Plotting with ggplot2

2031 - Introduction to Statistical Computing

Sam Clifford

2022-11-18

Introduction

About this practical session

In the lecture session we introduced visualisation with the histogram, x-y plots and other scatter plot techniques, and touched on some key principles from the work of both Tukey and Tufte.

We will be looking at data on the forced expiratory volume in the first second of breath, FEV1, a measure of lung function. The data are from the "Six Cities" study, which aimed to investigate the relationship between air pollution and mortality described in Dockery et al. (1983). Re-analysis of this data is detailed in Krewski, Burnett, Goldberg, Hoover, Siemiatycki, Abrahamowicz, and White (2005) and Krewski, Burnett, Goldberg, Hoover, Siemiatycki, Abrahamowicz, Villeneuve, et al. (2005), and a discussion of the original study and the replicability of its results is found in Choirat, Braun, and Kioumourtzoglou (2019).

Getting help: You may find it useful to look at the RStudio cheatsheets (RStudio 2012) and ggplot2 documentation for hints on how to implement particular graphical ideas. Another good resource is "R for Data Science" (Wickham and Grolemund 2020), particularly Chapter 3, "Data Visualisation," and Chapter 7, "Exploratory Data Analysis."

Learning outcomes

- Assumed skills
    - Writing code into a script file
    - Understanding of x-y plots
    - Reading documentation, including the help file
- Learning objectives
    - Creating a graph using a layered grammar of graphics
    - Being able to critique a graph that you have created
- Professional skills
    - Creating graphics which are reproducible and clear
    - Documenting code by commenting

Group formation

Assign the following roles. Someone may need to take on multiple roles or you may need to split a role depending on the size of your group.

A reminder of expectations in the prac:

- Keep a record of the work being completed with a well-commented R script in your forked github repository
- Allow everyone a chance to participate in the learning activities
- All opinions are valued provided they do not harm others
- Everyone is expected to do the work, learning seldom occurs solely by watching someone else do work

Obtaining and loading the data

Activity 1 - Forking the git repository

Browse to the git repository at https://github.com/samclifford/2031_eda

If you feel comfortable with git already, click the “Fork” button in the top right corner. This will give you a copy of the repository under your own account that you’ll have write access to. Clone the remote repository (on the GitHub server) to your local machine.

If you do not feel comfortable with git, download the repository instead by clicking the green Code button and selecting Download ZIP.

There are partially worked code solutions in the `ggplot2.R` script in the git repository (or its download) to get you started on some of these questions. After answering each question, save your script file (and commit the changes in git if using it).

Activity 2 - Loading the data

Load the tidyverse package and then read the data in with `read_csv()`. Ensure you give your data object a meaningful name and that `id` is a factor variable.

The next block of code samples the data table to ensure that we have the data for 20 individuals with more than 6 records. We will be looking at a subset of the data for now.

Simple plots

Activity 3a - A simple scatter plot

Build a plot that shows the relationship between FEV1 and age.

Question: Given the strength of the linear association between these two variables, do you think a linear trend would be an appropriate model?

Activity 3b - Labels and theme

You may wish to save the plot object in 3a and add to it, continuing to do so from here on, or copy-paste the code. Add meaningful labels for the x and y axes, including units, and change the theme from the default. You may want to consult the suggested reading or search online for the included ggplot2 theme choices.

Activity 3c - Scatterplot smoother

Add a smooth line of best fit to the plot. You may wish to change its colour, turn off the standard error ribbon, or make other changes to it to help show the data and improve contrast with the background colour of your plot.

The default behaviour is to use a LOESS smoother (Cleveland, Grosse, and Shyu 1992) which can be set with `method = 'loess'` as an argument to `geom_smooth()`. You could also use a generalised additive model (S. N. Wood 2017) with `method = 'mgcv'`.

Question: Given the strength of the linear association between these two variables, do you think a linear trend would be an appropriate model?

Extension activities (choose one)

You are expected to do at least one of these, ideally as a group, and show your work in the end of class discussion.

Activity 4a (extension) - Showing further structure

We have repeat measurements on 20 individuals. Through either small multiples, geometry grouping, or other aesthetic options, determine a way to highlight which observations belong to the same individual.

Activity 4b (extension) - How many observations per individual?

Many of the 300 individuals in the downloaded data set have been measured multiple times over the years. Count the number of times that each `id` is measured and make a bar plot to summarise the proportion of individuals who have 1, 2, etc. measurements.

Activity 4c (extension) - Incorporating height

Make a plot that shows both FEV1 and age but also includes height. There are a number of ways to do this.

Activity 4d (advanced extension) - Accounting for repeat measurement

If you intend to do this activity please be aware that it uses an R package you may not be familiar with to fit Generalised Additive Mixed Models.

Build a regression model for the change in FEV1 with age that accounts for repeat measurement of individuals. Using the `gamm` function (S. N. Wood 2004) in the mgcv package in R (S. N. Wood 2017), we can fit a mixed effects model that uses a spline for the effect of age and has a random effects mean to account for the differences in baseline FEV1 across individual. When building your prediction data frame, make sure that you give the predicted values the name `FEV1` so you can more easily reuse the aesthetics inherited from the base plot.

Finishing up

Ensure that you have saved all your work. If you have used git, push all your committed changes up to your remote repository and share the link with your group so they have access to it. If you haven't used git, ZIP and save your folder.

Further reading

A lot of the key ideas in data visualisation that we investigate arose with Tufte (1983), and are summarised by Pantoliano (2012). Tufte's website is well worth exploring, particularly the discussion on how the visual presentation of information could have helped avert the *Challenger* disaster (Tufte 1997). For some more guidance on using ggplot2 for data visualisation, check Chapter 3 of Wickham and Grolemund (2020), the RStudio cheatsheets (RStudio 2012), and Chang (2017).

References

Chang, Winston. 2017. *R Graphics Cookbook: Practical Recipes for Visualizing Data*. 2nd ed. O'Reilly Media. http://www.cookbook-r.com/Graphs/.

Choirat, Christine, Danielle Braun, and Marianthi-Anna Kioumourtzoglou. 2019. "Data Science in Environmental Health Research." *Current Epidemiology Reports* 6 (3): 291–99. https://doi.org/10.1007/s40471-019-00205-5.

Cleveland, WS, E Grosse, and WM Shyu. 1992. "Local Regression Models." In *Statistical Models in s*, edited by JM Chambers and TJ Hastie. Wadsworth & Brooks/Cole, Pacific Grove, CA.

Dockery, DW, CS Berkey, JH Ware, FE Speizer, and BG Ferris Jr. 1983. "Distribution of Forced Vital Capacity and Forced Expiratory Volume in One Second in Children 6 to 11 Years of Age." *American Review of Respiratory Disease* 128 (3): 405–12.

Krewski, D., R. T. Burnett, M. Goldberg, K. Hoover, J. Siemiatycki, M. Abrahamowicz, P. J. Villeneuve, and W. White. 2005. "Reanalysis of the Harvard Six Cities Study, Part II: Sensitivity Analysis." *Inhalation Toxicology* 17 (7-8): 343–53. https://doi.org/10.1080/08958370590929439.

Krewski, D., R. T. Burnett, M. Goldberg, K. Hoover, J. Siemiatycki, M. Abrahamowicz, and W. White. 2005. "Reanalysis of the Harvard Six Cities Study, Part i: Validation and Replication." *Inhalation Toxicology* 17 (7-8): 335–42. https://doi.org/10.1080/08958370590929402.

Pantoliano, Mike. 2012. "Data Visualization Principles: Lessons from Tufte." 2012. https://moz.com/blog/data-visualization-principles-lessons-from-tufte.

RStudio. 2012. "RStudio Cheat Sheets." 2012. https://www.rstudio.com/resources/cheatsheets/.

Tufte, Edward R. 1983. *The Visual Display of Quantitative Information*. Graphics Press.

———. 1997. "Visual and Statistical Thinking." In *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press. https://www.edwardtufte.com/tufte/books_textb.

Wickham, Hadley, and Garrett Grolemund. 2020. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. http://r4ds.had.co.nz.

Wood, S. N. 2017. *Generalized Additive Models: An Introduction with r*. 2nd ed. Chapman; Hall/CRC.

Wood, S. N. 2004. "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models." *Journal of the American Statistical Association* 99 (467): 673–86.