9th International Conference on Higher Education Advances (HEAd'23)
Universitat Politècnica de València, València, 2023
DOI: http://dx.doi.org/10.4995/HEAd23.2023.16045

Online Repository for Facilitating Teaching and Learning of Undergraduate Statistical Modeling Tools

Qing Wang[1], Xizhen Cai[2]
[1]Department of Mathematics, Wellesley College, USA, [2]Department of Mathematics and Statistics, Williams College, USA.

Abstract

This paper presents an online repository created for facilitating teaching and learning statistical modeling tools at the undergraduate level. Statistical models and modeling techniques have always been considered the backbone in data analysis and statistical learning. Over the past decade, teaching of such topics has also gained an increasing attention in the undergraduate statistics and data science curricula. The developed online repository aims at improving the teaching and learning of statistical modeling tools in various undergraduate statistical modeling courses. We present the four core components of the repository, showcase some of its functionalities, and exhibit available resources online. Through our informal assessments after incorporating the online repository into our classrooms, our students seemed to have a uniformly better understanding of the related concepts and methods, which was reflected during the in-class discussions as well as in the subsequent tests in the courses.

Keywords: *Interactive web applications; linear regression; online repository; statistical models.*

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

1. Introduction

Statistical models play a vital role in data analysis and have been applied to an abundance of practical problems (Cox, 1990). Teaching statistical modeling tools with an emphasis on regression analysis has always been an essential part of the undergraduate statistics education. In the American Statistical Association report on Curriculum Guidelines for Undergraduate Programs in Statistical Science (Horton et al., 2014), statistical modeling, including regression analysis, has been identified as one of the core topics in undergraduate statistics curriculum. Moreover, in the current Guidelines for Assessment and Instruction in Statistics Education (Carver et al., 2016), learning and using statistical models is listed as one of the important learning goals, even for introductory statistics. Here the goal is not only to ensure students' mastery of the general knowledge of statistical models, but also to facilitate the development of students' statistical maturity, which serves as a crucial building block for them to acquire some more advanced topics in their statistics curriculum. Given the rapid development and growing interest of data science over the past decade, regression analysis has also been universally recognized and widely adopted as a central component in the data science curriculum (De Veaux et al., 2017; Donoghue et al., 2020).

Even though students often get exposed to the topic of (simple) linear regression in a first course in statistics, such as introductory statistics, taking some more advanced modeling courses is necessary to fill them in the technical details, expose them to other modeling techniques, introduce them to inferential procedures of different models, and improve their computational skills through the use of certain statistical software or programming languages. These more advanced courses in statistical modeling are especially vital for students who are considering pursuing a graduate degree in statistics or a related field. Most undergraduate programs in the US offer applied statistics modeling courses in addition to introductory statistics. Although slightly different in content, such courses often require students to have certain mathematical background, and their prerequisites usually include calculus or linear algebra. The mathematical components of such courses further solidify students' understanding of the fundamental concepts and properties of various statistical models. However, at the same time, they may unfortunately place a barrier in the learning process for students with a relatively weaker quantitative background, especially for under-represented minorities, first-generation college students, students of color, and indigenous students. Consequently, those students may feel discouraged from further pursing statistics or a field in STEM in general.

When teaching abstract or challenging concepts, it is a common practice for statistics educators to take advantage of visualization in order to break each concept into several pieces of information for ease of understanding. It is also crucial to connect each concept with real-world applications through concrete yet interesting examples. Some recent pedagogical efforts on this front often involve active learning (Gelman and Nolan, 2017; Green et al., 2018; Cai and Wang, 2020) and interactive web applications (Tintle et al., 2020). Although the use of those applications saves students from undertaking the hard coding of simulations, there are only a small number of such online applications available for teaching or learning statistical modeling at the undergraduate level.

Another important learning goal we often emphasize in our classrooms is the introduction and clarification of formal inferential procedures of statistical models. Since the true population model is not observable, one of the critical steps in the modeling process is to draw conclusions of the true model based on its sample estimate. This procedure is referred to as statistical inference. Due to a lack of mathematical preparations for some students, one of the challenges in teaching such concepts is to convey the fundamental idea behind statistical inference and explain intuitively why a certain distribution is assumed in each of the inferential procedures. To overcome this hurdle, one popular approach nowadays is to adopt the so-called simulation-based inferential procedures (Tintle et al., 2018; Hildreth et al., 2018). There are a number of statistics textbooks with an emphasise on simulation-based methods, especially for introductory statistics (Tintle et al., 2020). However, there seems to be insufficient resources available that are thoughtfully designed for more advanced modeling topics in statistics.

Given all the challenges discussed above, we believe that there is a pressing need to develop an online repository to engage students in active learning, help students build connections and identify differences between various modeling tools, as well as to incorporate simulation-based inference into the teaching of statistical modeling at the undergraduate level. In what follows, we will detail the components of our designed online repository.

2. Components of the Proposed Repository

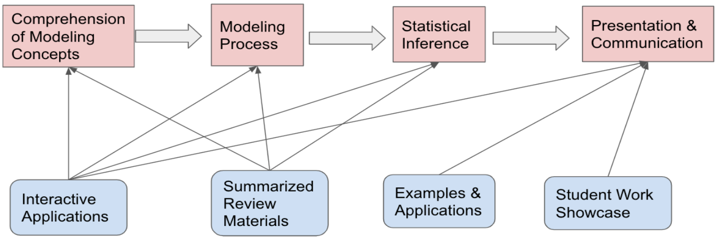
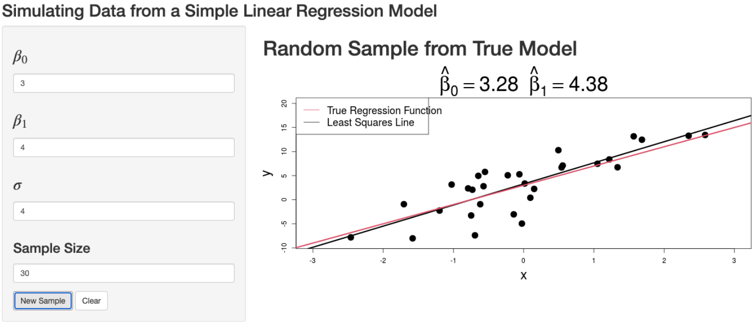


Figure 1: Overview of the four components in the repository.

Figure 1 presents the four core components of our designed online repository, namely, Comprehension of Modeling Concepts, Modeling Process, Statistical Inference, and Presentation and Communication. In the following, we will provide more details for each of these four components.

2.1. Comprehension of Modeling Concepts

The objective of the first component, Comprehension of Modeling Concepts, is to facilitate the introduction of statistical models, discuss important elements and assumptions of each model, and demonstrate the differences and similarities between a statistical model that describes the relationship of variables in the target population and the corresponding fitted model estimated based on a sample of data drawn from the population.



Simulating Data from a Simple Linear Regression Model

β_0: 3

β_1: 4

σ: 4

Sample Size: 30

Figure 2. Example of an interactive application for simple linear regression.

Interactive web-based applications can well meet this objective. Over the past several years, we have been working on developing a number of interactive applications for our statistics modeling courses, using the R Shiny package (R Core Team, 2022). For example, Figure 2 showcases one application we created when first introducing the simple linear regression model in our classes. In this application, users can input values of the true parameters of the model, and then generate a random sample of data based on the underlying true model. Given the simulated data set, the estimated regression coefficients, $\widehat{\boldsymbol{\beta_0}}$ and $\widehat{\boldsymbol{\beta_1}}$, obtained by the Ordinary Least Squares Criterion are displayed on top of the scatterplot. Moreover, both the true regression line and the estimated regression line are presented in the plot to clarify the difference between the population relationship and its corresponding sample estimate.

A set of thoughtfully designed interactive applications are available in the first component of our online repository, ranging from topics on multiple linear regression, analysis of variance (ANOVA) models, logistic regression, receiver operating characteristic (ROC) curve, to support vector classifiers. We plan to continue our efforts in this direction, and develop more interactive tools that are useful for teaching and learning more challenging concepts in statistical modeling.

2.2. Modeling Process

The second component of our repository focuses on providing more details of the modeling process of a given model. It provides students with a road map of mastering different modeling tools in a more systematic way and aims at enhancing students' understanding of various statistical modeling techniques.

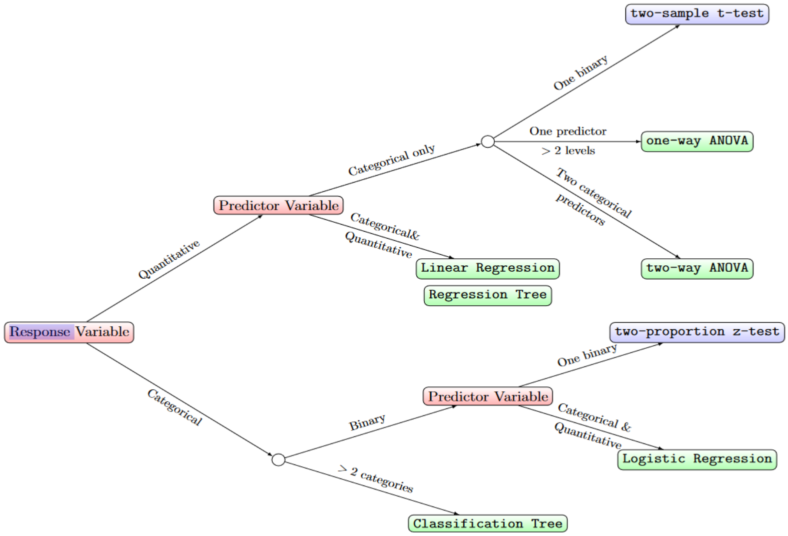


Figure 3. Diagram used to identify a proper statistical model.

The Modeling Process component includes an overview of various types of statistical models, typically introduced at the undergraduate level. Students start with a diagram, as displayed in Figure 3. By following the diagram, students are guided through the process of selecting an appropriate model given a real data set and some research question of interest. In addition, the diagram is interactive, i.e. each model is linked to a module, where students can further explore and review some relevant concepts for the selected model.

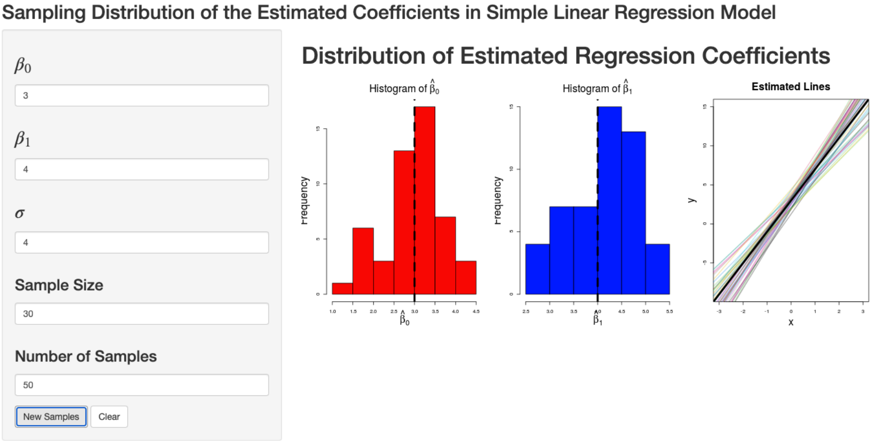
In order to illustrate the order, connection, and cyclic nature between steps of the modeling process, we designed another diagram to reflect the modeling steps in a sequential manner. The diagram also offers a head-to-toe comparison between different models. With the help of the hyperlink feature, when selecting one of the models the user will be directed to a new web page that presents the comparison across various models. For example, Table 1 presents an example that compares multiple linear regression model, two-way ANOVA model, and multiple logistic regression model. By comparing these models side by side, it is easy for students to recognize that linear regression and ANOVA models share similar model assumptions, while linear regression and logistic regression have more in common in their formulations, as both relate the (transformed) mean response with a linear combination of predictors.

Table 1. Sample Comparisons of Different Models.

	Linear Regression	ANOVA	Logistic Regression
Model	$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$ • model linear relationship; predict the value of Y • Y is quantitative • X_j is either quantitative or categorical (e.g. indicator) • error: $\epsilon \sim N(0, \sigma)$, independent • k=number of predictors	$Y = \mu_j + \epsilon = \mu + \alpha_i + \beta_j + \epsilon$ • comparing means across combinations of factor levels • Y is quantitative • factors are categorical • error: $\epsilon \sim N(0, \sigma)$, independent • I, J=number of levels for each factor	$\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ • adjusted relationship that predicts the probability of $Y = 1$ • Y binary; $\pi = P(Y = 1)$ • X_j's are either categorical or quantitative • no explicit error term • k=number of predictors

2.3. Statistical Inference

Over the past decade, simulation-based inference has gained much attention and become a popular approach for introducing statistical inferential tools to undergraduate students. However, most existing work and interactive applications were built for introductory statistics. To bridge the gap, we construct the following modules in order under the Statistical Inference section in our repository: Module 1. Sampling distribution and computing p-value for a hypothesis test of the population mean based on a simulation; Module 2. Commonly used distributions and their properties; Module 3. Sampling variation and inference for model parameters and predictions; Module 4. Other simulation-based methods and their applications to statistical models.


Sampling Distribution of the Estimated Coefficients in Simple Linear Regression Model

β_0: 3

β_1: 4

σ: 4

Sample Size: 30

Number of Samples: 50

Distribution of Estimated Regression Coefficients

Figure 4. Sample Applications of Statistical Inference.

As an example, Figure 4 presents one of the developed interactive applications for Module 3. It visualizes the variability of the estimated coefficients in a simple linear regression model, which also echoes back to the application displayed in Figure 2. By increasing the sample size as well as the number of samples drawn from the true model, the histogram will display a distributional shape that is approaching a bell-shaped curve. This observation naturally guides students to the topic of hypothesis testing of model parameters.

2.4. Presentation and Communication

One of the important learning goals for a statistical modeling course is to improve students' communication skill in delivering statistical results to a broad audience across different disciplines. This skill is usually honed by having students work on a course project, give a oral presentation, and/or complete a written project report. We have noticed from our previous teaching experience that some students have limited prior experience in working on data analysis projects, such as students who are not traditionally well represented in the STEM fields. Hence, they often face more challenges during the completion of such class assignments.

To overcome this issue, we offer useful resources in our repository on how to present statistical results and write a professional project report. We include useful tips on best practices and related literature. Moreover, with our former students' permission, we created a library that showcases past students' work (e.g., papers or presentation videos).

3. Summary and Conclusion

In this paper, we discussed an online repository designed for teaching and learning statistical modeling tools at the undergraduate level. In particular, we presented the four core components of the repository and showcased some examples among its functionalities. We anticipate that the four components in our repository will complement each other to engage students in the classroom, facilitate their understanding, and make the learning of statistical modeling more enjoyable and effective. Through our informal assessments when integrating some of the applications in the repository to our own classes, we received uniformly positive feedback from our students. Furthermore, by incorporating the online repository over the past semesters, our students seemed to have a generally better understanding of the related concepts and methods, which was reflected during the in-class discussions as well as in the subsequent tests of the courses. We plan to continue our efforts in this direction, enriching the available tools and web applications in the online repository. In addition, we hope to make the respository publicly available in the near future so that it can benefit the broader statistics and data science education community across the globe. As a future project, we plan to conduct formal analysis on the feedback received from students and educators who have experimented the developed online repository.

References

Cai, X. & Wang, Q. (2020). Educational tool and active-learning class activity for teaching agglomerative hierarchical clustering. *Journal of Statistics Education*, 1–9.

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Roswell, G. H., Velleman, P., Witmer, J., & Wood, B. (2016). Guidelines for assessment and instruction in statistics education (GAISE) college report.

Cox, D. (1990). Role of models in statistical analysis. *Statistical Science*, 5(2):169–174.

De Veaux, R., Agarwal, et al. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, 4:15–30.

Donoghue, T., Voytek, B., & Ellis, S. E. (2020). Teaching creative and practical data science at scale. *Journal of Statistics Education*, 1–22.

Gelman, A. & Nolan, D. (2017). Teaching statistics: A bag of tricks. Oxford University Press.

Green, L. B., McCormick, N., McDaniel, S., Rowell, G. H., & Strayer, J. (2018). Implementing active learning department wide: a course community for a culture change. *Journal of Statistics Education*, 26(3):190–196.

Hildreth, L., Robison-Cox, J., & Schmidt, J. (2018). Comparing student success and understanding in introductory statistics under consensus and simulation-based curricula. *Statistics Education Research Journal*, 17(1).

Horton, N., Chance, B., Cohen, S., Grimshaw, S., Hardin, J., Hesterberg, T., Hoerl, R., Malone, C., Nichols, R., & Nolan, D. (2014). Curriculum guidelines for undergraduate programs in statistical science. *American Statistical Association*.

R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Tintle, N., Chance, B. L., Cobb, G. W., Rossman, A., Roy, S., Swanson, T., &VanderStoep, J. (2020). *Introduction to statistical investigations*. John Wiley & Sons.

Tintle, N., Clark, J., Fischer, K., Chance, B., Cobb, G., Roy, S., Swanson, T, & VanderStoep, J. (2018). Assessing the association between precourse metrics of student preparation and student performance in introductory statistics: Results from early data on simulation-based inference vs. nonsimulation-based inference. *Journal of Statistics Education*, 26(2):103–109.