Bringing Visual Inference to the Classroom

Adam Loy

Department of Mathematics and Statistics, Carleton College

June 17, 2020

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

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, Tintle et al. 2014, Hildreth et al. 2018). 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 the computational toolkit we use throughout the statistic curricula. At the introductory level, numerous toolkits are commonplace depending on the objectives and audience of the course. Web apps are commonly used when students may not have access to their own computers, or to lower the techinical barriers to entry. Examples include StatKey (Lock et al. 2017), the *Introduction to Statistical Investigations* applets (Tintle, Chance, Cobb, Rossman, Roy, Swanson & VanderStoep 2015), and the shiny apps from Agresti et al. (2017). These apps allow students to explore course concepts without getting into the computational weeds. For courses exploring both the concepts and implementation in a realistic data analysis workflow R (R Core Team 2019) is a common open-source choice, and packages have been developed to lower the barriers to entry for students, notably the mosaic (Pruim et al. 2017), ggformula (Kaplan & Pruim 2019), and infer (Bray et al. 2019) R packages.

The above developments have enabled the statistics education community to address key recommendations made in the GAISE report (GAISE College Report ASA Revision Committee 2016). The simulation-based curriculum has focused attention on teaching statistical thinking and fostering conceptual understanding before delving into the mathematical details. An improved computational toolkit has enabled students to use technology to explore concepts, such as a sampling and permutation distributions, and to analyze data.

While the use of the simulation-based curriculum has helped students focus on the underlying ideas of statistical inference, little has changed about the way we help students *visualize* inference. Specifically, we still have students grapple with null/reference distribu-

tions in hypothesis testing and sampling distributions while they are trying to hone their intuition. These distributions are very abstract ideas, and while the web apps we use to demonstrate their construction can help make sense of a single "dot" on the distribution, students commonly lose the forest for the trees. Wild et al. (2017) propose the use of scaffolded animations to help students hone their intuition about sampling/randomization variation and discover the utility of the bootstrap and permutation distributions. While the visualizations are similar to what we see in common web apps that build these distributions, Wild et al. (2017) focus on making the concepts accessible visually. The animations discussed by Wild et al. (2017) to visualize randomization variation appear to be quite useful in communicating this complex idea to students, but a formal distribution of is not necessary to introduce the core ideas behind hypothesis testing. Instead, we can ask our students to try to identify the data plot among a small set of plots decoy plots generated by permutation resampling and link this simple perceptual task with fundamental ideas of statistical inference.

In this article, we discuss how to use the lineup protocol from visual inference to help students differentiate between different forms of signal and noise and to better understand the nuances of statistical significance. Section 2 presents an overview of 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 6

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 based on the new GAISE guidelines regardless of whether a simulation-based approach is used (cf. Lock et al. 2017, Tintle, Chance, Cobb, Rossman, Roy, Swanson & VanderStoep 2015, De Veaux et al. 2018).

In visual inference, the *lineup protocol* provides a direct analog for each step of a hy-

pothesis test (Buja et al. 2009).

1. **Competing claims**: Similar to a traditional hypothesis test, a visual test begins by clearly stating the competing claims about the model/population parameters.

2. **Test statistic**: A plot displaying the raw data or fitted model (call the *observed plot*) serves as the test statistic. This plot must be chosen to highlight features of the data that are relevant to the hypotheses. For example, a scatterplot is a natural choice to examine whether or not two quantitative variables are correlated.

3. **Reference (null) distribution**: *Null plots* are generated consistently with the null hypothesis and the set of all null plots constitutes the *reference* (or *null*) *distribution*. To facilitate comparison of the observed plot to the null plots, the observed plot is randomly situated in the field of null plots, just like a suspect is randomly situated amongst decoys in a police lineup. This arrangement of plots is called a *lineup*.

4. **Assessing evidence**: If the null hypothesis is true, then we expect the observed plot to be indistinguishable from the null plots. If the observer is able to identify the observed plot in a lineup, 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, it is not a productive discussion in most introductory courses that don't explore probability theory.

2.1 Example: Comparing groups

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 explore whether motivation type (intrinsic or extrinsic) impacted creativity scores. To evaluate 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.
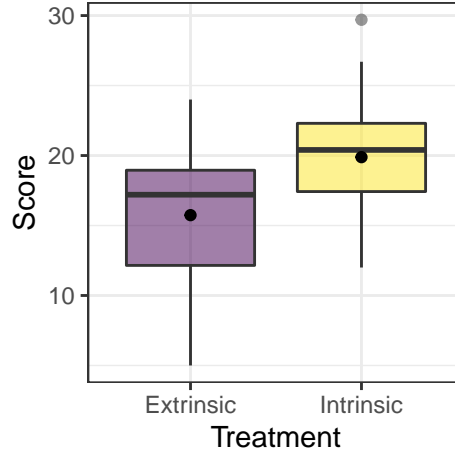
Figure 1: Boxplots of the original creative writing scores by treatment group. The dot represents the mean of each group.

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. In a visual test, plots take the role of test statistics (Buja et al. 2009). In this situation, we must choose a plot (test statistic) that can highlight the difference in average creativity scores between the intrinsic and extrinsic treatment groups. Figure 1 displays boxplots of creative writing scores by treatment group where a dot is used to represent the sample mean for each group, though other graphics could be used. There is an apparent difference in the distribution of the scores—the average score for the intrinsic group appears to be larger–but is it an important (i.e. significant) difference?

3. To understand whether the observed (data) plot provides evidence of a significant difference, we must understand the behavior of our test statistic under the null hypothesis. To do this, we generate *null plots* consistent 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 is randomly situated in the field of null plots. This arrangement of plots is called a *lineup* (named after a "police lineup" for

criminal investigations). 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, if 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 pedagogical value of the lineup protocol is in visualizing signal and noise.

3 Using visual inference in introductory statistics

In this section, we discuss how to use the lineup protocol in the introductory setting to introduce students to the logic of hypothesis testing and to new statistical graphics. 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 minutiae/controversy of p-values, or the technical issues of describing a simulation procedure before students understand why that is important. All students understand the question "which one of these plots is not like the others," and this common understanding generates fruitful discussion about the underlying inferential thought process without the need for a slew of definitions. Below is an outline of a class activity discussing the creative writing experiment to introduce the inferential thought process.

3.1.1 Outline of activity

This activity is designed to be completed in groups of three or four students. We have found that this group size allows all students to contribute to the discussion. This group
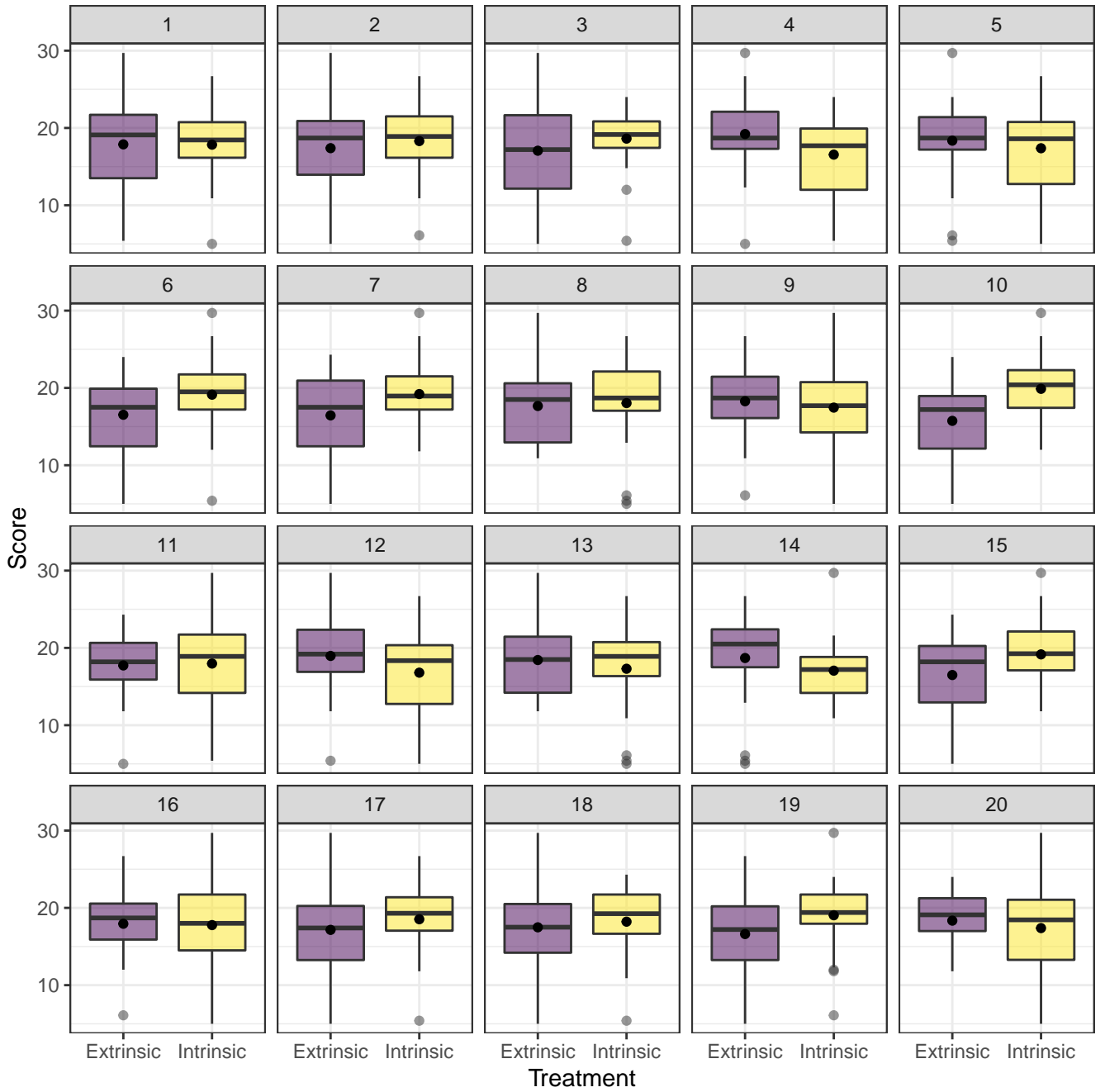
Figure 2: A lineup consisting of 19 null plots generated via permutation resampling and the original data plot for the creative writing study. The data plot was randomly placed in panel #4.

size is also conducive to assigned roles (Roseth et al. 2008), if you find that helps foster discussion in your classroom.

Competing claims. To begin, we have students discuss what competing claims are being investigated. We encourage them to write these in words before linking them with the mathematical notation they saw in the reading prior to class. The most common answer is: "there is no difference in the average creative writing scores for the two groups vs. there is a difference in the average creative writing scores for the two groups." During the debrief, I make sure to link this to the appropriate notation.

EDA review. Next, we have students discuss what plot types would be most useful to investigate this claim. It's important to ask why students selected a specific plot type, as this will make links to key ideas from exploratory data analysis.

Lineup evaluation. Most students recognize that side-by-side boxplots, faceted histograms, or density plots are reasonable choices to display the relevant aspects of the distribution of creative writing scores for each group. We then give them a lineup of side-by-side boxplots to evaluate (we do place a dot at the sample mean for each group), such as the one shown in Figure 1. At this point, we do not give the details behind the creation of null plots, we simply tell them that one plot is the observed data while the other nineteen agree with the null hypothesis. We ask students to (i) choose which plot is the most different from the others, and (ii) explain why they chose that plot. Once each student has had time to make this assessment (usually about one minute) we ask the groups to discuss and defend their choices.

Lineup discussion. Once all of the groups have evaluated their lineups and discussed their reasoning, we regroup for a class discussion. During this discussion, we reveal which panel contain the observed data (panel #4 of Figure 2), and display these data on a slide so that we can point to particular features of the plot as necessary. After revealing the observed data, we have students return to their groups to discuss whether they chose the real data, and whether their choices support either of the competing claims. Once the class regroups and thoughts are shared, we make sure that the class realizes that an identification of the observed data provides evidence against the null hypothesis (though I always hope students will be the ones saying this).
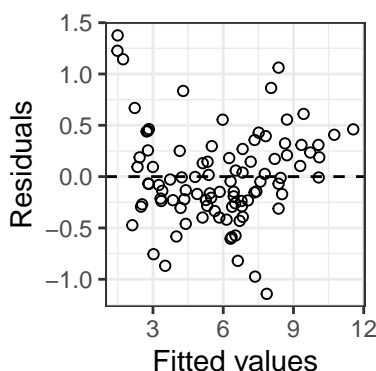
8

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

3.2 Interpreting unfamiliar plots

A second place to utilize the lineup protocol in the first course when introducing new and unfamiliar plot types. For example, we have found many introductory students struggle to interpret residual plots. In this situation, the lineup protocol helps students tune their understanding of what constitutes an "interesting" pattern (i.e. signal).

3.2.1 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 after a brief introduction to residual plots is given in class (or video if in a flipped classroom). Again, we suggest that students complete such an activity in small groups.

Model fitting. 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.

Interpreting residual plots. 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.

Lineup evaluation. Once students have carefully interpreted the observed residual

9

plot, we have them generate (or present them with) a lineup where their data plot has been randomly situated 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. We ask students to answer the first question on their own before discusing with their group, and typically give one or two minutes of individual "think time" prior to group discussion.

Debrief. Once all of the groups have evaluated their lineups and discussed their reasoning, it is important to regroup for a class discussion. This allows you to reveal the observed residual plot and revisit key points about residual plots and their interpretation.

Teaching tips

- In Figure 4, 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 course 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 the smoother highlights.

- An alternative activity first has students use the *Rorschach protocol* (Buja et al. 2009) to look through a series of null plots, describing what they see, and then looking at a single residual plot.

3.2.2   Other plot types

Similar activities can be designed to introduce other statistical graphics. Specifically, we have also found lineups help students learn to read normal quantile-quantile and mosaic
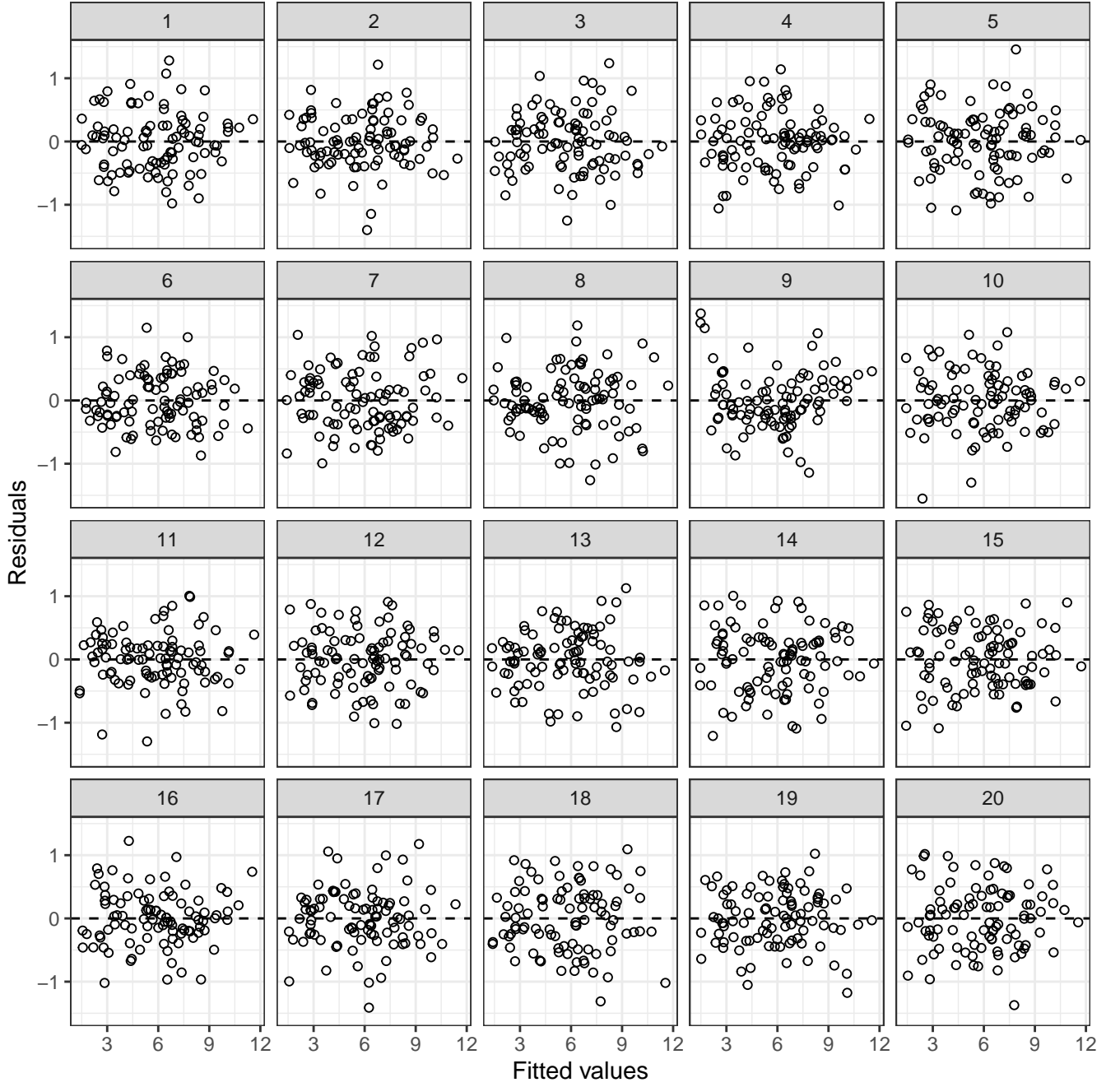
Figure 4:   A lineup of residual plots.   The null plots are generated via a parametric bootstrap from the fitted model. The observed data are shown in panel #9.

plots (or stacked bar charts).

4 Using visual inference in other courses

The utility of visual inference is not limited to introductory courses. Whenever a new model is encountered intuition about diagnostic plots must be rebuilt, so the lineup protocol is useful throughout the statistics curricula to build student intuition. As an example, consider diagnostics for binary logistic regression models, a common topic in a second course.

4.1 Diagnostics for binary logistic regression

Interpreting residual plots from binary logistic regression is 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 for this discussion. For example, you can simulate data from a model where a quadratic effect is needed, but fit the data to a model with only a linear effect and extract the Pearson residuals. Then, you can simulate the null plots from the model with only the linear effect and extract the Pearson residuals. Figure 5 shows a lineup created in this way. Having a discussion surrounding this lineup in class will help pinpoint the difficulty using conventional residual plots for model diagnoses.

After establishing the pitfalls of "conventional" residual plots for binary logistic regression, you can introduce alternative strategies (i.e., new diagnostic plots) and again use the lineup protocol to calibrate student intuition. Below are two such examples.

4.1.1 Binned residual plots

Gelman & Hill (2007) recommend using *binned residual plots* to explore possible violations of linearity for binary logistic regression. A binned residual plot is created by calculating the average residual value with bins that partition the x-axis. Figure 6 shows 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. The number of bins is set to $\lfloor \sqrt{n} \rfloor$, but can be adjusted as with a histogram. Gelman & Hill (2007) claim that these plot should behave much like the familiar standardized residual plots from regression. If this
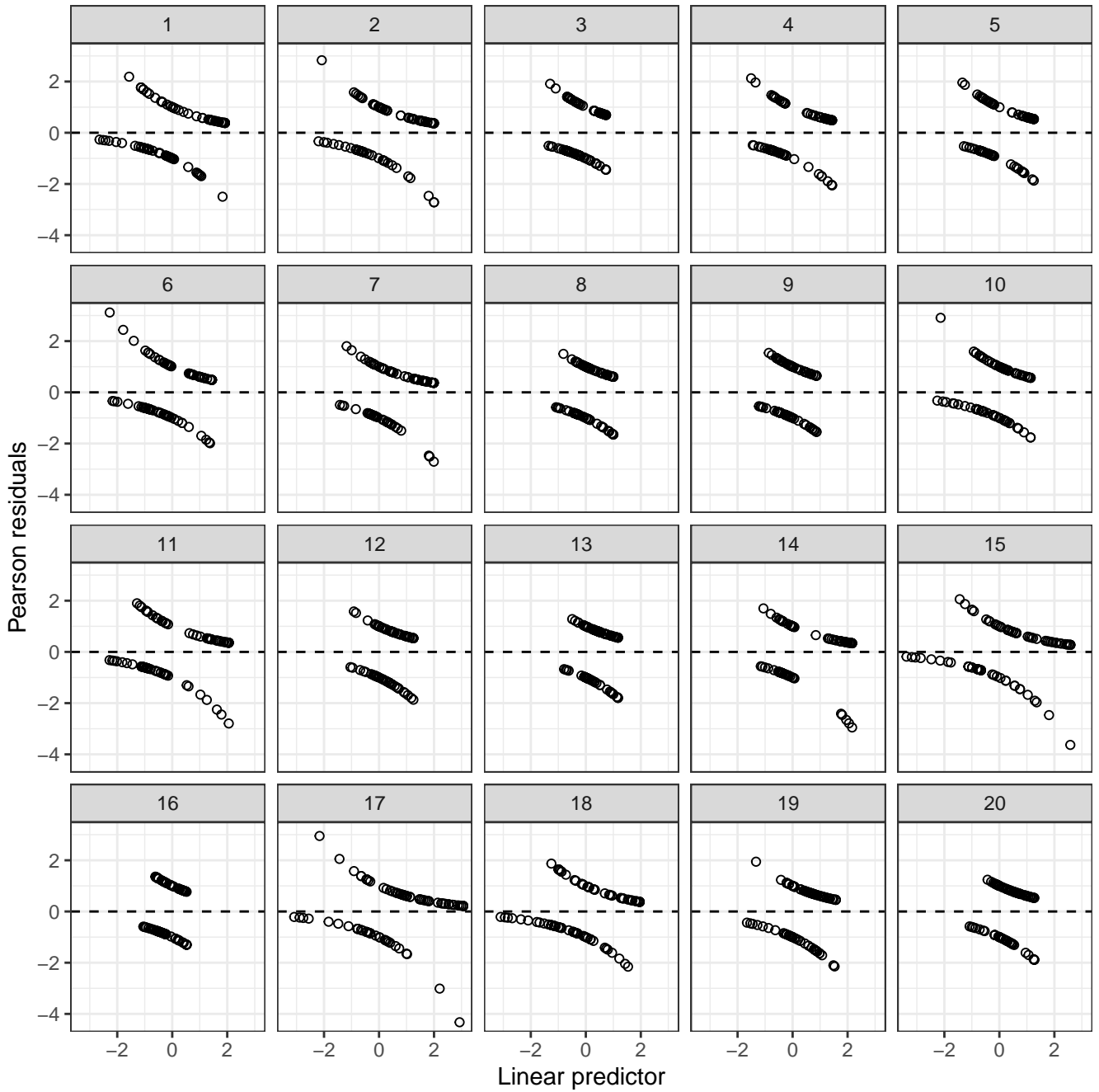
Figure 5: A lineup for a deficient logistic regression model. The data plot are simulated from a model with a quadratic effect, while the null plots are simulated from a model with only a linear effect. Can you identify the deficient plot?
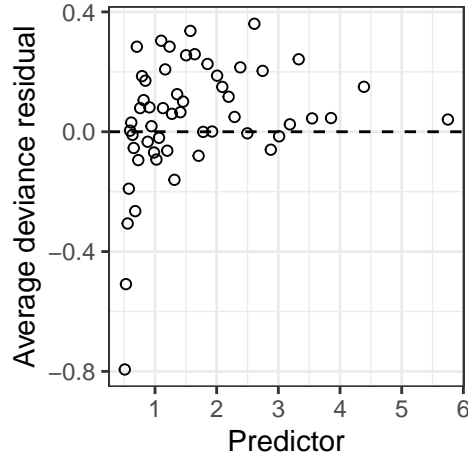
Figure 6: A binned residual plot from a simple binary logistic regression model. The average deviance residual is plotted on the y-axis for each of 54 bins on the x-axis.

is the case, then Figure 6 is indicative of nonlinearity. However, rather than simply citing Gelman & Hill (2007) to students, a lineup empowers them to investigate the behavior of this new plot type. A lineup for these residuals is given in Figure 7. As suspected, the data plot (panel #8) stands out from the field of null plots, indicating a deficiency with the model. This investigation via a lineup can be framed as a whole class discussion or as a group activity, similar to the activities already outlined.

4.1.2 Empirical logit plots

A more-common alternative to the binned residual plot is the *empirical logit plot* (c.f., Cannon et al. 2018, Ramsey & Schafer 2013). An empirical logit plot can be constructed for each explanatory variable by calculating the adjusted proportion of "successes" within each "group" as

$$\widehat{p}_{\text{adj}} = \frac{\text{number successes} + 0.5}{\text{number of cases} + 1},$$

and plotting $\log\left(\widehat{p}_{\text{adj}}/(1 - \widehat{p}_{\text{adj}})\right)$ against the average value of a quantitative explanatory variable, or the level of a categorical explanatory variable. For quantiative variables, it is common to form groups by forming bins of roughly equal size.

While an empirical logit plot is quite straightforward to create, it can be hard to in-terepret for smaller data sets where few groups are formed. For example, Cannon et al.
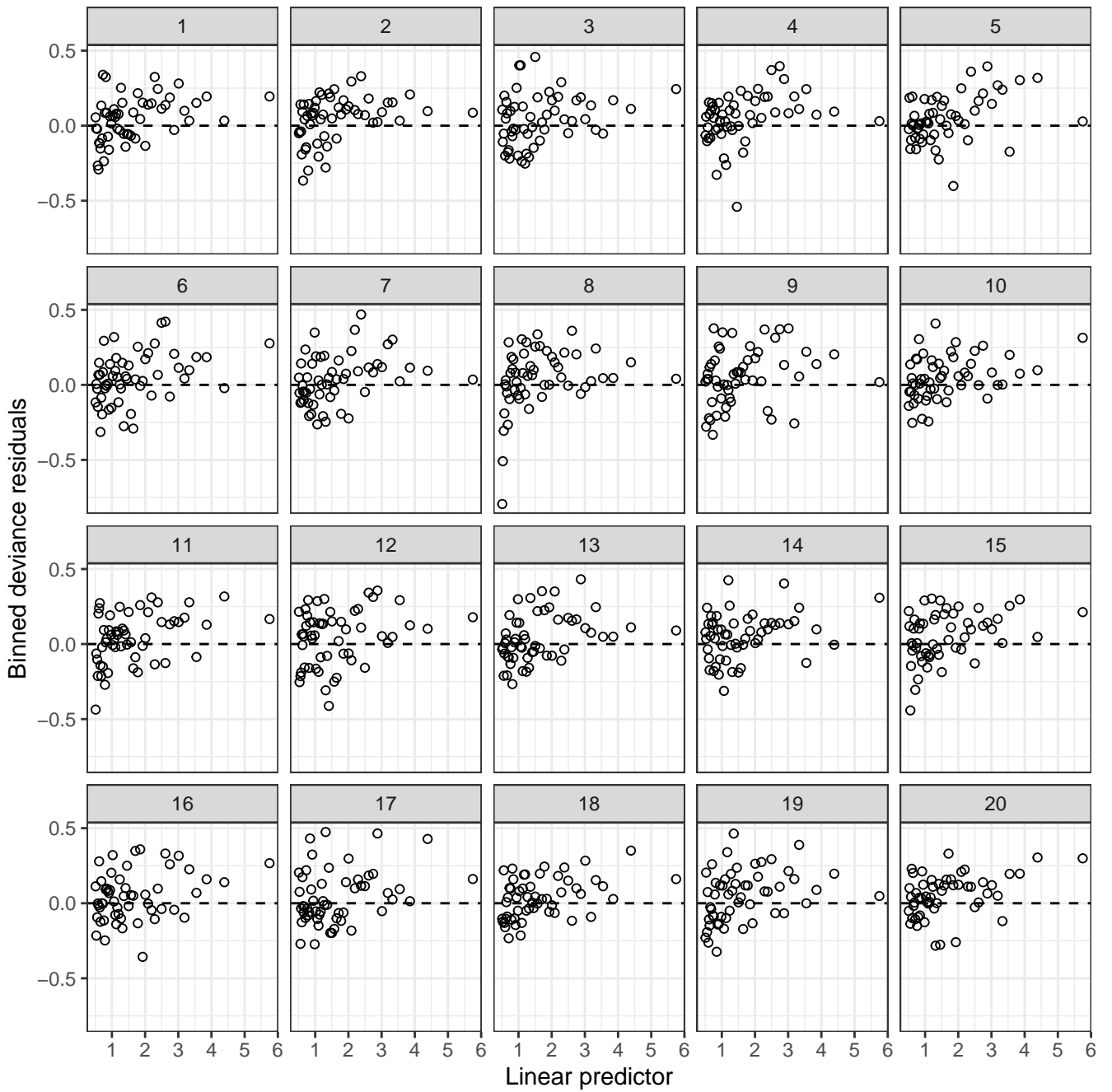
14

Figure 7: 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.
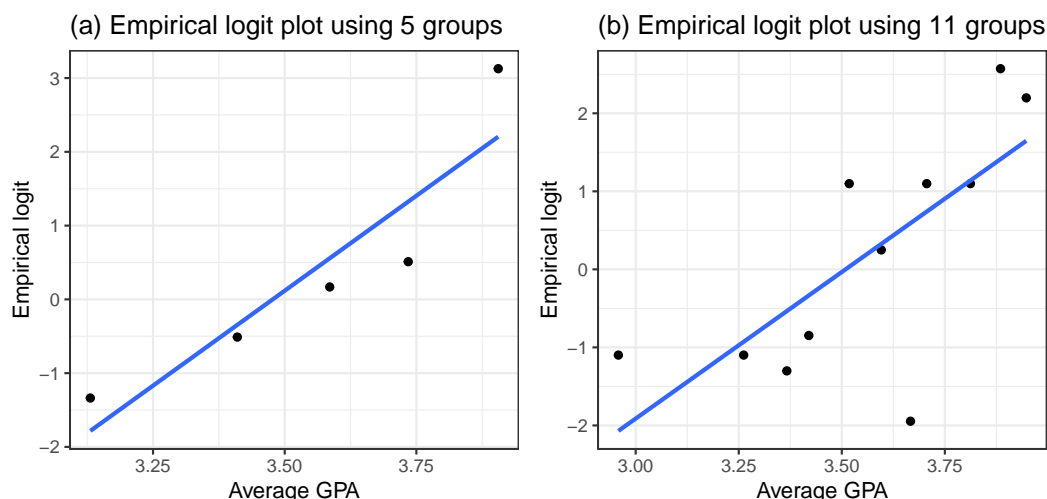
Figure 8: Two empirical logit plots rendered for the same data set with $n = 55$ observations. Panel (a) is rendered using 5 groups while panel (b) is rendered using 11 groups. The appearance of the plots changes substantially, often leading to confusion in intrepretation.

(2018) use empirical logit plots to explore a binary logistic regression model for medical school admission decisions based on an applicant's average grade point average and render empirical logit plots based on both 5 and 11 bins. Figure 8 shows recreations of these plots. Experimenting with the bin width reveals the difficulty students may have determining whether linearity is reasonable: the plot can change substantially based on the binwidth. In our experience, students often see some indication of non-linearity in the plot with 5 bins (Figure 8 (a)), whereas they think the plot with 11 bins (Figure 8 (b)) is reasonably linear.

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'

## `geom_smooth()` using formula 'y ~ x'
```

To help students interpret whether observed patterns on empirical logit plots are problematic, we again appeal to the lineup protocol to enforce a comparison between the observed plot and what is expected under the model. Figure 9 displays a lineup of the empirical logit plot created using 5 groups. The observed plot (panel #2) is difficult to pick out from the field of null plots, providing no evidence that against linearity.
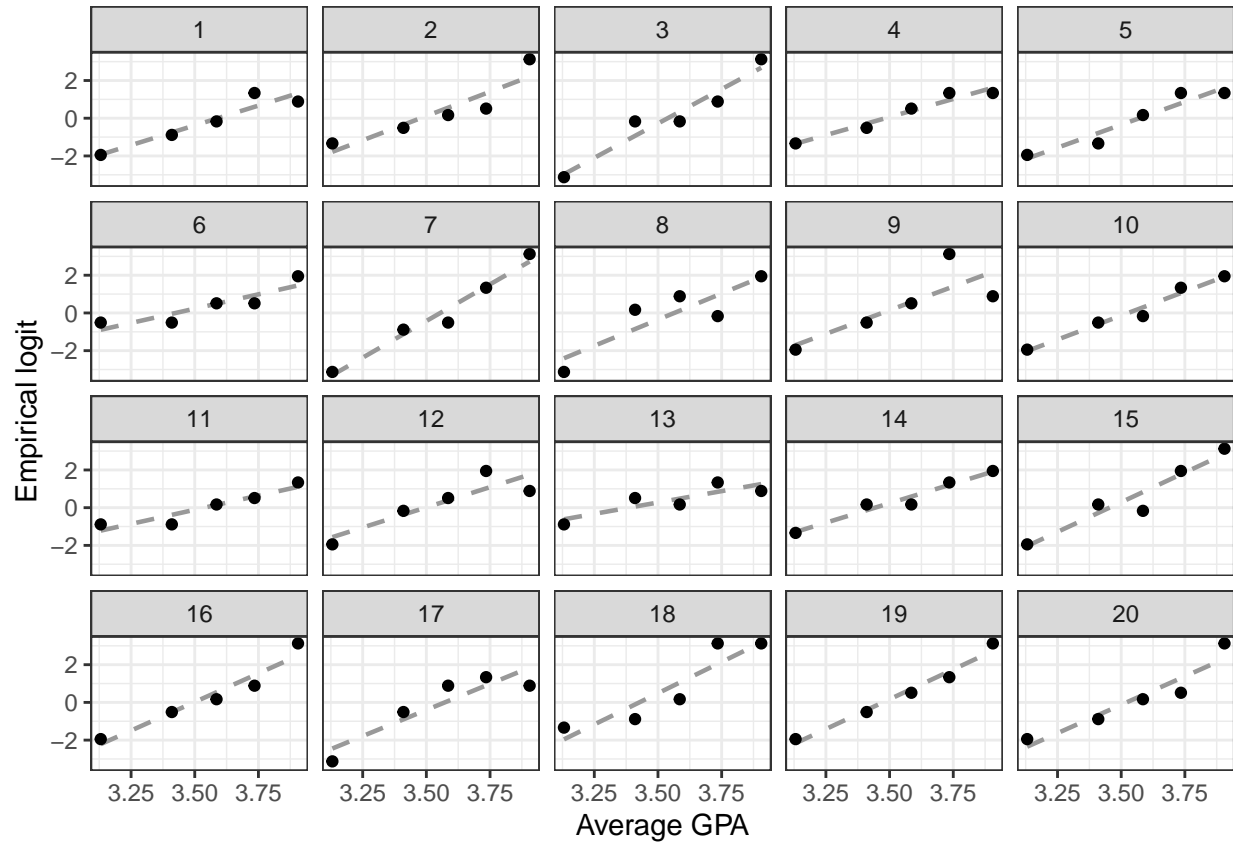
Figure 9: A lineup of empirical logit plots from a simple binary logistic regression model. The observed plot is shown in panel #2 and does not stand out from the field of null plots, indicating no problem with linearity.

4.2   Diagnosing other models

In this section, we focused on using the lineup protocol to help diagnose logistic regression models, but the approach is generally applicable. 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 et al. (2017) discuss how visual inference can be used to diagnose multilevel models.

5   Implementation

All of the lineups presented in this paper were rendered in R (R Core Team 2019). A tutorial outlining this process using the ggplot2 (Wickham 2016) and nullabor (Wickham et al. 2014) R packages is provided in the supplementary materials. These tools allow you to customize lineups for class use, but we do not recommend having introductory students grapple with this code. We recommend either providing handouts or slides with pre-rendered lineups during class activities. Alternatively, we have created a suite of Shiny apps (Chang et al. 2019) where students can upload data sets and render lineups. The current suite include apps to generate lineups to explore associations between groups, normal Q-Q plots, and residual plots for simple linear regression models. Links to the shiny apps can be found at `https://aloy.rbind.io/project/classroom-viz-inf/`.

While we have found in-class activities where students use the lineup protocol to explore new plots types and the inferential thought process to be useful, these could also be assigned as homework problems or pre-class exercises. Regardless of the venue, scaffolding the activity to guide students and foster discussion/reflection is key. In the above examples, we illustrated the approach that has worked well for our students, but each instructor should adjust this to their own class, teaching style, and student population. In addition, our discussion has not been exhaustive, so we encourage instructors to identify additional places in their curriculum where the lineup protocol can help their students build intuition.

6 Conclusion

The lineup protocol provides a framework to help students learn to interpret new statistical graphics and hone their intuition about what constitutes an interesting feature/pattern. This is achieved by randomly embedding the observed data plot into a field of decoy (null) plots, just as a suspect is randomly situated in a police lineup. Lineups provide a natural way to introduce new statistical graphics throughout the statistics curriculum. At the introductory level, lineups can help students learn to detect association in side-by-side boxplots and mosaic plots, and detect problematic patterns in residual plots for regression models. In more advanced courses, lineups can be used to frame conversations about why conventional residual plots are problematic for certain models and can improve a student's diagnostic ability as they investigate new models.

The shift to permutation tests in introductory courses has lowered the initial technical barriers to hypothesis testing; however, it still requires an explanation of *why* we need to resample and *how* we resample. Exploring lineups that you provide and making the analogy to the police lineup (or alternatively the Sesame Street question: "which one of these is not like the others") introduces students to the the logic behind testing without the need for these technical discussions. This allows initial focus to be on the core concepts of hypothesis testing rather than simultaneous focus on the core concepts and the technical details. We have found that a wide range of students understand why an inferential process is needed and what the findings imply at a more intuitive level after grappling with questions such as "which one of these plots is not like the others?", "how do you know?", and "what does this mean about your initial claim?" In addition, permutation tests logically follow the lineup protocol, providing students with the details behind the generation of the null/decoy plots and ways to formalize the strength of evidence against an initial claim.

Finally, the lineup protocol equips students with a rigorous tool for visual investigation that is applicable outside of the classroom. This not only prepares students to explore unfamiliar models or graphics in their own statistical analyses, but can also enable "teaser" conversations about advanced models for majors. For example, if you introduce your students to the lineup protocol in a modeling course, then you can show a lineup of cholorpleth maps and discuss spatial statistics as a potential area of future study.

7 Supplementary materials

8 Acknowledgements

References

Agresti, A., Franklin, C. A. & Klingenberg, B. (2017), *Statistics: The Art and Science of Learning from Data*, 4th edn, Prentice Hall.

Bray, A., Ismay, C., Chasnovski, E., Baumer, B. & Cetinkaya-Rundel, M. (2019), *infer: Tidy Statistical Inference.* R package version 0.5.1.
URL: *https://CRAN.R-project.org/package=infer*

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.

Cannon, A., Cobb;, G. W., Hartlaub, B. A., Legler, J. M., Lock, R. H., Moore, T. L., Rossman, A. J. & Witmer, J. A. (2018), *STAT2: Modeling with Regression and ANOVA*, 2nd edn, MacMillan.

Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. (2019), *shiny: Web Application Framework for R.* R package version 1.4.0.
URL: *https://CRAN.R-project.org/package=shiny*

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

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.

GAISE College Report ASA Revision Committee (2016), 'Guidelines for assessment and instruction in statistics education college report 2016'.
URL: *http://www.amstat.org/education/gaise*

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

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.

Kaplan, D. & Pruim, R. (2019), *ggformula: Formula Interface to the Grammar of Graphics*. R package version 0.9.1.
URL: *https://CRAN.R-project.org/package=ggformula*

Lock, R., Frazer Lock, P., Lock Morgan, K., Lock, E. & Lock, D. (2017), *Statistics: Unlocking the Power of Data*, 2nd edn, John Wiley & Sons, Hoboken.

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.

Pruim, R., Kaplan, D. T. & Horton, N. J. (2017), 'The mosaic package: Helping students to 'think with data' using R', *R Journal* **9**(1), 77–102.
URL: *https://journal.r-project.org/archive/2017/RJ-2017-024/RJ-2017-024.pdf*

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Ramsey, F. & Schafer, D. (2013), *The Statistical Sleuth: A course in Methods of Data Analysis*, 3 edn, Cengage Learning, Boston, MA.

Roseth, C. J., Garfield, J. B. & Ben-Zvi, D. (2008), 'Collaboration in learning and teaching statistics', *Journal of statistics education: an international journal on the teaching and learning of statistics* **16**(1).
URL: *https://doi.org/10.1080/10691898.2008.11889557*

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., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T. & VanderStoep, J. (2014), Quantitative evidence for the use of simulation and randomization in the introductory statistics course, *in* K. Makar, B. de Sousa & R. Gould, eds, 'Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.'.

Tintle, N., VanderStoep, J. & Holmes, V. L. (2011), 'Development and assessment of a preliminary randomization-based introductory statistics curriculum', *Journal of Statistics Education* **19**.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.

  URL: *https://ggplot2.tidyverse.org*

Wickham, H., Chowdhury, N. R. & Cook, D. (2014), *nullabor: Tools for Graphical Inference.* R package version 0.3.1.

  URL: *http://CRAN.R-project.org/package=nullabor*

Wild, C. J., Pfannkuch, M., Regan, M. & Parsonage, R. (2017), 'Accessible conceptions of statistical inference: Pulling ourselves up by the bootstraps', *International Statistical Review* **85**(1), 84–107.

  URL: *https://onlinelibrary.wiley.com/doi/full/10.1111/insr.12117*