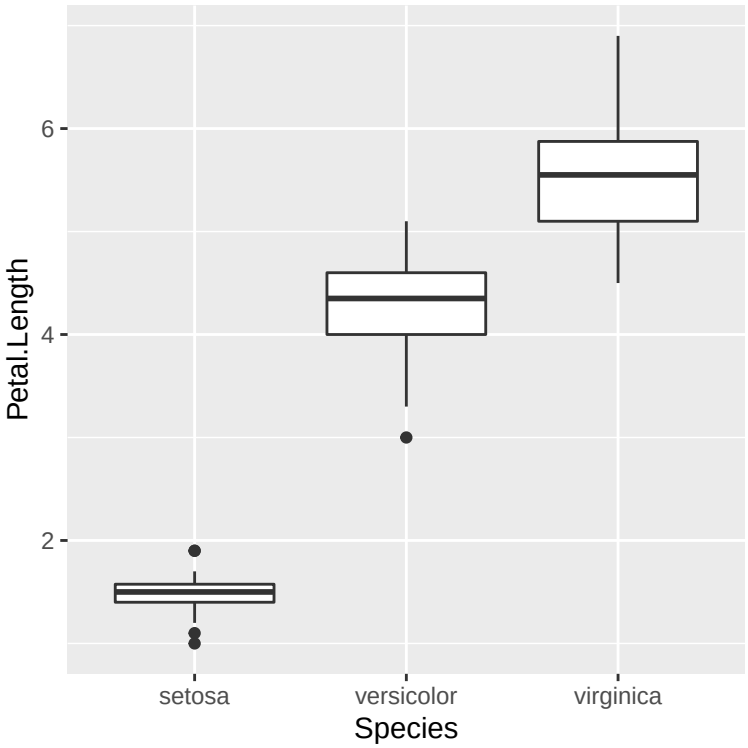
The ANOVA test has the following assumptions:

- normally distributed values in groups
  - especially if groups have different sizes
- homogeneity of variance of values in groups
  - if groups have different sizes
- independence of groups

Example

The example seen in the lecture illustrates how ANOVA can be used to verify that the three different species of iris in the `iris` dataset have different petal length.

```
iris %>%
  ggplot2::ggplot(
    aes(
      x = Species,
      y = Petal.Length
    )
  ) +
  ggplot2::geom_boxplot()
```



ANOVA is considered a robust test, thus, as the groups are of the same size, there is no need to test for the homogeneity of variance. Furthermore, the groups come from different species of flowers, so there is no need to test the independence of the values. The only assumption that needs testing is whether the values in the three groups are normally distributed. As there are 50 flowers per species, we can set the significance threshold to `0.05`.

The three Shapiro–Wilk tests below are all not significant, which indicates that all three groups have normally distributed values.

```
iris %>% dplyr::filter(Species == "setosa") %>% dplyr::pull(Petal.Length) %>% stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.95498, p-value = 0.05481
```

```
iris %>% dplyr::filter(Species == "versicolor") %>% dplyr::pull(Petal.Length) %>% stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.966, p-value = 0.1585
```

```
iris %>% dplyr::filter(Species == "virginica") %>% dplyr::pull(Petal.Length) %>% stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.96219, p-value = 0.1098
```

We can thus conduct the ANOVA test using the function `aov`, and the function `summary` to obtain the summary of the results of the test.

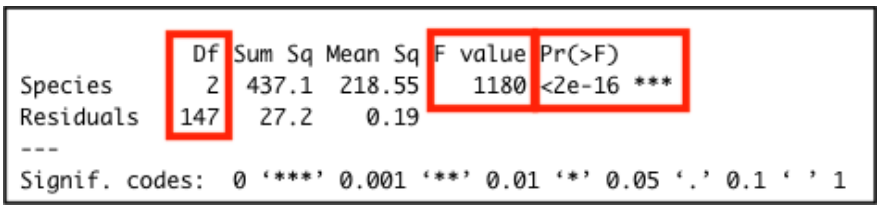
```
# Classic R coding approach (not using %$%)
# iris_anova <- aov(Petal.Length ~ Species, data = iris)
# summary(iris_anova)

iris %$%
  stats::aov(Petal.Length ~ Species) %>%
  summary()
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  437.1  218.55    1180 <2e-16 ***
## Residuals   147   27.2    0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The difference is significant $\mathrm{F}(2, 147) = 1180.16$, $p < .01$.

The image below highlights the important values in the output: the significance value `Pr(>F)`; the F-statistic value `F value`; and the two degrees of freedom values for the F-statistic in the `Df` column.



Exercise 314.1

Question 314.1.1: Load the `2011_OAC_Raw_uVariables_Leicester.csv` dataset. Check whether the values of mean age (`u020`) are normally distributed, and whether they can be transformed to a normally distributed set using logarithmic or inverse hyperbolic sine functions.

Question 314.1.2: Check whether the values of mean age (`u020`) are normally distributed when looking at the different 2011OAC supergroups separately. Check whether they can be transformed to a normally distributed set using logarithmic or inverse hyperbolic sine functions.

Question 314.1.3: Is the distribution of mean age (`u020`) different in different 2011OAC supergroups in Leicester?

Correlation

The term **correlation** is used to refer to a series of a standardised measures of covariance, which can be used to statistically assess whether two variables are related or not.

Furthermore, if two variables are related, such measures can identify whether they are:

- positively related:
  - entities with *high values* in one tend to have *high values* in the other;
  - entities with *low values* in one tend to have *low values* in the other;
- negatively:
  - entities with *high values* in one tend to have *low values* in the other;
  - entities with *low values* in one tend to have *high values* in the other.

Correlation can be calculated in many ways, but there are three approaches which are by far the most common. They all start from the null hypothesis that there is no relationship between the variables. Thus, if the p-value is above a pre-defined significance threshold, the null hypothesis is rejected, and the conclusion is that there is a relationship between the two variables.

If the test is significant is the case:

- a **positive** correlation value indicates a positive relationship;
- a **negative** correlation value indicates a negative relationship;
- the **square** of the correlation value can be taken as an indication of the percentage of shared variance between the two variables.

However, each one has different assumptions about the variables' distribution and thus implements the same general ideal measure in a different way:

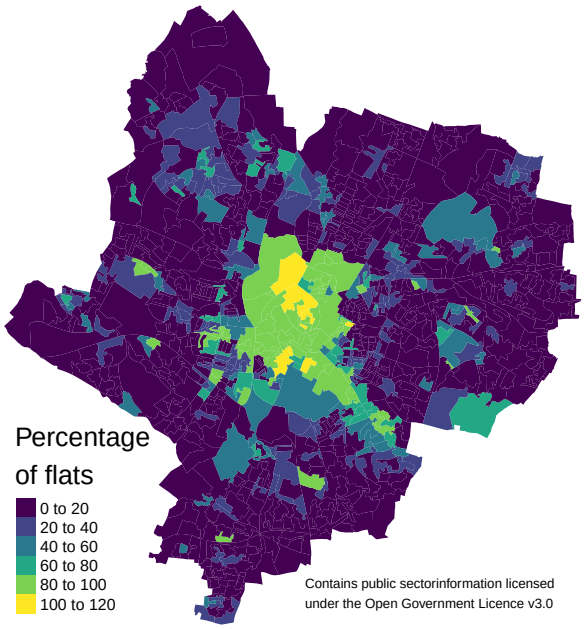
- if two variables are **normally distributed**:
  - *Pearson's r*;
- if two variables are **not normally distributed**:
  - if there are **no ties among values**:
    - *Spearman's rho*;
  - if there are **ties among values**:
    - *Kendall's tau*.

Example

When studying how people live in cities, a number of questions might arise about where the live and how the move around the city. For instance, looking at a map of Leicester, it is clear that (has in many English cities) there seems to be a very high concentration of flats in the city centre. At the same time, there seems to be almost no flats at all in the suburbs. This might led us to ask: *"do households living in flats (and thus mostly in the city centre) own the same amount of cars as households living in the city center?"*

That could be due to many reasons. As the suburbs in England are largely residential, whereas most working places are located in the city centre. As such people living in flats might be more likely to walk or cycle to

work, or commute using public transportation within the city or to other cities. City centres usually afford less spaces for parking. Many flats are rented to students, who might be less likely to own a car. The list could continue, but these are still hypothesis based on a certain (probably biased) view of the city. Can we use data analysis to explore whether there is any ground to such an hypothesis?



The dataset used to create the 2011 Output Area Classification (2011OAC) contains two variables that might help explore this issue. These data are not very current anymore, and they are not they values we might collect if we were to conduct a fresh survey for this specific study. However, they can still provide some insight.

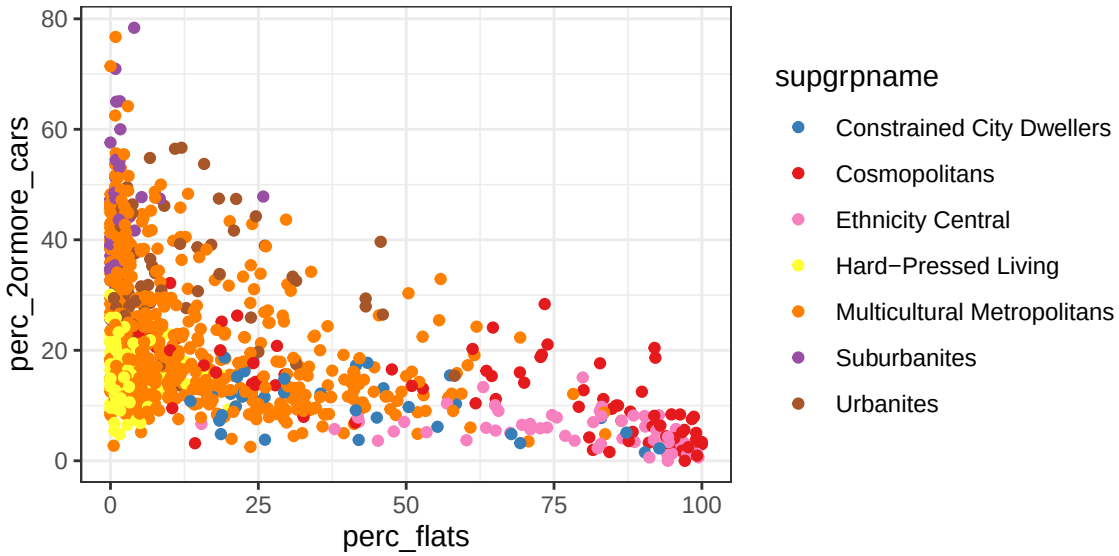
- `u089`: count of flats per Output Area (OA). The statistical unit for this variable is `Household_Spaces`. As OAs vary in size and composition, we can use `Total_Household_Spaces` to calculate the percentage of flats per OA, which is a more stable measure.
  - `perc_flats = (u089 / Total_Household_Spaces) * 100`
- `u118`: 2 or more cars or vans in household. The statistical unit for this variable is `Household`. As OAs vary in size and composition, we can use `Total_Household_Spaces` to calculate the percentage of households per OA with 2 or more cars or vans, which is a more stable measure.
  - `perc_2ormore_cars = (u118 / Total_Households) * 100`

The process of transforming variables to be within a certain range (such as a percentage, thus using a `[0..100]` range, or a `[0..1]` range) is commonly referred to as **normalisation**. The process of transforming a variable to have mean zero and standard deviation one (z-scores) is commonly referred to as **standardisation**. However, note that these terms are sometime used interchangably.

```
flats_and_cars <-
  leicester_2011OAC %>%
  dplyr::mutate(
    perc_flats = (u089 / Total_Household_Spaces) * 100,
    perc_2ormore_cars = (u118 / Total_Households) * 100
  ) %>%
  dplyr::select(
    OA11CD, supgrpname, supgrpcode,
    perc_flats, perc_2ormore_cars
  )
```

Plotting the two variables together in a scatterplot reveals a pattern. Indeed, a very low percentage of

households living in flats own two or more cars. However, the proportion of households owning two or more cars who live in the suburbs seem to span almost throughout the whole range, from zero to 80%. That seems to indicate some level of negative relationship, but the picture is clearly far less clear-cut as we might have initially assumed. The initial assumption about car ownership for households living in flats seems to hold, but we probably didn't consider the situation in the suburbs with sufficient care.



The first step in establishing whether there is a relationship between the two variables is to assess whether they are normally distributed, and thus which correlation test we should use for the analysis. The scatterplot already seem to suggest that the variables are rather skewed.

As there are 969 OAs in Leicester, we can set the significance threshold to `0.01`. The results of the `stats::shapiro.test` functions below show that neither of the two variables are normally distributed. Transforming the variables using the *inverse hyperbolic sine* still does not result in normally distributed variables. Thus, we should discard *Pearson's r* as an option to explore the correlation between the two variables.

```
flats_and_cars %>%
  dplyr::select(perc_flats, perc_2ormore_cars) %>%
  dplyr::mutate(
    ihs_perc_flats = asinh(perc_flats),
    ihs_perc_2omcars = asinh(perc_2ormore_cars)
  ) %>%
  pastecs::stat.desc(basic = FALSE, desc = FALSE, norm = TRUE) %>%
  knitr::kable()
```

	perc_flats	perc_2ormore_cars	ihs_perc_flats	ihs_perc_2omcars
skewness	1.5621906	0.9075026	-0.0927406	-0.9460022
skew.2SE	9.9417094	5.7753049	-0.5901967	-6.0203149
kurtosis	1.3282688	0.4588571	-1.1009004	1.6988166
kurt.2SE	4.2308489	1.4615680	-3.5066270	5.4111309
normtest.W	0.7443821	0.9328442	0.9572430	0.9514757
normtest.p	0.0000000	0.0000000	0.0000000	0.0000000

The next step is to assess whether there are ties among the values in the two variables. The code below fist counts the number of cases per value. Then it counts the number of values for which the number of cases is greater than one.

```
ties_perc_flats <-
  flats_and_cars %>%
  dplyr::count(perc_flats) %>%
  dplyr::filter(n > 1) %>%
  # Specify wt = n() to count rows
  # otherwise n is taken as weight
  dplyr::count(wt = n()) %>%
  dplyr::pull(n)

ties_perc_2ormore_cars <-
  flats_and_cars %>%
  dplyr::count(perc_2ormore_cars) %>%
  dplyr::filter(n > 1) %>%
  # Specify wt = n() to count rows
  # otherwise n is taken as weight
  dplyr::count(wt = n()) %>%
  dplyr::pull(n)
```

The variable `perc_flats` has 127 values with ties and `perc_2ormore_cars` has 115 values with ties. As such, using *Spearman's rho* is not advisable and *Kendall's tau* should be used. As above, we can set the significance threshold to `0.01`.

Finally, we can run the `stats::cor.test` function to assess the relationship between the two variables. The code below saves the results of the test to a variable. This afford to subsequent actions. First, we can show the full results by simply invoking the name of the variable (term used in the programming-related meaning here) in the final line of the code. Second, we can extract and square the estimate value in RMarkdwon in the following paragraph, to show the percentage of shared variace.

```
flats_and_cars_corKendall <-
  flats_and_cars %$%
  stats::cor.test(
    perc_flats, perc_2ormore_cars,
    method = "kendall"
  )

flats_and_cars_corKendall
```

```
##
##  Kendall's rank correlation tau
##
## data:  perc_flats and perc_2ormore_cars
## z = -19.026, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.4094335
```

```
The percentage of flats and the percentage of households
owning 2 or more cars or vans per OA in the city of Leicester
are negative related, as the relationship is significant
(‘p-value < 0.01‘) and the correlation value is negative
(‘tau =‘ -0.41). The two variables share 16.8% of variance. We can thus conclude
that there is significant but very weak relationship
between the two variables.
```

The percentage of flats and the percentage of households owning 2 or more cars or vans per OA in the city of

Leicester are negative related, as the relationship is significant (`p-value < 0.01`) and the correlation value is negative (`tau = -0.41`). The two variables share 16.8% of variance. We can thus conclude that there is significant but very weak relationship between the two variables.

Exercise 314.2

Question 314.2.1: As mentioned above, when discussing movement in cities, there is an assumption that people living in the city centre live in flats and work or cycle to work, whereas people living in the suburbs live in whole houses and commute via car. Study the correlation between the presence of flats (`u089`) and people commuting to work on foot, bicycle or other similar means (`u122`) in the same OAs. Consider whether the values might need to be normalised or otherwised transformed before starting the testing procedure.

Question 314.2.2: Another interesting issue to explore is the relationship between car ownership and the use of public transport. Study the correlation between the presence of households owning 2 or more cars or vans (`u118`) and people commuting to work via public transport (`u120`) or on foot, bicycle or other similar means (`u122`) in the same OAs. Consider whether the values might need to be normalised or otherwised transformed before starting the testing procedure.