DEVELOPING STATISTICAL INTUITION THROUGH SIMULATION STUDIES IN A SECOND STATISTICS COURSE

Aimee Schwab-McCoy
Institute of Technology, Sligo
Schwab-McCoy.Aimee@itsligo.ie
Creighton University
AimeeSchwab-McCoy@creighton.edu

Changes in early statistics and data science education have a ripple effect across the curriculum: as the introductory courses are modernized, the later courses must too. The class described in this paper is a second-semester statistical modeling course with a modern, post-data science flair. Regression models are introduced separately (multiple regression, Poisson regression, logistic regression) before being synergized as the generalized linear model (GLM). In this class, learners studied the patterns and behaviors of these models through targeted labs leaning heavily on simulated data. This course emphasizes the development of statistical intuition through hands-on learning experiences, rather than a set of rules for each situation. The addition of introductory data science, coupled with an increased emphasis on statistical computing in the first statistics course, make realistic simulation studies possible earlier in the curriculum.

INTRODUCTION

Modern introductory data science courses provide students with a computational foundation beyond traditional introductory courses. Even in the first statistics course, many students are now using R to support their coursework rather than applet or "point-and-click" software systems. As the introductory courses change and grow -- what comes next? The class described in this paper is a second-semester statistical modeling course with a modern, post-data science flair. Regression models are introduced separately (multiple regression, Poisson regression, logistic regression) before being *generalized* as the generalized linear model (GLM). Students study the patterns and behaviors of these models and develop statistical intuition about model performance through simulation studies. These labs are greatly enhanced through the strong foundation in statistical programming, reproducible research, and data visualization provided by an introductory data science course.

LITERATURE REVIEW

In the past decade, student interest and enrollments in statistics and data science programs has skyrocketed. As a result, statistics educators have found themselves teaching more advanced statistical content sooner. For example, generalized linear mixed models, or GLMMs, were once considered a graduate-level topic. However, their wide flexibility and applicability make them an ideal topic from which to build a second undergraduate statistics course.

Effective use of GLMMs requires a high level of statistical intuition. The theoretical and practical aspects of these models, such as scope of inference, understanding the link scale v. data scale, design considerations, and incorporation of random effects mean that many of the "rules of thumb" students learn in introductory statistics go out the window. So how can statistics instructors help learners develop the intuition needed? As with any new subject, real (or realistic) and authentic experiences are key. Simulation studies and hands-on model testing can provide these experiences, and mimic modern statistical research workflows.

Even though there is now a general consensus in statistics education community on what belongs in the introductory statistics course (Garfield et al., 2002) and the guiding principles for instruction (Carver et al., 2016), there is more variation in what comes next. Generally, a "second semester" statistics course includes topics such as multiple regression, analysis of variance, and design of experiments (Blades et al., 2015; McGaughey et al., 2018), but the level of coverage and rigor can vary. As the introductory statistics course changes, and the introductory data science course grows, students will be coming to later courses with stronger backgrounds.

Recently, a group of faculty members at Loyola Marymount University developed a set of learning objectives for developing undergraduate data acumen, called the Undergraduate Data Pathways (Bargagliotti et al. 2020). The thirteen final learning objectives covered elements of the entire

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. iase-web.org/Conference_Proceedings.php

undergraduate curriculum, not just a single course. These and other sets of learning outcomes such as the Park City Math Institute (DeVeaux et al., 2017) can provide a "road map" for what comes after the first statistics course in the age of data science. Loy, Kuiper, and Chihara shared a set of hands-on lab activities that could be used to bolster data science concepts such as visualization, data wrangling, and databases in a traditional statistics course or curriculum (Loy et al., 2019). These data science concepts are integral to simulation studies in statistical research, and student familiarity with them sooner means that meaningful simulations can be introduced at the undergraduate level.

Simulation studies can be effective tools at the post-secondary level to develop statistical competencies and explore nuances of new and complex models. Several comparisons of simulation-based introductory courses against traditional courses found that teaching via simulation was associated with improved learning outcomes related to statistical inference (Maurer and Lock, 2016; Hildreth et al., 2018). However, most of the simulation-based curriculums focus on applets or web-based approaches, and not simulation studies as we would think of them in statistical research. With increased programming and data knowledge, upper-level courses in statistics are well-suited to take that next step toward learning through simulation.

Learning through simulation can be an effective teaching tool at the undergraduate level. Students often ask when examining residual plots, or evaluating model assumptions, **what should this plot look like**? What values would we expect if a model is a "good fit"? With simulation studies, we can actively explore with our students what happens when assumptions about our model are in fact met, and what happens when they are not.

MTH 362: STATISTICAL MODELING

MTH 362: Statistical Modeling is a second semester statistics course offered as part of the Data Science Major and Minor at Creighton University, a private Jesuit Catholic institution in the Midwestern United States. Introductory Statistics is a pre-requisite for this course, and most students take an introductory course focused on the health sciences (due to Creighton's large pre-medical student population). In previous semesters, 50-75% of the students have also completed Creighton's Introduction to Data Science course. The course presents the usual conceptual challenges and must also bridge the gap between two sets of students: those who are planning to complete the Data Science Major or Minor, and those in other fields seeking to improve their statistical background.

The learning outcomes in MTH 362 can be split into two components, described in Table 1. To meet these learning outcomes, a series of simulation labs has been incorporated into the course. Each lab is explicitly designed around using a simulation study to understand a new concept related to model fitting and evaluation. Topics include selection bias, residual plots for non-normal data, model diagnostic behavior, and correlation structures. In Spring 2021, labs took place approximately every other week, and like the rest of the course, were completely online due to the pandemic.

Table 1. Learning outcomes for MTH 362: Statistical Modeling

Subcategory	Outcome
Statistics	1. Understand the basic components and theoretical assumptions of a (generalized) linear (mixed) model 2. Consider whether theoretical assumptions of a model are (not) a good fit to reality 3. Create a reasonable statistical model based on both characteristics of the response and experimental design considerations, as well as the relevant research objectives of a study
Computation	1. Translate a statistical model into R code using base statistical packages and "lme4" (Bates et al., 2015) 2. Generate predictions, residual plots, fitted value plots, and summaries of model 3. Write professional, reproducible analyses using R and RMarkdown **4. Use simulated data to understand a model's performance when conditions are "met" and when they are "violated"**

LAB TOPICS AND STRUCTURE

Each lab was designed to take about 60 minutes of a 75-minute class period, leaving time for discussion. Since the course took place entirely online, labs were completed in Zoom Breakout Rooms with groups of 3-4. Complete summaries of each lab are in Table 2. Initially, labs focused more on the computation and reproducible workflow aspects of the course. This allowed students entering from multiple introductory statistics courses, or that had not taken introductory data science, a chance to "catch up".

Each of the labs focused on a particular aspect of interpreting residual plots and model fit from a (generalized) linear (mixed) model. Throughout the semester, students worked up to a unified approach to modeling through the generalized linear mixed model. To effectively use GLMMs, statisticians need to develop a strong sense of intuition, especially when diagnosing potential issues in a model. Changes at each level of the GLMM, such as categorical explanatory variables instead of numeric, or a negative binomial model instead of a Poisson, often manifest in different "correct" patterns in a residual plot or "best" values of a diagnostic. Rather than discuss each model separately, simulation enabled students to compare what should happen when a model is "correct" compared to what happens in reality.

Sample simulation activity: residual plots in Poisson models

Lab 2: Model Diagnostics was the first foray into exploring model performance, and diagnosing whether a model was a "good fit". In this lab, students were given examples of two models: a "naïve" model and a "better" model. Each group simulated their own data under one of several violated conditions, then reported back to the entire class. In one instance, two explanatory variables, $X_1 \sim Uniform(0, 1)$ and $X_2 \sim Binomial(n = 1, p = 0.6)$ were used to generate a Poisson distributed response with 100 observations, $Y \sim Poisson(\ln(\lambda) = 1 + 2X_1 + X_2)$.

Figure 1 shows some sample findings. Based on the fit statistics, like multiple R-squared, there is no obvious problem with the linear model. However, the residual plots show clear curvature and heteroscedasticity, and are vastly improved by switching to the Poisson model. Each group explored their own changes to the correct data generating model, such as adding interaction terms, non-significant variables, or quadratic terms, and explored how this manifested in the correct Poisson and incorrect linear models. This example also gives an early introduction to the idea of scale in a generalized linear model – specifically the difference between the data scale and link scale as shown by the different axes in the fitted value plots.

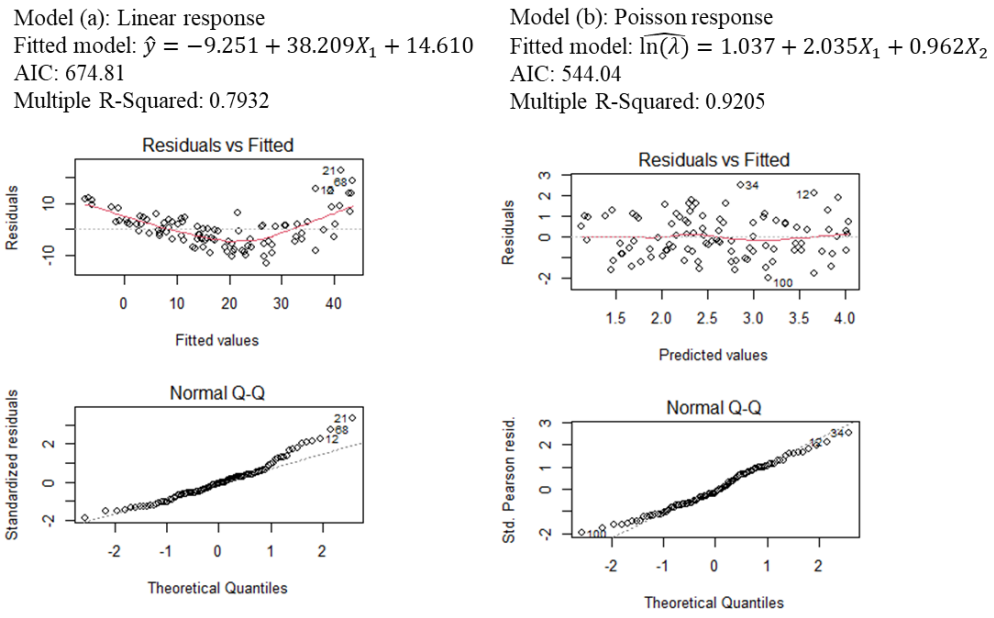


Figure 1. Sample simulation from Lab 2: Model Diagnostics

Table 2. Descriptions of each lab activity and placement during a 15-week semester

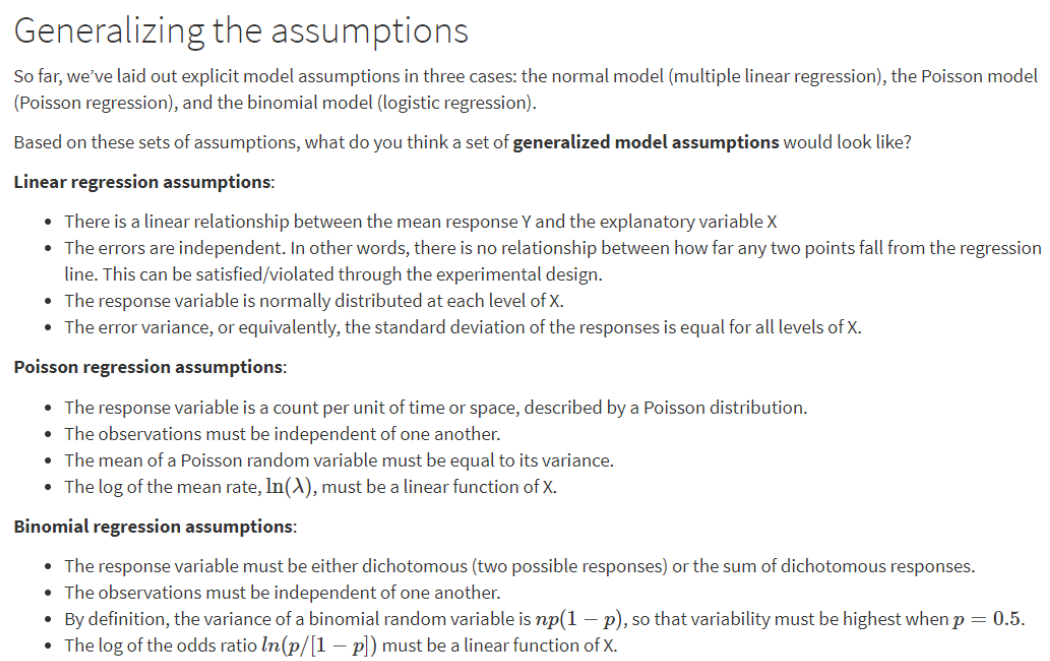
Lab	Week	Description
Lab 1: Interaction Terms	2	This lab explored the addition of categorical variables to a linear model using interaction terms. Students demonstrated that interaction terms allow the slope and intercept of a model to change from one group to another and can be used to compare the relationships between groups.
Lab 2: Model Diagnostics	3	In this exercise, students simulated data from a known population model, and generated residual plots to evaluate the *true* model's fit. Once that was completed, each group was assigned a different misspecification, such as adding a non-linear term to the true model, removing an interaction term, incorporating a Poisson response, or adding a non-constant variance term. Each group then created residual plots for the misspecified model and presented their findings to the group using Google Slides. This lab allowed students to see concrete examples of assumption violations manifested through residual plots and was intended to begin building intuition about when a residual plot doesn't "look right".
Lab 3: Exploring Poisson Regression Models	5	This lab was similar in structure to Lab 2, with the addition of a Poisson-distributed response. Rather than introduce a new set of rules of thumb, we simulated Poisson data under a variety of parameter combinations and examined the resulting residual plots from the *true* model. We then used these to look for similarities and identify some expected patterns with Poisson data.
Lab 4: Exploring Binomial Models	9	This lab had similar structure and goals to Lab 3, however the *true* model was a binomial GLM, or logistic regression. This lab was completed after a review of probability distributions and an overview of the structure of a GLM.
Lab 5: The Missing Link (Function)	10	In this lab, we used two case studies to explore the behavior of proportional data under varying link functions. In one data set, the choice of link function made a meaningful difference in the performance of the model. For the other data set, all models were equally "unsuccessful.
Lab 6: Don't Mix Up Your Fixed and Random Effects	12	In the final simulation-driven lab, students were presented with six experimental design scenarios. For each one, students identified the treatment and topographical structure of the design, the appropriate type of effect for each variable (fixed or random), and a suitable response distribution. Then, students translated that to a reasonable model formulation in R. During the next class period, students tested their proposed models on simulated data.

DISCUSSION

Introducing simulation-based labs to target model diagnostics in a second statistics course enriched the learning experience. Students were able to work together, even online, to explore difficult concepts in a hands-on way. For example, one topic that students often struggle with is comparing models using selection criterion such as AIC (Akaike's Information Criterion). Direct comparison of AIC across different response distributions, such as a normal response to a binomial response, is invalid due to theoretical differences in the structure of the likelihood function. To motivate the theoretical reasonings to an undergraduate audience without a background in calculus or mathematical statistics would be difficult to say the least! However, simulating data and fitting two competing models is a more tangible experience, shows the same results (the normal distribution has the lower

AIC regardless of whether it is the "correct" model), and is more memorable for students than the theoretical justification.

Another benefit of the simulation labs was the ability to demystify the R output, especially for more complex models such as mixed effects models. Without a deep exploration of the documentation for the various mixed models and GLMMs in R, and the various packages used to implement them, it's difficult to know what exactly is represented in the output. By simulating data with known variance components, students were better able to connect each part of the output back to the data generation process. Students also gained a better understanding of the functional components of a GLMM. Working through the simulation process explicitly demonstrated how the model believes the data has been generated. Not only that, but repeated investigation eventually allowed for generalization. For example, Figure 2 shows a prompt taken from Lab 4. After students had explored linear regression, Poisson regression, and logistic regression in detail, they were asked to write their own *generalized* set of assumptions for a generalized linear model. Since students had worked so closely with each of the three models in previous assignments, they were able to construct thoughtful sets of generalized assumptions for a new model.



Generalizing the assumptions

So far, we've laid out explicit model assumptions in three cases: the normal model (multiple linear regression), the Poisson model (Poisson regression), and the binomial model (logistic regression).

Based on these sets of assumptions, what do you think a set of **generalized model assumptions** would look like?

Linear regression assumptions:

- There is a linear relationship between the mean response Y and the explanatory variable X
- The errors are independent. In other words, there is no relationship between how far any two points fall from the regression line. This can be satisfied/violated through the experimental design.
- The response variable is normally distributed at each level of X.
- The error variance, or equivalently, the standard deviation of the responses is equal for all levels of X.

Poisson regression assumptions:

- The response variable is a count per unit of time or space, described by a Poisson distribution.
- The observations must be independent of one another.
- The mean of a Poisson random variable must be equal to its variance.
- The log of the mean rate, $\ln(\lambda)$, must be a linear function of X.

Binomial regression assumptions:

- The response variable must be either dichotomous (two possible responses) or the sum of dichotomous responses.
- The observations must be independent of one another.
- By definition, the variance of a binomial random variable is $np(1-p)$, so that variability must be highest when $p = 0.5$.
- The log of the odds ratio $ln(p/[1-p])$ must be a linear function of X.

Figure 2. Discussion prompt from Lab 4

Finally, completing simulation studies in small groups and reporting back encouraged students to accept natural variability in both simulated data and parameter estimates. Due to the inherent randomness of simulated data all student groups obtained slightly different results. In this particular semester, students shared their findings in online documents such as Google Slides or Google Docs, so they were able to see the differences in the online environment. In fact, Google Slides worked so well that the author would recommend using it in face-to-face classes as well.

One of the major limitations to implementing advanced simulation-based labs are the expanded prerequisites. Neither introductory statistics nor introductory data science is enough on its own to prepare students for such a course. For institutions with multiple pathways, multiple instructors, or both, ensuring all students have sufficient background may be a challenge. To address this, simulation labs should be scaffolded with respect to both difficulty and level of independence expected of students. Relying on learning through simulation also means additional technical challenges: more code means more troubleshooting. In general, instructors should try to mimic the student's environment as much as possible. This could mean maintaining a separate user profile on your computer for teaching or updating already installed R packages to have the same versions available as students.

GLMM theory can be intimidating, especially the structure of the model and the link function. Working through some examples of finding the canonical link of a distribution can help, but there is also a need to balance the mathematical prerequisites with the course material. One possible way to address this is to set aside 1-2 days early in the semester to discuss probability distributions. At Creighton, some introductory statistics courses cover the usual probability distributions, and others only cover the normal distribution. Spending extra time early in the semester can help get students that are new to probability theory up to speed, as well as provides a refresher for experienced students.

In a time-restricted setting, students shouldn't reinvent the wheel. The goal was to use simulation as a tool for increasing understanding, not as the final result. Providing students with sample code took the cognitive focus away from writing the code and shifts it to the results. As the semester progressed, students got less and less code as a starting point, but in the beginning almost no new lines of code were written during the lab times. Lessons learned from the labs should extend beyond the lab time. Students periodically reflected on their findings in weekly quizzes, and applied what they learned about model performance and interpreting results on their weekly homework assignments.

With the rise of data science courses, and the increase in computing acumen developed in introductory statistics classes, statistics faculty can incorporate more meaningful computing experiences earlier in the curriculum. Doing so can increase student participation and engagement, further develop computing acumen, and foster statistical intuition.

RESOURCES

All labs and data sets are available at: https://github.com/aimeeschwab-mccoy/IASE2021_Labs. Instructor solution guides are available on request.

REFERENCES

Bargagliotti, A., Binder, W., Blakesley, L., Eusufzai, Z., Fitzpatrick, B., Ford, M., Huchting, K., Larson, S., Miric, N., Rovetti, R., Seal, K., & Zachariah, T. (2020). Undergraduate Learning Outcomes for Achieving Data Acumen. *Journal of Statistics Education*, *28*(2), 197–211. https://doi.org/10.1080/10691898.2020.1776653

Bates, D., Maechler, M., Bolker, B., & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48. doi:10.18637/jss.v067.i01.

Blades, N. J., Schaalje, G. B., & Christensen, W. F. (2015). The Second Course in Statistics: Design and Analysis of Experiments? *The American Statistician*. https://doi.org/10.1080/00031305.2015.1086437

Christou, N. (2021). Spatial Data in Undergraduate Statistics Curriculum. *Journal of Statistics and Data Science Education*, *29*(1), 27–38. https://doi.org/10.1080/10691898.2020.1844104

DeVeaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., et al. (2017), "Curriculum Guidelines for Undergraduate Programs in Data Science," *Annual Review of Statistics and Its Application*, 4.

Hildreth, L. A., 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), 103–120.

Loy, A., Kuiper, S., & Chihara, L. (2019). Supporting Data Science in the Statistics Curriculum. *Journal of Statistics Education*, *27*(1), 2–11. https://doi.org/10.1080/10691898.2018.1564638

Maurer, K., & Lock, D. (2016). Comparison of Learning Outcomes for Simulation-Based and Traditional Inference Curricula in a Designed Educational Experiment. *Technology Innovations in Statistics Education*, *9*(1). https://doi.org/10.5811/westjem.2011.5.6700

McGaughey, K., Chance, B., Tintle, N., Roy, S., Swanson, T., & VanderStoep, J. (2018). Finding Meaning in a Multivariable World: A Conceptual Approach to an Algebra-Based Second Course in Statistics. *Proceedings from the 10th International Conference on Teaching Statistics*.

Woodard, V., Lee, H., & Woodard, R. (2020). Writing Assignments to Assess Statistical Thinking. *Journal of Statistics Education*, *28*(1). https://doi.org/10.1080/10691898.2019.1696257