LEARNR MODULES FOR SELF-PACED LEARNING OF LINEAR REGRESSION CONCEPTS

Katherine Daignault and Mohammed Kaviul Khan
University of Toronto, Canada
katherine.daignault@utoronto.ca

The Methods of Data Analysis 1 course at the University of Toronto is a theoretical and applied presentation of linear regression with a heavy emphasis on the use of the R statistical software. Prerequisite competence in inference and R programming vary dramatically, causing many students to struggle with new material. To address this disparity, learnR modules were developed for five course topics to provide guided practice in programming and review of relevant prerequisite topics. The modules further present new course concepts commonly misunderstood in an exploratory manner prior to formal introduction in class. The goal of these modules is to create opportunities for low-stakes R practice, review of concepts needed to build further notions in the course, and to develop an intuitive understanding of the theory and core concepts of linear regression. We anticipate students will gain confidence in the necessary skills to be successful in the course.

INTRODUCTION

In the current data landscape, students graduating with a statistics degree are required to be proficient in data analytical skills in at least one statistical software in addition to having comprehensive knowledge of common statistical methods (Nolan & Temple Lang, 2010). However, in programs where students have substantial choice in their statistics courses and program streams such as at the University of Toronto, situations arise where students entering required upper-year statistics courses vary in their exposure to statistical software and in their depth of prerequisite knowledge of statistical inference. Without consistent integration of statistical software and data analysis in early statistics courses alongside instruction focusing on formulae and calculation, students face steep learning curves when confronted with a course that requires a solid foundation in these areas to be successful (Tucker et.al., 2022).

One such course is the Methods of Data Analysis 1 course that teaches linear regression analysis with a central focus on theory, application of methods with software, and communication of analytical results to various stakeholders. As a consequence of the inconsistent exposure to statistical programming and communication, students frequently struggle with course concepts such as appropriate interpretation of results (Peterson & Ziegler, 2021) or the importance of assumptions for valid inference (Greenland et.al., 2016), while having difficulty transferring what they learn through instruction to an application with statistical software (Tucker et.al., 2022). The goal of this project was to create a series of learnR modules (Aden-Buie et.al., 2023) in the form of self-paced worksheets involving knowledge-check multiple choice questions and code chunks for completion of provided R code that aim to build an intuitive understanding of commonly misunderstood course concepts prior to their formal instruction.

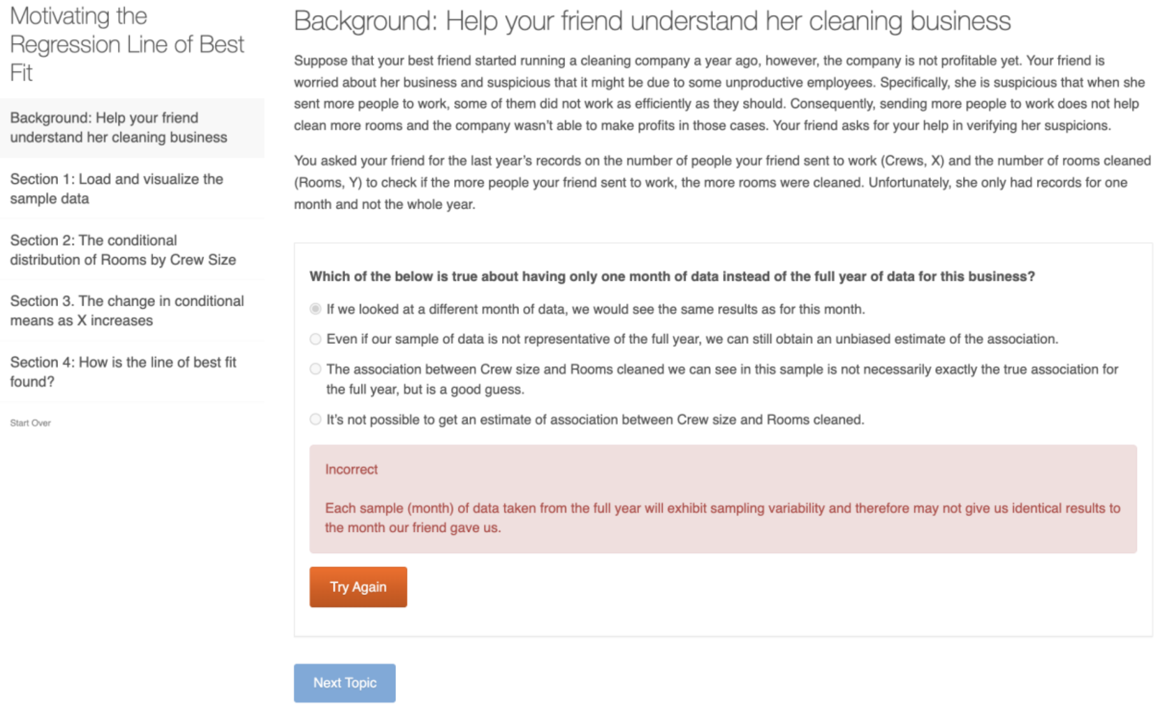
The use of software and computers for the purpose of teaching has grown in popularity to the point where there is an abundance of resources available to instructors. Laviolette (1994) noted the potential for software to be a valuable teaching tool for linear regression concepts for its graphical abilities. Recently, Shiny applets (Chang et.al., 2021) have gained in popularity for their interactivity and ability to allow students to explore statistical concepts on their own. Many such Shiny applets focus on concepts related to probability theory and basic statistical inference (Doi et.al., 2016). However, some exist that address more complicated topics, such as linear and logistic regression, time series, stochastic processes (Wang et. al., 2021), and extreme value distributions and analyses under a Bayesian framework (Fawcett, 2018). Others have developed software outside of the Shiny framework for teaching linear regression concepts (Marasinghe, Duckworth, & Shin, 2004), or further topics for probability and inferential statistics using learnR (Stoudt, Scotina, & Luebke, 2022). Few address the linear regression concepts with which students in this course have historically struggled. Further, Shiny applets do not allow students to gain practice in coding and equalize the R programming experience of the students while simultaneously exploring the course concepts. The modules outlined in the remainder of the paper are designed to fill this unique role.

In: EM Jones (Ed.), *Fostering Learning of Statistics and Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
July 2023, Toronto, Canada. ©2023 ISI/IASE https://iase-web.org

OVERVIEW OF LEARNR MODULES

Table 1: Current series of learnR modules in development for commonly misunderstood linear regression concepts.

Module Title	Overview of Module Tasks	Motivation
1. Motivating the Regression Line of Best Fit	Display scatterplot, conditional means, and line of best fit for a given dataset, and connect mathematical notion of a slope of a line to the line of best fit displayed.	Safner (n.d.)
2. Interpretation and Use of Indicators and Interactions	Estimate simple linear relationships in subgroups and visualize on scatterplot. Compare difference in slopes/intercepts in simple models to estimated coefficients from a multiple linear model in which subgroup variable included.	Peterson & Ziegler (2021)
3. Impact of Violated Assumptions on Confidence Intervals	Execute a simulation study on the coverage of confidence intervals of coefficients under satisfied assumptions, and upon violation of each assumption of linear regression.	—
4. Understanding Multicollinearity	Generate data with varying levels of correlation between different predictors and visualize the changes in confidence interval widths/significance of coefficients.	Waskom (2021)
5. Visualizing Problematic Observations	Use scatterplots and delete-one models to visualize the changes leverage points, outliers, and influential points have on estimation in linear models.	Dudek (2016)


Motivating the Regression Line of Best Fit
Background: Help your friend understand her cleaning business
Section 1: Load and visualize the sample data
Section 2: The conditional distribution of Rooms by Crew Size
Section 3. The change in conditional means as X increases
Section 4: How is the line of best fit found?
Start Over
Try Again
Next Topic

Background: Help your friend understand her cleaning business

Suppose that your best friend started running a cleaning company a year ago, however, the company is not profitable yet. Your friend is worried about her business and suspicious that it might be due to some unproductive employees. Specifically, she is suspicious that when she sent more people to work, some of them did not work as efficiently as they should. Consequently, sending more people to work does not help clean more rooms and the company wasn't able to make profits in those cases. Your friend asks for your help in verifying her suspicions.

You asked your friend for the last year's records on the number of people your friend sent to work (Crews, X) and the number of rooms cleaned (Rooms, Y) to check if the more people your friend sent to work, the more rooms were cleaned. Unfortunately, she only had records for one month and not the whole year.

Which of the below is true about having only one month of data instead of the full year of data for this business?

● If we looked at a different month of data, we would see the same results as for this month.

○ Even if our sample of data is not representative of the full year, we can still obtain an unbiased estimate of the association.

○ The association between Crew size and Rooms cleaned we can see in this sample is not necessarily exactly the true association for the full year, but is a good guess.

○ It's not possible to get an estimate of association between Crew size and Rooms cleaned.

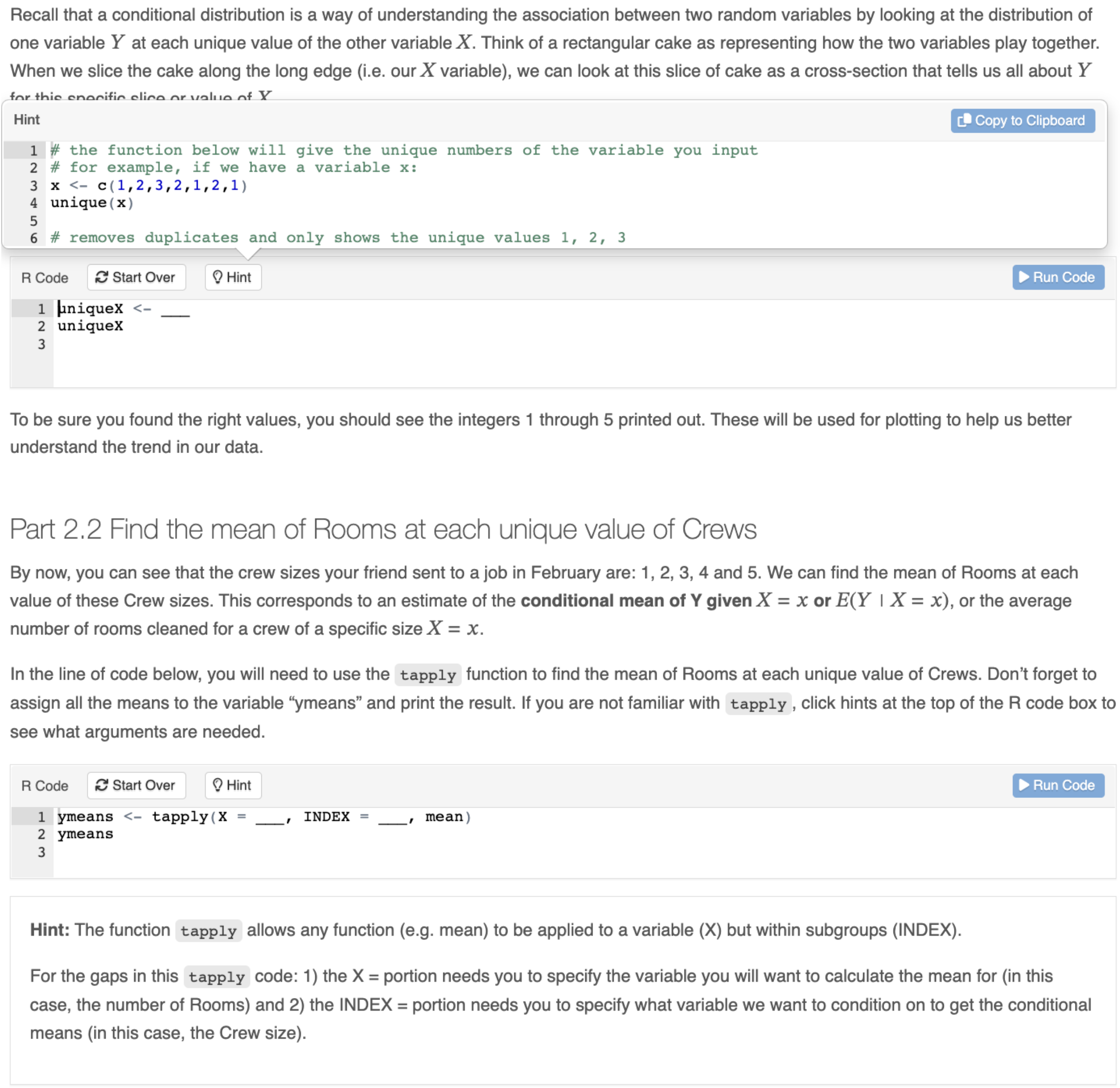
Incorrect

Each sample (month) of data taken from the full year will exhibit sampling variability and therefore may not give us identical results to the month our friend gave us.

Figure 1: Screenshot of the landing page for learnR Module 1: Motivating the Regression Line of Best Fit. Includes setup for the data to be used, and a knowledge check question on estimation, with feedback for each answer.

Table 1 summarizes the modules that have been developed or are in development as part of the first round of this project. These five areas were chosen as they are either concepts that past iterations of the course have highlighted as difficult concepts for students to explain or interpret correctly (e.g., modules 1, 2, and 3) or are concepts that would benefit from further illustration to gain a deeper understanding (e.g. modules 3, 4 and 5). All modules were adapted from existing R Markdown code used in previous iterations of the course in live-coding sessions (Çetinkaya-Rundel, 2021) during class to illustrate concepts and to see applications of the methodology to a dataset. Many of the original live coding frameworks were inspired by applets or the work of others (see Table 1).

Each learnR module features on its landing page an introduction or motivation to the module and a multiple-choice question meant to encourage recollection of a specific concept from a prerequisite course needed for the module. For example, Figure 1 showcases the landing page for Module 1, in which the students would work through the four sections of the module to discover through data exploration and visualization that a simple linear model not only estimates a linear association between two variables, but specifically estimates the change in conditional means arising from unit increases in the independent variable. In this instance, the knowledge needed to begin the module is an understanding of the concept of estimation, sampling variation, and the distinction between a sample and a population. Each knowledge check question allows for multiple attempts to be made to encourage students to use the feedback from incorrect answers to revise their understanding of the concepts before moving forward.



Recall that a conditional distribution is a way of understanding the association between two random variables by looking at the distribution of one variable Y at each unique value of the other variable X. Think of a rectangular cake as representing how the two variables play together. When we slice the cake along the long edge (i.e. our X variable), we can look at this slice of cake as a cross-section that tells us all about Y for this specific slice or value of X.

```
# the function below will give the unique numbers of the variable you input
# for example, if we have a variable x:
x <- c(1,2,3,2,1,2,1)
unique(x)

# removes duplicates and only shows the unique values 1, 2, 3
```

```
uniqueX <- ___
uniqueX

```

To be sure you found the right values, you should see the integers 1 through 5 printed out. These will be used for plotting to help us better understand the trend in our data.

Part 2.2 Find the mean of Rooms at each unique value of Crews

By now, you can see that the crew sizes your friend sent to a job in February are: 1, 2, 3, 4 and 5. We can find the mean of Rooms at each value of these Crew sizes. This corresponds to an estimate of the **conditional mean of Y given $X = x$ or $E(Y \mid X = x)$**, or the average number of rooms cleaned for a crew of a specific size $X = x$.

In the line of code below, you will need to use the `tapply` function to find the mean of Rooms at each unique value of Crews. Don't forget to assign all the means to the variable "ymeans" and print the result. If you are not familiar with `tapply`, click hints at the top of the R code box to see what arguments are needed.

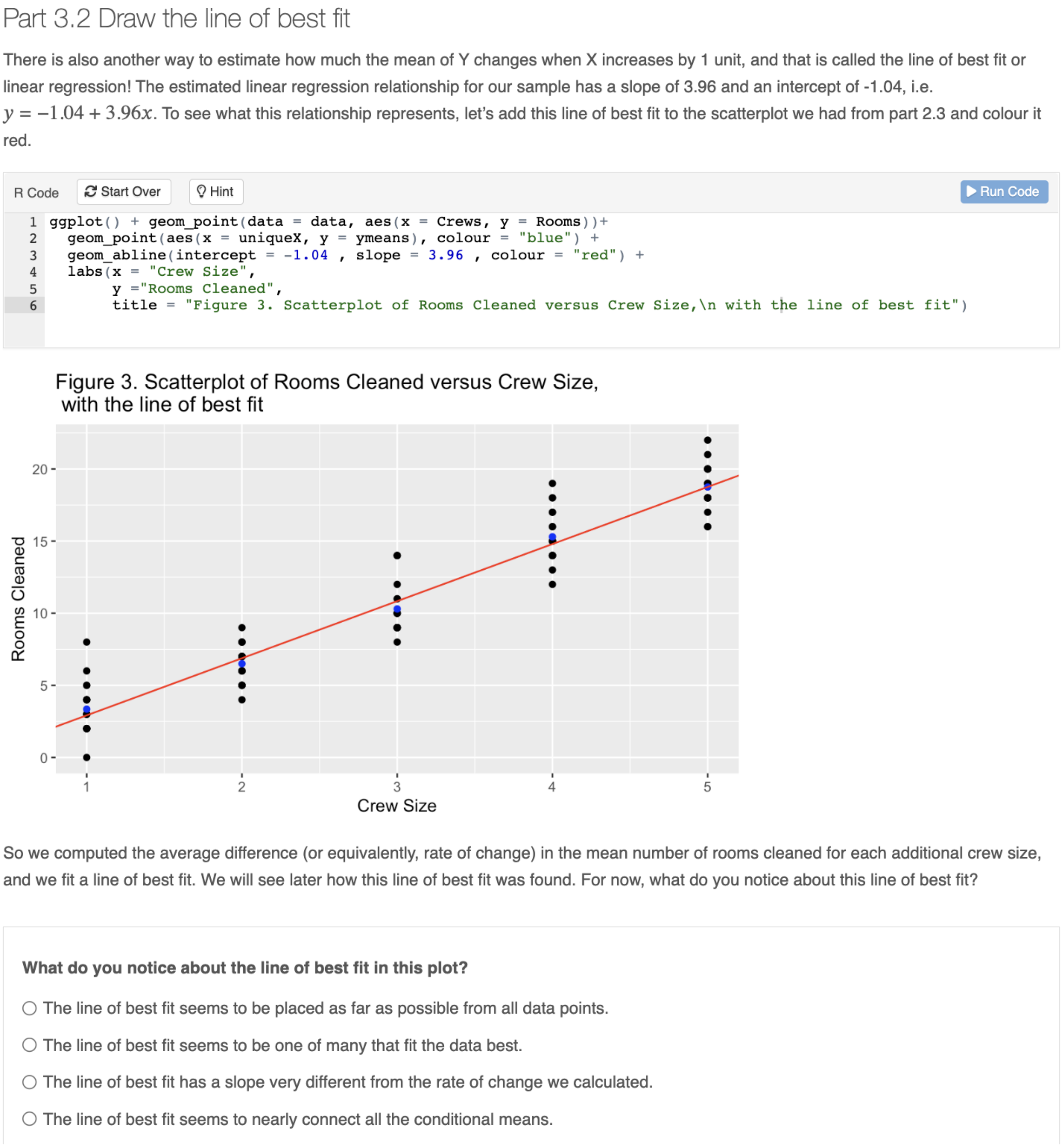
```
ymeans <- tapply(X = ___, INDEX = ___, mean)
ymeans

```

Hint: The function `tapply` allows any function (e.g. mean) to be applied to a variable (X) but within subgroups (INDEX).

For the gaps in this `tapply` code: 1) the X = portion needs you to specify the variable you will want to calculate the mean for (in this case, the number of Rooms) and 2) the INDEX = portion needs you to specify what variable we want to condition on to get the conditional means (in this case, the Crew size).

Figure 2: Screenshot of Module 1: Motivating the Regression Line of Best Fit displaying worksheet code chunks with two styles of hints for completing the requested code.

As students work through the module, they will have explanations of topics and tasks provided in the form of both examples and code chunks that have been pre-populated with template code and clear gaps students are asked to complete. Hints and solutions, as seen in Figure 2, can be provided to guide students through the worksheet. These hints can be in the form of an R chunk in which traditional code and syntax can be written and copied into their worksheet, or in the form of text. The text-based hint displays below the working code chunk, while the code-based hint displays above their working code chunk and overlaps instructions (see Figure 2). The advantage of the code-based hint is that multiple sequential hints can be included, allowing students to break down more complicated coding tasks into individual pieces, each with an associated hint to guide them towards an answer.



Part 3.2 Draw the line of best fit

There is also another way to estimate how much the mean of Y changes when X increases by 1 unit, and that is called the line of best fit or linear regression! The estimated linear regression relationship for our sample has a slope of 3.96 and an intercept of -1.04, i.e. $y = -1.04 + 3.96x$. To see what this relationship represents, let's add this line of best fit to the scatterplot we had from part 2.3 and colour it red.

```
ggplot() + geom_point(data = data, aes(x = Crews, y = Rooms))+
  geom_point(aes(x = uniqueX, y = ymeans), colour = "blue") +
  geom_abline(intercept = -1.04 , slope = 3.96 , colour = "red") +
  labs(x = "Crew Size",
       y ="Rooms Cleaned",
       title = "Figure 3. Scatterplot of Rooms Cleaned versus Crew Size,\n with the line of best fit")
```

Figure 3. Scatterplot of Rooms Cleaned versus Crew Size,
with the line of best fit
Rooms Cleaned
0
5
10
15
20
Crew Size
1
2
3
4
5

So we computed the average difference (or equivalently, rate of change) in the mean number of rooms cleaned for each additional crew size, and we fit a line of best fit. We will see later how this line of best fit was found. For now, what do you notice about this line of best fit?

What do you notice about the line of best fit in this plot?

○ The line of best fit seems to be placed as far as possible from all data points.

○ The line of best fit seems to be one of many that fit the data best.

○ The line of best fit has a slope very different from the rate of change we calculated.

○ The line of best fit seems to nearly connect all the conditional means.

Figure 3: Screenshot of the final coding task (shown as completed) in the Motivating the Regression Line of Best Fit learnR module.

Each module is meant to be completed prior to any formal discussion or instruction on each topic so as to build a visual and intuitive understanding of the concepts that would serve as a basis for understanding the theory. In the case of Module 1, students would not know yet how the line of best fit they are asked to draw (see Figure 3) is determined. But through discussion and estimation of conditional means, along with the plots they have created, it should be clear that this line of best fit connects these conditional means (more or less). The goal is that this process would illustrate why the correct interpretation of the linear regression coefficients involves the mean response since they have visually confirmed this concept in a sample.

DISCUSSION

The goal of this project was to create a series of low stakes R practice modules in the format of self-contained guided learnR worksheets to address the different levels of coding experience of students in this course. We anticipate that students lacking experience in R will benefit the most from these modules as they are intended to equalize coding experience across students, and in-class live-coding sessions have received feedback to this end. However, the modules have the potential to be valuable to all students in the course as they have been integrated into the course material with the intention of strengthening the connection between the theoretical results presented in lecture and a conceptual understanding of the methodology (Garfield, 1995). This was done not only to develop student intuition but also to further integrate programming as a tool for learning and understanding (Tucker et.al., 2022) and subsequently to reduce the cognitive load for students through these guided and partially complete worksheets (Stoudt, Scotina, & Luebke, 2022).

The learnR modules are slated to be introduced as a pilot in the Fall 2023 academic offering of the Methods of Data Analysis 1 course. They will form part of the guided pre-work (Talbert, 2017) alongside lecturette videos meant to be completed by the students at the start of each weekly module in a flipped environment. In the meantime, additional modules are being planned to complete a 10- to 12-week sequence of learnR modules that would accompany each module of the course. Formal assessment of the usefulness of these modules and to ensure that they satisfy the original goal of the project will be performed in the upcoming academic terms, pending ethics approval.

ACKNOWLEDGEMENTS

We wish to acknowledge our team of research assistants, Feiyu Li, Zhongyuan Liang, and Lijing Wei, who modified our existing course materials to create the learnR modules. We also wish to acknowledge the funding support we received from the Pedagogical Innovation and Experimentation fund from the Faculty of Arts and Science at the University of Toronto.

REFERENCES

Aden-Buie, G., Schloerke, B., Allaire, J., & Rossell Hayes, A. (2023). *LearnR: Interactive Tutorials for R*. https://rstudio.github.io/learnr/, https://github.com/rstudio/learnr.

Çetinkaya-Rundel, M. (2021, June 25 – July 1). *Going Live: Live Coding as an (Incredibly) Effective Tool for Teaching Programming* [conference presentation]. USCOTS 2021, virtual.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A. & Borges, B. (2021). *shiny: Web Application Framework for R.* R package version 1.7.1 https://CRAN.R-project.org/package=shiny.

Doi, J., Potter, G., Wong, J., Alcaraz, I., & Chi, P. (2016). Web Application Teaching Tools for Statistics Using R and Shiny. *Technology Innovations in Statistics Education*, *9*(1). http://dx.doi.org/10.5070/T591027492.

Dudek, B. (2016). *Visualization of 'Influence' in Regression*. R Shiny applet. https://shiny.rit.albany.edu/stat/outliers/.

Fawcett, L. (2018). Using Interactive Shiny Applications to Facilitate Research-Informed Learning and Teaching, *Journal of Statistics Education, 26*(1), 2-16.

Garfield, J. (1995). How Students Learn Statistics. *International Statistical Review, 63*(1), 25-34.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations, *European Journal of Epidemiology, 31*, 337-350.

Laviolette, M. (1994). Linear regression: The computer as a teaching tool. *Journal of Statistics Education, 2*(2), , DOI: 10.1080/10691898.1994.11910472 .

Marasinghe, M., Duckworth, W. M., & Shin, T-S. (2004). Tools for Teaching Regression Concepts Using Dynamic Graphics, *Journal of Statistics Education, 12*(2), , DOI: 10.1080/10691898.2004.11910733.

Nolan, D. & Temple Lang, D. (2010). Computing in the statistics curricula. *The American Statistician, 64*(2), 97-107.

Peterson, A. D. & Ziegler, L. (2021). Building a multiple linear regression model with LEGO brick data, *Journal of Statistics and Data Science Education, 29*(3), 297-303.

Safner, R. (n.d.). *Visualizing Linear Regression*. R Shiny applet. https://ryansafner.shinyapps.io/ols_estimation_by_min_sse/.

Stoudt, S., Scotina, A. D., & Luebke, K. (2022). Supporting statistics and data science education with learnR. *Technology Innovations in Statistics Education, 14*(1), , http://dx.doi.org/10.5070/T514156264

Talbert, R. (2017). *Flipped Learning: A Guide for Higher Education Faculty.* Stylus Publishing, LLC.

Tucker, M. C., Shaw, S. T., Son, J. Y., & Stigler, J. W. (2022). Teaching statistics and data analysis with R, *Journal of Statistics and Data Science Education, 31*(1), 18-32.

Wang, S. L., Zhang, A. Y., Messer, S., Wiesner, A. & Pearl, D. K. (2021). Student-Developed Shiny Applications for Teaching Statistics**,** *Journal of Statistics and Data Science Education, 29*(3), 218-227, DOI: 10.1080/26939169.2021.1995545.

Waskom, M. (2021). *Multicollinearity in Multiple Regression*. R Shiny applet. https://gallery.shinyapps.io/collinearity/.