Bringing Visual Inference to the Classroom

Adam Loy
Department of Mathematics and Statistics, Carleton College

October 15, 2019

Abstract

In the classroom, we traditionally visualize inferential concepts related to inference using static graphics or interactive apps. For example, there is a long history of using apps to visualize sampling distributions. Recent developments in statistical graphics have created an opportunity to bring additional visualizations into the classroom to hone student understanding. Specifically, the lineup protocol (Buja et al. 2009) provides a framework for students see the difference between signal and noise. This protocol involves embedding a plot of observed data in field of null plots. This approach has proved valuable in visualizing randomization/permutation tests, diagnosing models, and even conducting valid inference when distributional assumptions break down. This paper provides an overview of the lineup protocol for visual inference and how it can be used to hone understanding of key statistical topics.

Keywords: Statistics education, Statistical graphics, Simulation-based inference, Visualizing uncertainty, Lineup protocol

1 Introduction

Could I use the GAISE guidelines to frame the start of this introduction???

- Recommendation 1: Teach statistical thinking
    - simulation-based inference has helped here
- Recommendation 2: Focus on conceptual understanding
    - simulation-based inference has helped here
    - applets help students visualize resampling, but not always the big picture idea behind inference… that's where sesame-street logic comes into play…
- Recommendation 5: Use technology to explore concepts and analyze data
    - apps help
    - lot's of great options (StatKey, Rossman/Chance, Many Shiny apps…)

Can we link to hypothetical outcome plots somehow?? Essentially we get 19ish hypothetical outcomes under a model…

Recent years have seen a great deal of innovation in how we teach statistics as we strive to overcome what Cobb (2007) termed "the tyranny of the computable." Most notably, simulation-based pedagogies for the first course have been proposed and validated (Cobb 2007, Tintle et al. 2011, 2012, Maurer & Lock 2014, Nathan L. Tintle, Dordt College et al. 2014). These simulation-based pedagogies have also been used in mathematical statistics (Chihara & Hesterberg 2011, Cobb 2011), and Tintle, Chance, Cobb, Roy, Swanson & VanderStoep (2015) argue that they should be used throughout the entire curriculum.

In addition to changes in how we introduce inference, there have also been changes in

Additionally, the GAISE guidelines have led instructors to use visualization to grapple with higher-dimensional data sets and messier data…

Something about apps…

Something about sampling distributions…

The goal of this article is to discuss how to incorportate visual inference into your classroom to help your students differentiate between different forms of signal and noise, and better understand the nuances of statistical significance. Section 2 presents an overview of visual inference, specifically the lineup protocol. Section 3 presents examples of how the lineup protocol can be used in the first course, and Section 4 presents additional examples throughout the curriculum. We conclude with a brief summary and discussion in Section 5

2 Visual inference

As outlined by Cobb (2007), most introductory statistics books teach that classical hypothesis tests consist of (i) formulating null and alternative hypotheses, (ii) calculating a test statistic from the observed data, (iii) comparing the test statistic to a reference (null) distribution, and (iv) deriving a p-value on which a conclusion is based. This is still true for the first course after adapting it to address the new GAISE guidelines regardless of whether a simulation-based approach is used (cf. Lock et al. 2013, Tintle, Chance, Cobb, Rossman, Roy, Swanson & VanderStoep 2015, De Veaux et al. 2018).

The *lineup protocol* for visual inference has a direct analog for each of these four steps, as outlined by Buja et al. (2009). As a first example of visual inference via the lineup protocol, consider the creative writing experiment discussed by (Ramsey & Schafer 2013, pp. 2–14). The experiment was designed to exploer whether creativity scores were impacted by the type of motivation (intrinsic or extrinsic). To do this creative writers were randomly assigned to a questionnaire where they ranked reasons they write: one questionnaire listed intrinsic motivations and the other listed extrinsic motivations. After completing the questionnaire, all subjects wrote a Haiku about laughter which was graded for creativity by a panel of poets. Ramsey & Schafer (2013) discuss how to conduct a permutation test for the difference in mean creativity scores between the two treatment groups. Below, we illustrate the steps of a visual test.

1. A visual test begins identically to a traditional hypothesis test, by clearly stating the competing claims about the model/population parameters. In a first course, this could be written as: $H_0 : \mu_{\text{intrinsic}} - \mu_{\text{extrinsic}} = 0$ vs. $H_0 : \mu_{\text{intrinsic}} - \mu_{\text{extrinsic}} \neq 0$.

2. The test statistic is a plot displaying an aspect of the data or model that allows the observer to differentiate scenarios under the null hypothesis from scenarios under

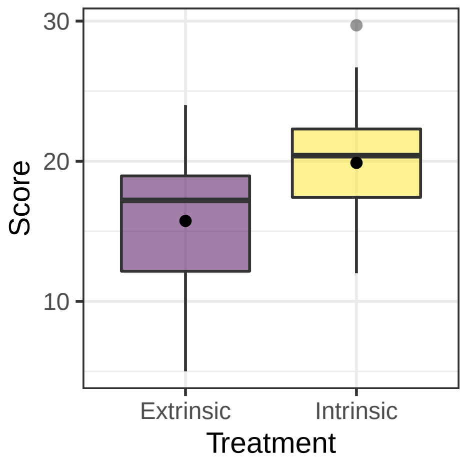

Figure 1: Boxplots of the original creative writing scores by treatment group. The dot represents the mean of each group.

alternative hypotheses. Here, side-by-side boxplots, or faceted histograms or density plots are reasonable choices to display the relevant aspects of the distribtions. Figure 1 are boxplots of the original creative writing scores by treatment group where a dot is used to represent the sample mean for each group.

3. *Null plots* are generated consistently with the null hypothesis and the set of all null plots constitutes the reference distribution. To facilitate comparison of the data plot to the null plots, the data plot should be randomly situtated in the field of null plots. This arrangement of plots is called a *lineup*. Figure 2 shows one possible lineup for the creative writing experiment. The 19 null plots were generated via permutation resampling, and the data plot was randomly assigned to panel #4.

4. If the null hypothesis is true, then we expect the data plot to be indistinguishable from the null plots. Thus, is one is able to identify the data plot in panel #4 of Figure 2, then this provides evidence against the null hypothesis. If one wishes to calculate a *visual p-value*, then lineups need to be presented to a number of independent observers for evaluation. While this is possible, we will not discuss this process as the pedagaogical value of the lineup protocol is in visualizing signal and noise.

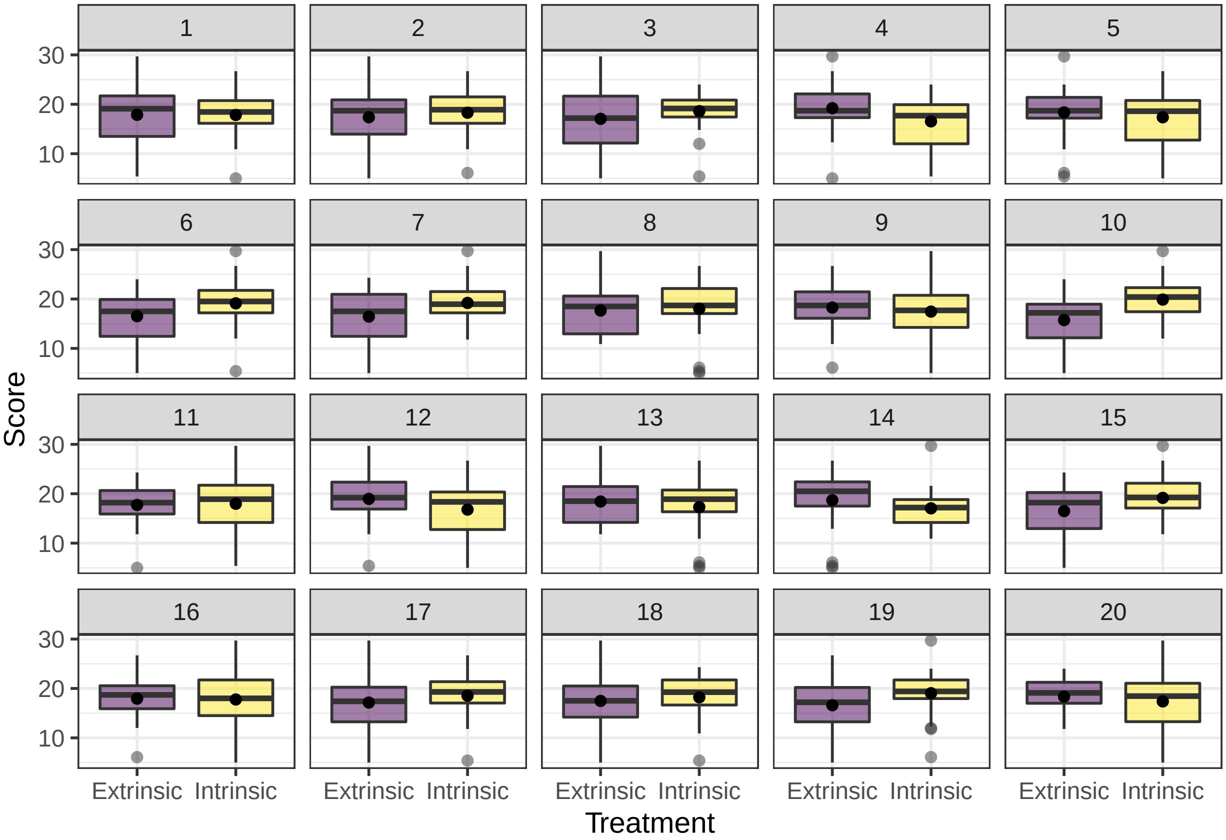

Figure 2: A lineup consisting of 19 null plots generated via permutation resampling and the original data plot for the creative writing study. The data plot was randomly placed in panel #4.

3 Using visual inference in introductory statistics

In this section, we discuss how to incorportate the lineup protocol into your classroom to clarify common points of confusion. The goal is to provide examples of how this could be done, not to provide an exhaustive list of possibilities.

3.1 Introducing (simulation-based) inference

The strong parallels between visual inference and classical hypothesis testing make it a natural way to introduce the idea of statistical significance without getting bogged down in the minutia of p-values. In this section, we will outline how we use visual inference to introduce the concepts behind hypothesis testing without devling into formal details. To do this, we will continue to discuss the reative writing experiment introduced in Section 2.

3.2 Interpreting unfamiliar plots

A second way to utilize visual inference in the first course is to help students build intuition about new and unfamiliar plot types. Two plot types that we have found introductory students struggle to interpret are the normal quantile-quantile (Q-Q) plot and the residual plot. In both situations, the lineup protocol helps students tune their understanding of what constitutes an "interesting" pattern (i.e. signal).

3.2.1 Q-Q plots

Teaching a novice to interpret a normal Q-Q plot is no easy task, especially in an algebra-based first course where students can get lost in the calculation of quantiles rather than focusing on the bigger picture. Further, Q-Q plots already suffer from a perception problem, since humans have a tendency to evaluate the shortest (i.e. orthogonal) distance, even when asked to evaluate vertical distances (Cleveland & McGill 1984, Robbins 2005, VanderPlas & Hofmann 2015). While a detrended version of the Q-Q plot has been proposed to overcome this difficulty (Loy et al. 2016), the plot still requires students to calibrate their intuition. In this paper we will discuss "classical" Q-Q plots for simplicity.

...show for smaller sample sizes... do i need to cite meeker and cook??....

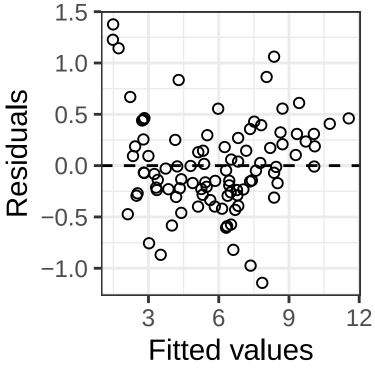

Figure 3: A residual plot for a simple linear regression model. Is there evidence that the model is insufficient?

3.2.2 Residual plots

Interpreting residual plots is also fraught with common errors. We have found that, regardless of our valient attempts to explain what "random noise" or "random deviations from a model" might look like, there is no substitute for first hand experience. In this section we outline a class activity/discussion that we use to help train students to interpret residual plots. The full activity can be found in the supplemental materials.

To begin, we have students fit a simple linear regression model, write down what a residual is (in both words and using notation), and then create a first residual plot, such as Figure 3. Next, we pose the question: "Does this residual plot provide evidence of a model deficiency?" This provides students time to formalize their decision, especially what features of the residual plot they based their decision upon.

Once students have carefully interpreted the observed residual plot, we have them generate a lineup where their data plot has been randomly embedded in a field of null plots, as shown in Figure 4. Here, the null plots have been generated using the parametric bootstrap, but the residual or non-parametric bootstraps are other viable choices. We avoid the details of how the null plots were generated, but this depends on the goals for your class. Once the lineup has been generated, we ask students to (i) identify which panel contains the observed residual plot, (ii) describe patterns they observed in three null plots, and (iii) decide whether/how the observed residual plot is systematically different from the null plots.

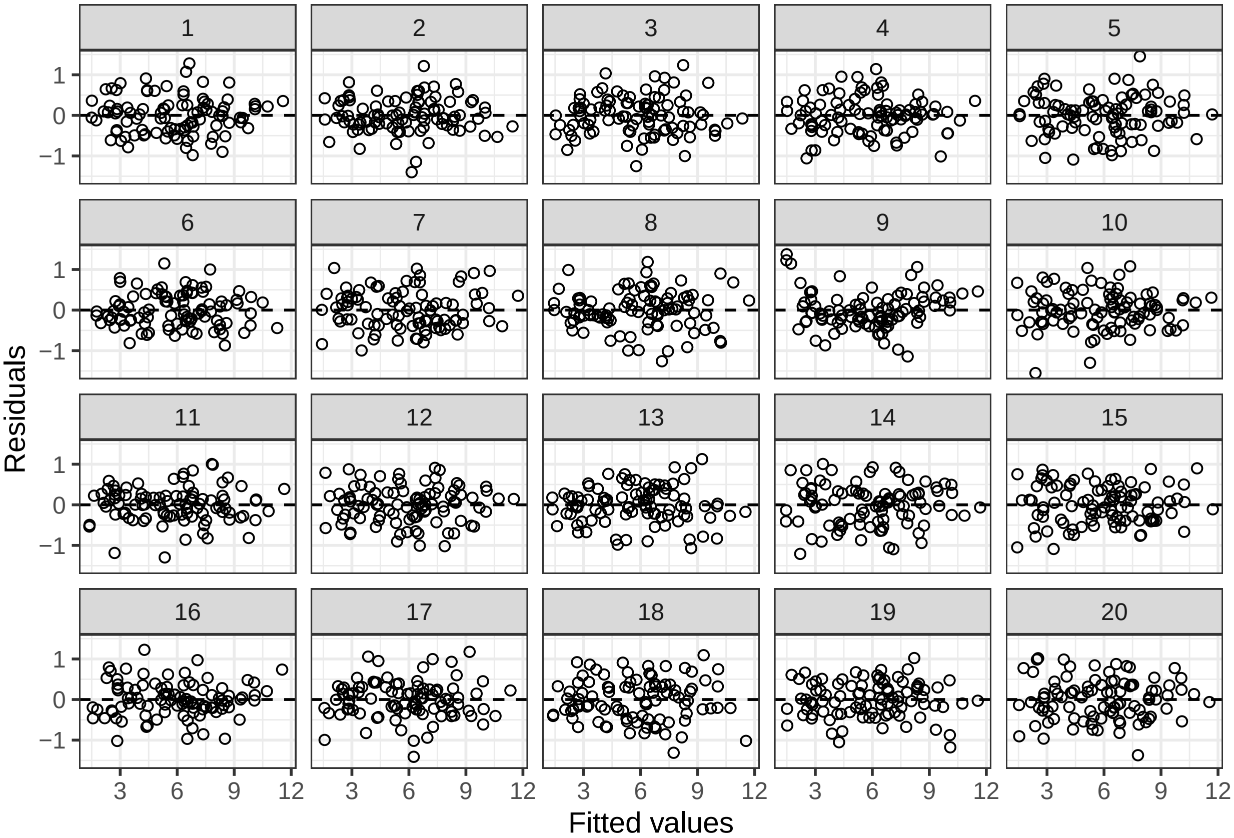

Figure 4: A lineup of residual plots. The null plots are generated via a parametric bootstrap from the fitted model. The observed data are shown in panel #9.

Teaching tips…

- In the above example, the observed residual plot in panel #9 is systematically different from the null plots. While this is one example we use in class, we also recommend a parallel example where there is no discrepancy between the data and the model.
- Depending on your prefernce and goals, follow-up discussions about the design of residual plots could be injected to the end of this activity. For example, you could provide students with a second version of the lineup where LOESS smoothers have been added to each panel and ask students what features of the residual plot this highlights.
- An alternative approach would be to have students first use the Rorschach protocol to look through a series of null plots, describing what they see, and then look at a single residual plot.

4 Using visual inference in other courses

The utility of visual inference is not restricted to introductory courses. We have found that whenever a new model is encountered intuition about diagnostic plots must be rebuilt. As an example we will consider diagnostics for binary logistic regression models, a common topic in a second course.

Interpreting residual plots from binary logistic regression is extremely difficult, as plots of the residuals against the fitted values or predictors often look similar for adequate and inadequate models. The lineup protocol provides a framework to have this dicussion. For example, you can simulate one data set that is appropriate for binary logistic regression and one that violates the linearity of the logit, and have your students try to identify the data plot in each situation.

After establishing the pitfalls of "conventional" residual plots for binary logistic regression, you can turn to alternative strategies, again using the lineup to calibrate student intution with new plot types. Below we present two examples.

Binned residuals. Gelman & Hill (2007) recommend using binned residual plots to explore possible violations of linearity for binary logistic regression. A binned residual plot

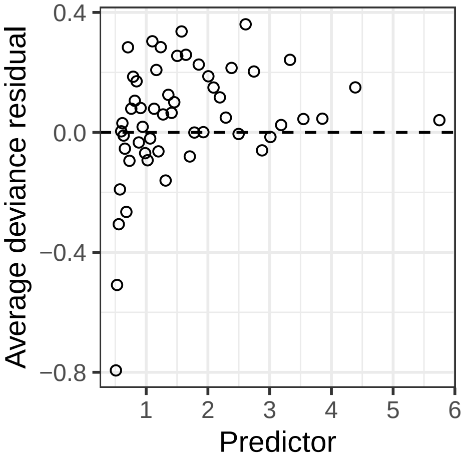

Figure 5: A binned residual plot from a simple binary logistic regression model. The average deviance residual is plotted on the y-axis for each of 54 bins on the x-axis.

is created by calculating the average residual for a number of bins along the x-axis. Figure 5 shows an example from a simple binary logistic regression model where the average deviance residual is plotted on the y-axis for each of 54 bins on the x-axis. The bins were set as with a histogram, and $\lfloor \sqrt{n} \rfloor$. Gelman & Hill (2007) claim that these plot should behave much like the familiar standardized residual plots from regression. If this is the case, then Figure 5 is indicative of nonlinearity. However, rather than simply citing Gelman & Hill (2007), a lineup empowers students to investigate the behavior of this new plot type. A lineup for these residuals is given in Figure 6. As suspected, the data plot (panel #8) clearly stand out from the field of null plots.

Empirical logit plots. A more-common alternative to the binned residual plot is the empirical logit plot…

Need to define briefly…

Something about small sample sizes and difficulty interpreting… Let's use one of the Stat2 example here… it should be well-known enough…

In this section we focused on using visual inference to help diagnose logistic regression models, but the approach is more general. If you have a plot highlighting some feature(s) of the fitted model, then after simulating data from a "correct" model (i.e. one without model deficiencies), you can create a lineup to interrogate the model. For example, Loy

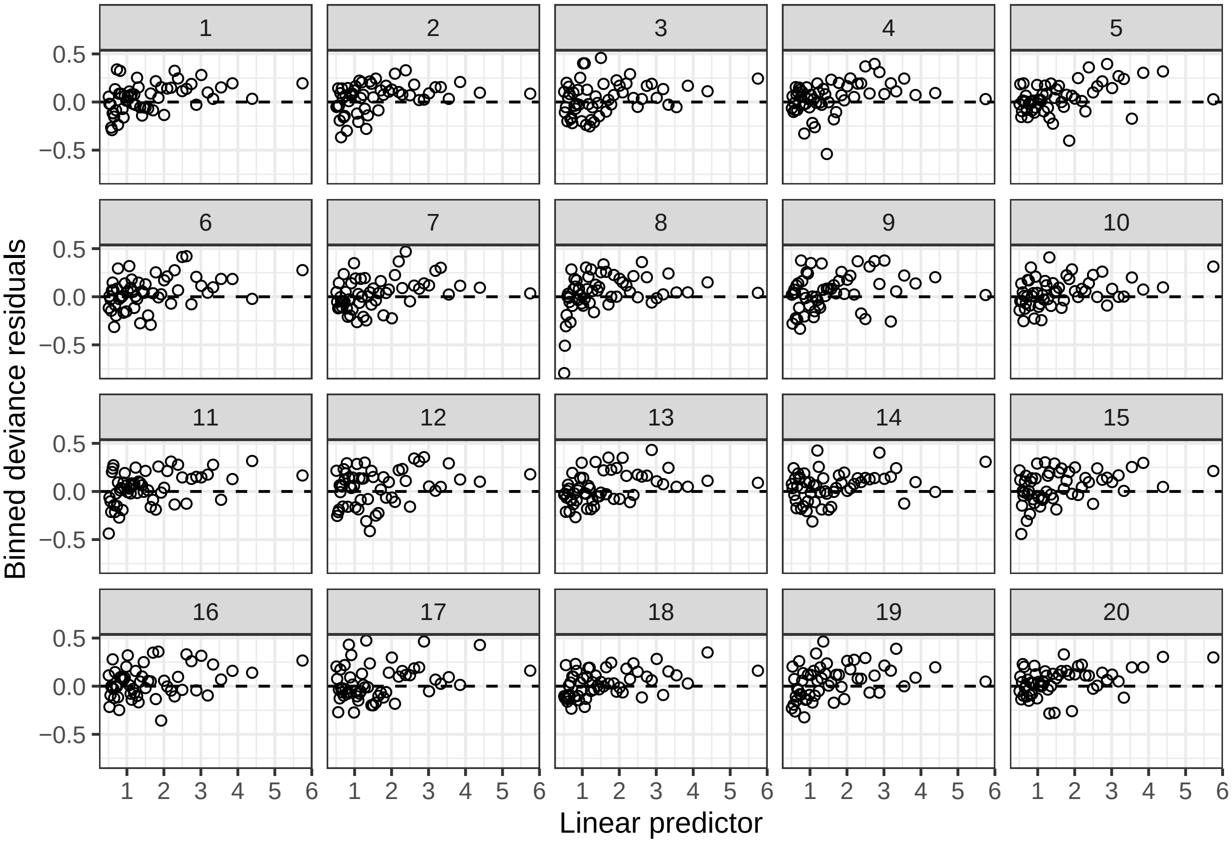

Figure 6: A lineup of binned residual plots from a simple binary logistic regression model. The observed residuals are shown in panel #8 and clearly stand out from the field of null plots, indicating a problem with linearity.

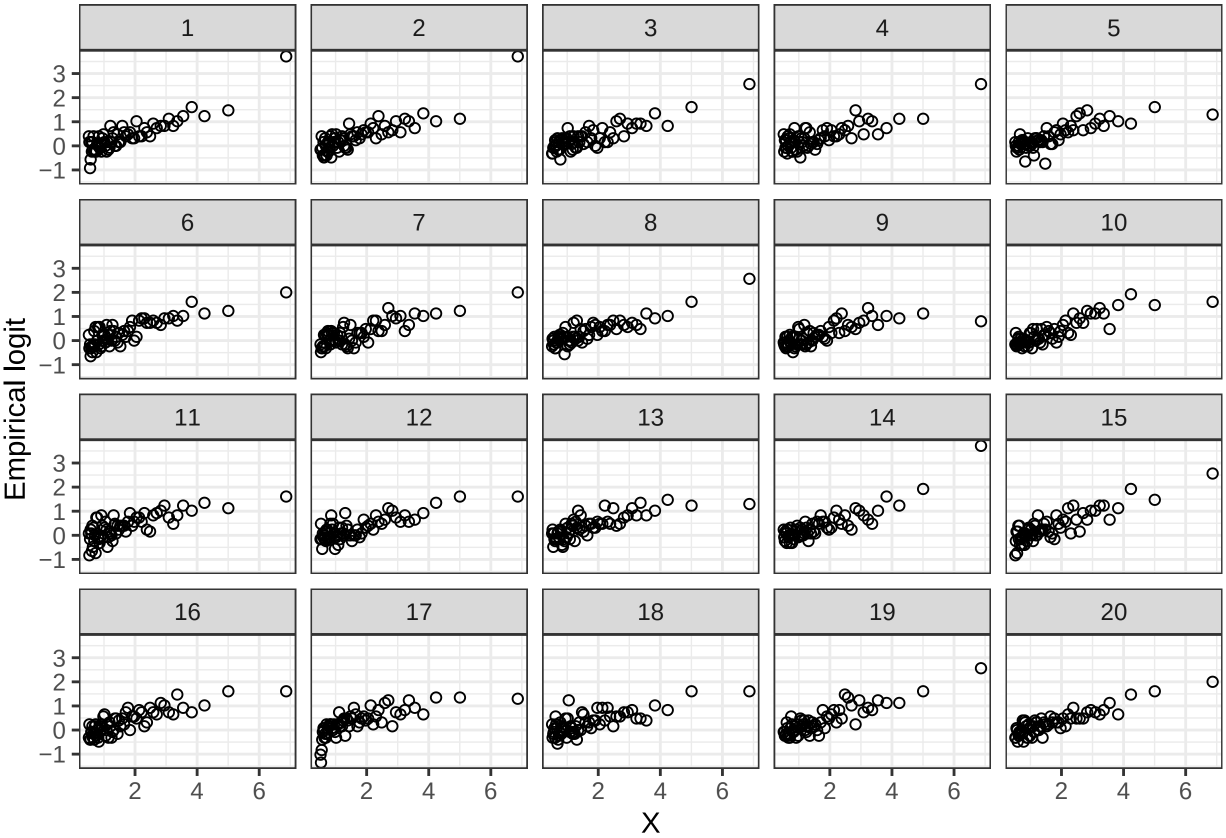

Figure 7: A lineup of empirical logit plots from a simple binary logistic regression model. The observed residuals are shown in panel #17 and clearly stand out from the field of null plots, indicating a problem with linearity.

et al. (2017) discuss how visual inference can be used to diagnose multilevel models.

5 Discussion

MENTION PEDAGOGICAL CONSIDERATIONS NOT YET DISCUSSED, HOW TO IMPLEMENT IN SOFTWARE OR AN APP, IN CLASS V. HOMEWORK, ETC.

A BRIEF GO AND DO PARAGRAPH TO REITERATE THE MESSAGE

References

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E. K., Swayne, D. F. & Wickham, H. (2009), 'Statistical inference for exploratory data analysis and model diagnostics', *Philosophical Transactions of the Royal Society A* **367**(1906), 4361–4383.

Chihara, L. & Hesterberg, T. (2011), *Mathematical Statistics with Resampling and R*, Wiley.

Cleveland, W. S. & McGill, R. (1984), 'Graphical perception: Theory, experimentation, and application to the development of graphical methods', *Journal of the American Statistical Association* **79**(387), 531–554.

Cobb, G. W. (2007), 'The introductory statistics course: a ptolemaic curriculum?', *Technology Innovations in Statistics Education* **1**.

Cobb, G. W. (2011), 'Teaching statistics: Some important tensions', *Chil. J. Stat.* **2**(1), 31–62.

De Veaux, R., Velleman, P. & Bock, D. (2018), *Intro Stats*, 5 edn, Pearson, Boston, MA.

Gelman, A. & Hill, J. (2007), *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press, New York.

Lock, R., Frazer Lock, P., Lock Morgan, K., Lock, E. & Lock, D. (2013), 'Statistics: Unlocking the power of data'.

Loy, A., Follett, L. & Hofmann, H. (2016), 'Variations of Q–Q plots: The power of our eyes!', *Am. Stat.* **70**(2), 202–214.
URL: *http://dx.doi.org/10.1080/00031305.2015.1077728*

Loy, A., Hofmann, H. & Cook, D. (2017), 'Model choice and diagnostics for linear Mixed-Effects models using statistics on street corners', *Journal of Computational and Graphical Statistics* **26**(3), 478–492.
URL: *http://dx.doi.org/10.1080/10618600.2017.1330207*

Maurer, K. & Lock, D. (2014), 'Comparison of learning outcomes for randomization-based and traditional inference curricula in a designed educational experiment', pp. 1–18.

Nathan L. Tintle, Dordt College, Ally Rogers, Dordt College, Beth Chance, C. P. S. U.-S. L. O., George Cobb, Mt. Holyoke College, Allan Rossman, C. P. S U.-S. L. O., Soma Roy, C. P. S. U.-S. L. O., Todd Swanson, Hope College, Jill VanderStoep, Hope College & Authors (2014), 'Quantitative evidence for the use of simulation and randomization in the introductory statistics course'.

Ramsey, F. & Schafer, D. (2013), *The statistical sleuth: a course in methods of data analysis*, 3 edn, Cengage Learning, Boston, MA.

Robbins, N. (2005), *Creating More Effective Graphs*, Wiley.

Tintle, N., Chance, B., Cobb, G., Roy, S., Swanson, T. & VanderStoep, J. (2015), 'Combating Anti-Statistical thinking using Simulation-Based methods throughout the undergraduate curriculum', *The American Statistician* **69**(4), 362–370.

Tintle, N., Chance, B. L., Cobb, G. W., Rossman, A. J., Roy, S., Swanson, T. & VanderStoep, J. (2015), *Introduction to statistical investigations*, John Wiley & Sons, Danvers, MA.

Tintle, N. L., Topliff, K. & VanderStoep, J. (2012), 'Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum', *Statistics Education Research Journal* **11**, 21–40.

Tintle, N., VanderStoep, J. & Holmes, V. L. (2011), 'Development and assessment of a preliminary randomization-based introductory statistics curriculum', *Journal of Statistics Education* .

VanderPlas, S. & Hofmann, H. (2015), 'Signs of the sine Illusion—Why we need to care', *Journal of Computational and Graphical Statistics* **24**(4), 1170–1190.