The Lineup Protocol: Using Simulation to Improve Model Diagnostics in Binary Logistic Regression

Jack L. Moran

Visual and numeric diagnostics for logistic regression frequently work with group-level data instead of individual observations. This leads to diagnostics that are well defined for binomial logistic regression where data are grouped but that can break down in binary logistic regression where there are no groups. This article explores how simulation-based diagnostics can be applied to binary logistic regression to remedy this problem. In particular, we conduct a study to estimate the power of the lineup protocol, a particular simulation-based diagnostic. We discover that the lineup protocol is more powerful than the Goodness-of-fit test when data violate the independence assumption of logistic regression but less powerful when data violate the log-odds linearity assumption.

1 Introduction

Model checking is a vital step in any statistical analysis. When fitting a model, the analyst must ensure that 1) the data follow all assumptions of a given model and that 2) the model strongly captures characteristics of the data. In most cases, this is done using a variety of graphical and numerical diagnostic tests that can be classified into two categories: exploratory data analysis and model diagnostics. Exploratory data analysis occurs before a model has been fit to the data and typically checks to ensure that assumptions of a proposed model are fulfilled by the available data, the first goal of the analyst. Model diagnostics address the second goal and ensure that the proposed model properly describes the characteristics of the data. [1]

Both exploratory data analysis and model diagnostics have pitfalls that can be harmful to a statistical analysis. Exploratory data analysis is often performed using graphical representations of the available data. However, we as humans are exceptional at picking patterns out of randomness which frequently leads to the over-interpretation of visual plots. Model diagnostics often have the opposite problem. If we only consider numeric test statistics and p-values, it is easy to overlook obvious model violations. This can lead to incorrect inferences made from a model that is poorly fit to the data. [1]

This article focuses on both exploratory data analysis and model diagnostics for logistic regression with a binary response variable. The binary logistic regression model is not considered a particularly complicated model and only assumes independence of observations and log-odds linearity of its data. Additionally, visualizing binary data can often be frustrating for an analyst as all values take either a 0 or 1, causing many points to appear stacked on top of one another. This creates a temptation to cut corners when working with logistic models and can lead to incorrectly specified models. For this purpose of this article, this frustration motivates the research for new diagnostic techniques for logistic regression that are easy to use and powerful diagnostic tools.

As we continue our discussion of simulation-based diagnostics in binary logistic regression, we review the model formulation and classical diagnostic techniques used in model checking in Sections 1 and 2. Then in Section 3, we introduce simulation-based model checking techniques proposed by Gelman (2004) [4]. This article focuses on one such diagnostic, the lineup protocol, first described by Buja et al. (2009) [1]. We run a simulation study, outlined in Section 4, to compare the power of the lineup protocol at detecting logistic model violations to that of the Goodness-of-fit test, a classical diagnostic test. The results of the study are presented in Sections 5 and 6, and Section 7 discusses possible direction for future research and gives recommendations for model checking in binary logistic regression.

1.1 The Lineup Protocol

The lineup protocol combines the visual nature of exploratory data analysis with the more formal hypothesis testing seen when diagnosing the fit of a model. In a visual exploration of data, the analyst looks at a plot containing the data set y and considers if there are any unusual attributes. While rarely formalized, we can think of the analyst comparing y to many many unspoken hypotheses about what "good" or "normal" data should look like. In the case of linear regression, if the analyst discovers that the data are skewed, they are rejecting the null hypothesis of symmetry or normality. If they note that there is a linear relation between an explanatory variable and the response, they are rejecting the null hypothesis of independence.

We can frame these same statements about normality and independence in terms of quantitative testing. Let $T^{(i)}(y)$ be a test statistic about the data y, where i is a feature of the data such as normality, outliers, sparseness, etc.. We can then simulate replicate data sets y^* that follow the null hypothesis for a given test statistic and compare our observed test statistic, $T^{(i)}(y)$, to the distribution of simulated test statistics $T^{(i)}(y^*)$. This method of comparing the observed data to data simulated from the fitted model using test statistics is a common form of quantitative testing and is discussed in both by Buja et al. (2009) [1] and Gelman (2004) [2]. When an analyst makes discoveries looking at a visual representation of the data y, they are noting for which test statistics $T^{(i)}(y)$ the null hypothesis is rejected in favor of some alternative. If I is the set of all features of the data such that any $i \in I$, then the analyst is considering all possible test statistics $T^{(i)}(y)$ with $i \in I$ at once when they make discoveries about a plot. [1]

Because it considers multiple test statistics at once, visual inference is more vague in its specification of null and alternative hypotheses than quantitative testing. This generality can be useful, as it often detects obvious model violations that might have been overlooked when considering individual test statistics in quantitative testing. However, as mentioned in the introduction, it can lead to over-interpretation of data and suggest violations that do not actual exist [1]. Over-interpretation occurs most frequently when we compare a visual diagnostic to an implicit rather than explicit reference distribution. For example, consider classical linear regression where we look at a residual plot. We assume that residuals are independently distributed around zero in our minds, but don't actually have a plot of independent residuals distributed around zero in front of us to compare the observed residuals to. The lineup protocol provides the analyst with these reference distributions explicitly. [1]

Assume that for data y, we have fit a model and have created a graphical visualization of the data. We will refer to this visual as the true plot. The lineup protocol simulates n null data sets according to the fitted model and, in the same manner as the true plot was made, creates n null

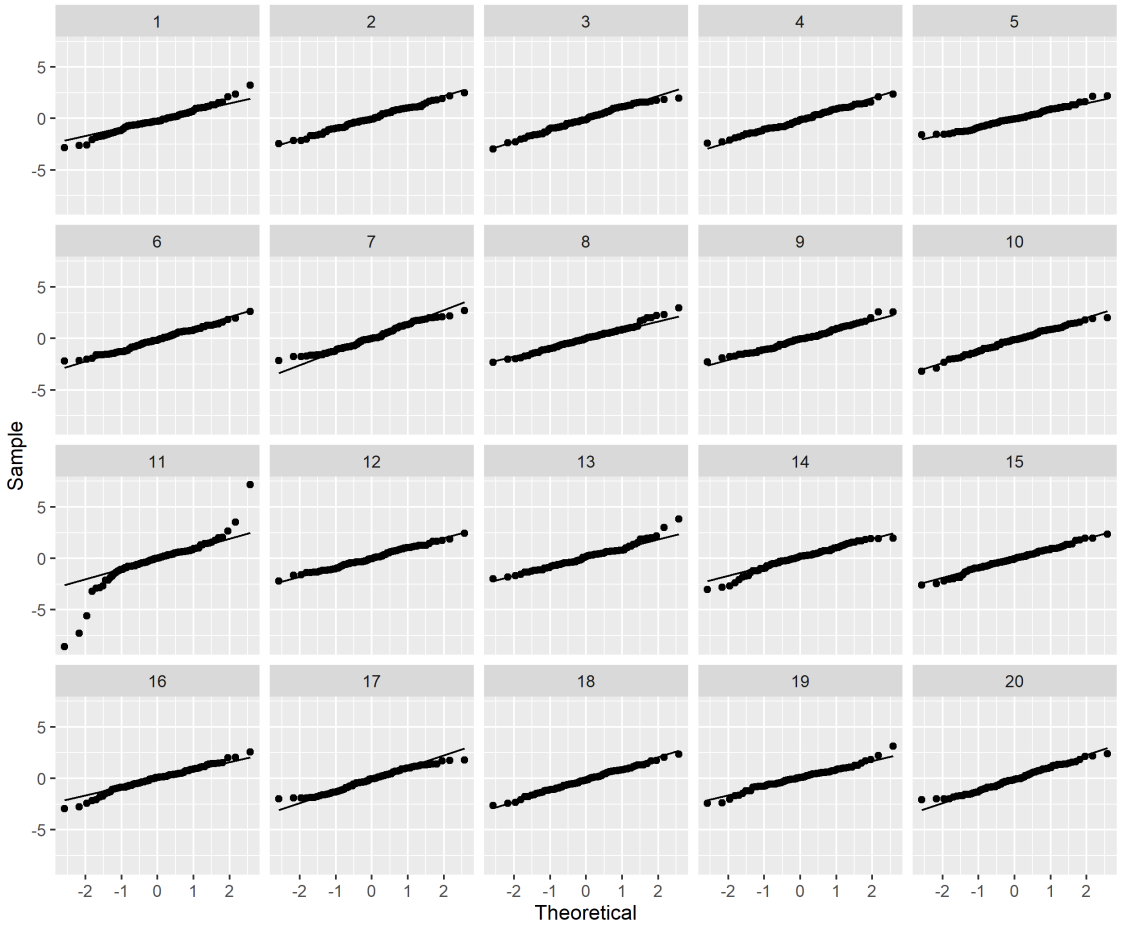


Figure 1: An example lineup of QQ-plots. The true plot is located in panel $(2^3 + 3)$.

plots. The specifics for simulating null data sets according to the model are discussed in Section 3. We use the terms null data and null plot because simulated data follow all assumptions of the fitted model. Thus the null hypothesis is true for the simulated data with any conceivable test statistic. The true plot and the n null plots are displayed next to one another, and a viewer is tasked with trying to pick the true plot out from the nulls. If the true plot stands out from the null plots, then it means that the true data is not consistant with the null hypothesis. Typically, we choose $n = 19$ so that there are 20 plots in all. We can think of the probability that the viewer picks the true plot purely at random as $p = 1/20 = 0.05$. Thus, if the viewer picks the true plot, we can assign a p-value of 0.05 to the discovery. Giving the lineup to K independent viewers can result in even smaller p-values. If we have k out of K independent observers choose the true plot, then the p-value is given by the probability $P(X \leq k)$ with $X \sim \text{Binom}(K, p = 1/20)$. [1]

An example of the lineup protocol is shown in Figure 1, where the true plot is located in position $(2^3 + 3)$. Each panel contains a QQ-plot that graphs the observed sample quantiles against the theoretical quantiles from a standard normal distribution. If the sample comes from the standard normal distribution, it should have a slope of approximately 1 when plotted against the theoretical

normal quantiles. For this lineup, the null hypothesis is that the data are normally distributed, and thus all of the null data sets are simulated from the standard normal distribution. If the viewer correctly chooses plot $(2^3 + 3)$ from the lineup, then we can reject the null hypothesis and say with 95% confidence that the sample is not from a standard normal distribution. This statement is true, as the sample comes from a t-distribution with $df = 3$. The lineup protocol paired with QQ-plots has been shown to be more powerful then leading quantitative tests at diagnosing non-normality with t-distributions [6] which raises the question of where else it might be applicable.

Because the lineup protocol is a form of visual inference, its hypotheses are often more vague than quantitative tests. Hence, it can be useful for the viewer to describe their reasons for choosing a panel. In terms of hypothesis testing, these reasons are equivalent to asking which test statistics resulted in the rejection of the model. In Figure 1, the viewer might picked plot $(2^3 + 3)$ because the left and right sides of the plot were different from the others. This corresponds to normality being violated in both tails of the distribution and would point to the t-distribution rather than other non-standard-normal distributions [1]. This article focuses on the lineup protocol applied to binary logistic regression. Therefore, it is essential to have a firm understanding of the logistic model as well as common visualizations and diagnostics for binary data.

1.2 Logistic Model Formulation

Binary logistic regression is the most common way to model binary data where observations are independent Bernoulli trials. Because the response only takes values of 0 and 1, we use a generalized linear model with link function $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$. The inverse-logit function transforms the continuous combination of linear predictors to the interval (0,1). Therefore, binary logistic regression models the probability that a trial is a success given its linear combination of explanatory variables.

$$P(y_i = 1) = \text{logit}^{-1}(\beta_1 X_{1i} + ... + \beta_p X_{pi}) \tag{1}$$

Equivalently, the model can be expressed in two parts with $\text{logit}(x) = \log(x/(1-x))$.

$$P(y_i = 1) = p_i \tag{2}$$

$$\text{logit}(p_i) = \beta_1 X_{1i} + ... + \beta_p X_{pi} \tag{3}$$

Because the inverse-logit function is curved, the response $P(y_i = 1)$ is not linearly related to its predictors. A change of n units in an explanatory variable x_i could be associated with varying amounts of change in $P(y_i = 1)$ depending on the initial value of x_i. [4]

Because the probability of success is not linearly related to the predictors, it is often easier to talk about the log-odds of success in logistic regression. Note that the odds of success for an observation $p_i/(1-p_i)$ are in the same form as $\text{logit}(p)$ as defined above. Thus, exponentiated logistic regression coefficients can be interpreted as the log-odds ratio of success for an observation.

$$\log\left(\frac{P(y_i = 1|x_i)}{P(y_i = 0|x_i)}\right) = \beta_1 X_{1i} + ... + \beta_p X_{pi} \tag{4}$$

This linear relationship between predictors and log-odds allows for easier interpretations of logistic regression coefficients and is a fundamental assumption of the logistic model. [4]

1.3 Model Assumptions

The two assumptions of binary logistic regression are log-odds linearity for explanatory variables and independence of observations. As shown above, the use of the inverse-logit link function implies that all predictors are linearly related to the log-odds of the response. This assumption can be violated if explanatory variables are missing necessary transformations such as when significant quadratic terms are omitted from a model.

The second assumption of binary logistic regression is independence of observations which can often be more difficult to detect than non-linearity. Independence of observations is violated any time that knowing the outcome of one observation gives information about the outcome of another. Independence is most commonly violated when observations are nested in groups. For example, if we were modeling the probability that an individual voted for a certain political candidate, it might be the case people who live in the same house have similar voting preferences. Thus, knowing how one family member voted would give information as to the vote of another family member, violating independence.

Often times when data are nested in groups, the best option is to use binomial logistic regression, in which we take a group to be an observation and use only group level predictors to model the proportion of success for a given group. Because data are only correlated within groups, the groups themselves are independent and satisfy the assumption. Binomial logistic regression has the added benefit that the response variable is now a proportion instead of the binary 0 or 1. This makes diagnostic plots more similar to classical regression and easier to interpret. However, binomial logistic regression uses group-level averages for explanatory variables, and thus does not utilize all information contained in the data.

An alternative to binomial logistic regression is to fit a hierarchical logistic regression model to the binary data that includes random-effects for groups. This approach allows us to incorporate both observation level and group level predictors and allows explanatory variables to be correlated differently with different groups. Let each observation i belong to some group j. A hierarchical model that includes a random-intercept for each group is given by

$$P(y_{ij} = 1) = \text{logit}^{-1}(\alpha_{ij} + \beta_1 X_{1ij} + ... + \beta_p X_{pij}) \tag{5}$$

$$\alpha_{ij} = \alpha_0 + u_j \tag{6}$$

with $u_i \sim N(0, \sigma^2)$. Here, α_0 is the fixed-effect intercept term, which is combined with a group random-intercept, u_j, that comes from a normal distribution with variance σ^2. This model accurately accounts for correlation within groups, allowing inferences to be made about the relationship between explanatory variables and the response [4]. In the following sections, we examine the problems that exist with classical diagnostics in binary logistic regression and how we can begin to solve these problems using simulation-based diagnostics.

2 Classical Diagnostics in Logistic Regression

Many diagnostics for binary logistic regression work by binning data into groups in order to approximate the group structure of binomial logistic regression. Within groups, the proportion of success and average value for exploratory variables are calculated and used in diagnostics designed for binomial logistic regression. Binning data is generally an effective diagnostic technique, but can sometimes create issues that need to be carefully considered.

2.1 Calculating Residuals

In logistic regression, the two most commonly used types of residuals are Pearson residuals and deviance residuals. Pearson residuals are calculated by subtracting the predicted probability of success from the observed outcome and standardizing by the standard error of each observation. The variance for an observation with a binary response is given by $\hat{p}_i(1-\hat{p}_i)$ where $\hat{p}_i$ is the estimated probability of success [4] [10]. Hence, the Pearson residuals are given by

$$\text{Pearson residual}_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} \tag{7}$$

The second type of residual is the deviance residual, which measures the individual contribution of an observation to the overall model deviance. The deviance of a model is given by negative two times the log-likelihood function up to an additive constant. Written explicitly,

$$\text{Deviance residual}_i = \text{sign}(y_i - \hat{p}_i) * \sqrt{2\Big\{y_i \log\Big(\frac{y_i}{\hat{p}_i}\Big) + (1-y_i)\log\Big(\frac{1-y_i}{1-\hat{p}_i}\Big)\Big\}} \tag{8}$$

Deviance is used in logistic regression rather than R^2 because the logistic model parameters are estimated using maximum likelihood instead of least squares. Therefore, least squares error is no longer the optimal measure of model error. This article uses Pearson residuals because they are easier to calculate and interpret than deviance residuals. Additionally, in Section 2.4 we discuss some difficulties in using deviance residuals for simulation when data are binary. The later half of the paper deals exclusively with simulation-based diagnostics, and so Pearson residuals are preferred. [10]

2.2 Residual Plots

Because responses are binary, residuals in binary logistic regression can only take two values given their explanatory variables. Say that the predicted probability of success for an observation is 0.6 based on the model. The observation's outcome can only be 1 or 0, and so the Pearson residual will be either $1 - 0.6/(0.6 * 0.4) = 1.667$ or $0 - 0.6/(0.6 * 0.4) = -2.5$. An example residual plot showing Pearson residuals against predicted probability for a correctly specified logistic model is shown on the left of Figure 2. The structure induced by the binary data makes the residual plot difficult to interpret. [4]

To work around this problem, we use a binned residual plot. A binned residual plot approximates the group structure seen in binomial logistic regression by dividing observations into groups (bins) based on the value of an explanatory variable. The average residual for observations in a bin is plotted against the average value of the predictor used to make the bin. This is illustrated on the right side of Figure 2 which shows a binned residual plot with bins created from the explanatory variable x_1. Additionally, the light grey lines indicate a 95% theoretical error bound for average binned-residuals. This error bound is approximated as two times the standard error of each bin, $2\sqrt{[p_j(1-p_j)]/[m_j(\hat{p_j}(1-\hat{p_j}))]}$ where m_j is the number of observations in the jth bin, p_j is the observed proportion of successes in bin j, and $\hat{p_j}$ is the mean predicted probability of success for observations in bin j. [4]

For a binned residual plot, it is important to have both a sufficient number of observations per bin so that the averaged residuals are not too noisy as well as sufficient number of bins so that the viewer can easily assess the structure of the residuals (i.e. curvature, outliers). In the remainder

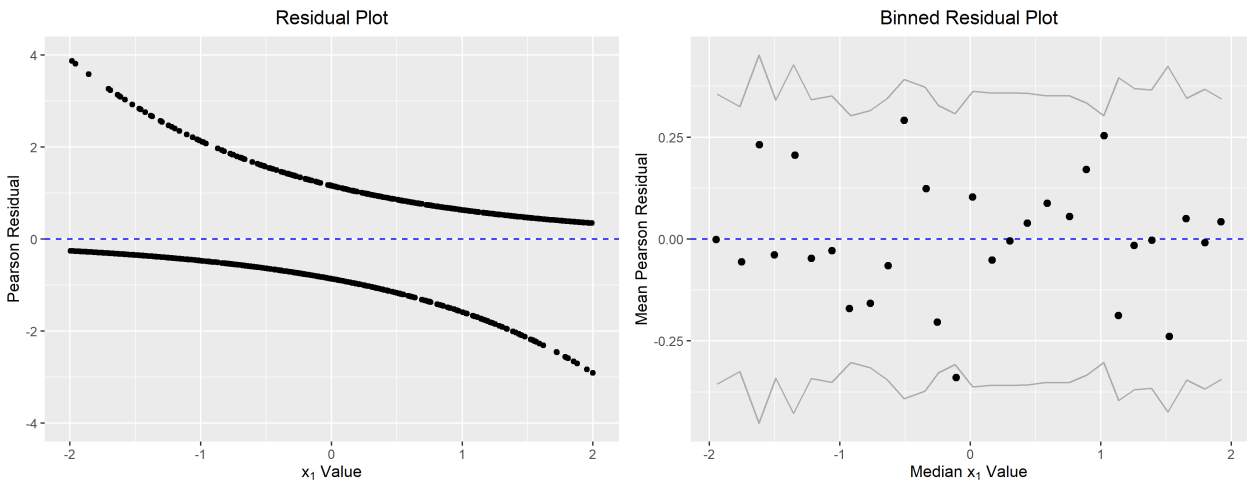


Figure 2: A comparison of a residual plot and binned residual plot. The residual plot is difficult to interpret due to clustering of points. Binning x_1 values creates an interpretable plot.

of this article, the number of bins is determined by taking the square root of the total number of observations unless otherwise specified. This is generally a good compromise between the two objectives and performs well for most sample sizes.

2.3 Empirical Logit Plots

Empirical logit plots are used to diagnose log-odds linearity for logistic regression. As with binned residual plots, observations are binned according to values of an explanatory variable. For each bin, the empirical logit is calculated by taking $\text{logit}(Y_j/m_j)$ where Y_j is the total number of successes in bin j and m_j is the number of observations in bin j. Note that from Section 1.2, we showed that this empirical logit is equivalent to the log-odds for group j. The group log-odds are plotted against the average value of the binning variable. Figure 3 shows an example empirical logit plot for a correctly specified model binned using explanatory variable x_1. In logistic regression, predictors are linearly related to log-odds (the empirical logit); therefore any non-linear relationship in the empirical logit plot indicates a violation in the linearity assumption. [10]

2.4 Goodness-of-fit Test

In binomial logistic regression, when there are replicate values for each explanatory variable in the model due to groups, a common diagnostic technique is to use the deviance goodness-of-fit test on the model. The deviance Goodness-of-fit test compares the model of interest to the saturated model using the predicted proportion of success (from the model of interest) and the observed proportion of success (from the saturated model). The test uses the $(n - p)$ additional parameters in the saturated model to calculate the deviance statistic (the sum of the squared deviance residuals from the model of interest) and gives a p-value indicating if the fitted model adequately characterizes the sample. Equivalently, the deviance statistic is defined as the log-likelihood of the fitted model minus the log-likelihood of the saturated model multiplied by negative two. For binomial models

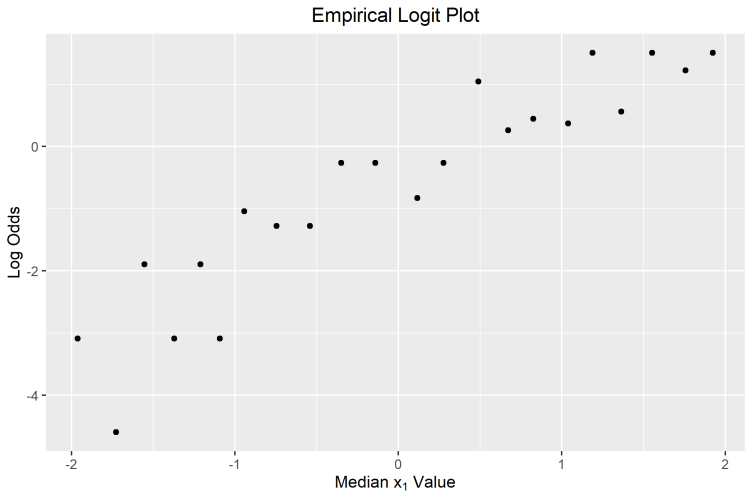


Figure 3: Example empirical logit plot. Binned values of the explanatory variable x_1 are linearly associated with the log-odds of the response.

with sufficiently large group sizes, the distribution of this deviance statistic is approximated by a chi-squared distribution with $(n - p)$ degrees of freedom. [10]

The Goodness-of-fit test begins to fail when the observations are binary instead of binomial. Firstly, the chi-squared approximation of the deviance statistic assumes large group sizes, but binary logistic regression has all group sizes equal to one (there are no groups). This implies that the distribution of the deviance statistic cannot be approximated with the chi-squared distribution and needs to be simulated. However, simulation also proves ineffective because the deviance statistic is not necessarily pivotal in binary logistic regression. Because there are no replicate observations, the saturated model has log-likelihood equal to zero. This causes the deviance to be centered at the log-likelihood of the fitted model, which is in turn approximately equal to the likelihood of the true model.

Because deviance is not necessarily pivotal, it is not a valid option to use the deviance Goodness-of-fit-test. Instead we need to use the less common Pearson Goodness-of-fit test. This test operates the same as the deviance version but calculates a Pearson statistic equal to the sum of squared Pearson residuals instead of the deviance statistic. The Pearson statistic is pivotal and thus can be simulated using the techniques discussed in Section 3. Moving forward, any time the Goodness-of-fit test is referenced in this article, we refer to a Pearson Goodness-of-fit test where the distribution of Pearson statistics is simulated. Note that although some version of the Goodness-of-fit test exists for binary data, it is more difficult to compute than for binomial data. Additionally, because deviance is not pivotal, simulating a deviance Goodness-of-fit test gives inflated p-values and wrongly indicates that the fitted model is adequate. The added simulation difficulties and possibility for incorrect inference are some of the motivating reasons for pursuing alternative tests for general model misfits.

3 Simulation-Based Diagnostics for Logistic Regression

All of the visual diagnostic plots mentioned in Section 2 have implicit rather than explicit reference distributions. For a binned residual plot, we expect residuals to be independent and distributed around zero. For an empirical logit plot, we expect a linear association between the log-odds and explanatory variable. However, what does it look like for residuals to be distributed around zero? And how much curvature in a logit plot can be attributed to randomness before we become suspicious that an explanatory variable is non-linear? The diagnostics discussed in this section attempt to answer these questions by simulating null data sets from the fitted model. Because the null sets come from the fitted model, they follow all assumptions of the model. Thus, we can use them to create explicit reference distributions for all of the visual diagnostics discussed above. Additionally, for any test statistic, we can use the null data sets to create a distribution of null test statistics. Comparing the observed test statistic to what would be expected under the model gives a simulated p-value for any test statistic. Simulation thus has potential to improve both visual diagnostics and quantitative tests for model checking. Note that the lineup protocol is an example of a simulation-based visual diagnostic.

3.1 Simulating Null Distributions

There are many possible ways to simulate a null distribution, with the three main methods being (i) conditional sampling, (ii) parametric bootstrap sampling, and (iii) Bayesian posterior predictive sampling [1]. We use conditional sampling when the test statistic can be sampled directly given the null hypothesis. The null data sets from the lineup in Figure 1 use this method. Because the null hypothesis is that data come from a standard normal distribution, we simply simulate data from a standard normal distribution. If the null hypothesis was more general, such as that the data were normally distributed with variance σ^2, conditional sampling would not be appropriate because the parameter σ^2 would need to be estimated. Another common example of conditional sampling is a permutation distribution. When we have two samples, we can calculate a test statistic for every possible permutation of observations within the two groups. In this case, we are simulating every possible combination of observations without having to estimate any parameters. [1]

Parametric bootstrap sampling and posterior predictive sampling are the Frequentist and Bayesian approaches to simulating data from a fitted model, and mirror one another in methodology. Consider the data set y which is used to estimate the parameters θ of a given model. We are interested in sampling from a null distribution of data y^* that come from the specified model. In the Frequentist setting, if the model parameters are estimated using maximum likelihood, then it makes sense to sample a null distribution y^* from the distribution $p(y \mid \hat{\theta})$. This process is known as parametric bootstrap sampling and gives an approximate posterior predictive distribution [1]. The maximum likelihood estimates for the parameters $\hat{\theta}$ contain uncertainty that parametric bootstrap sampling does not account for. In the Bayesian setting, we can skip the step of estimating θ and sample directly from the posterior predictive distribution $y^* \mid y$ using Monte Carlo approximation. If we draw a single posterior sample of the estimated parameters $\theta^* \sim p(\theta \mid y)$ and use this draw to take a prediction from the likelihood, $y^* \mid \theta^* \sim f(y \mid \theta^*)$, then y^* follows the posterior predictive distribution [11]. Monte Carlo simulation repeats this process many times so that data are simulated using different values for the model parameters. Because Bayesian posterior predictive sampling propagates uncertainty about the model parameters, the variance of the posterior predictive distribution

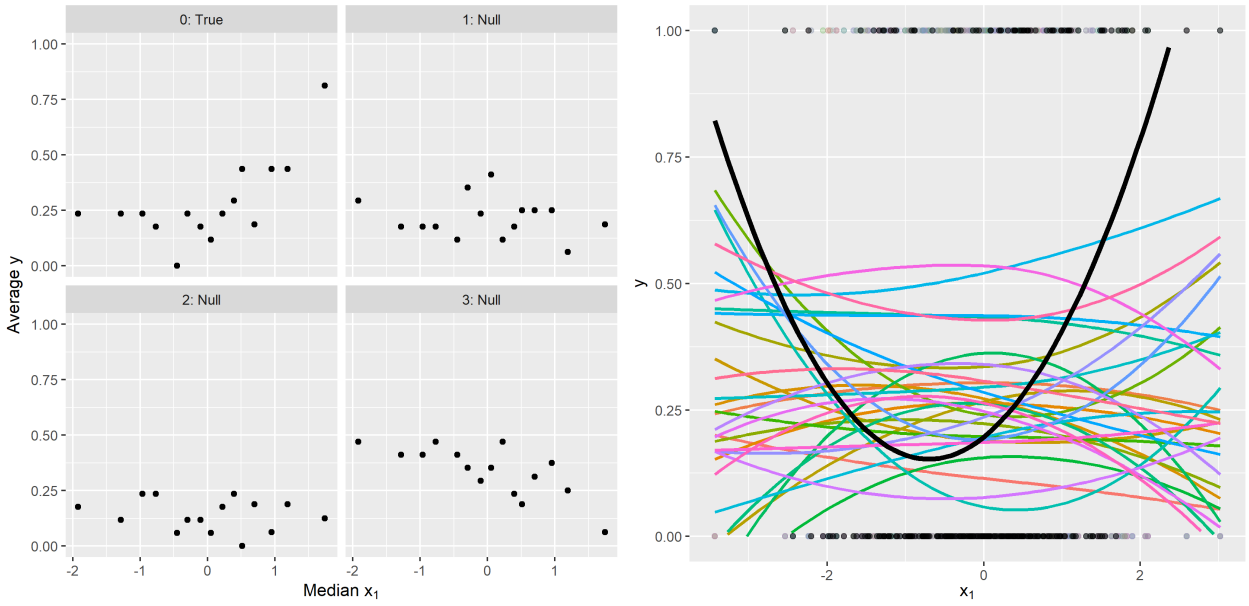


Figure 4: Visual diagnostics comparing observed binary data to null data simulated from the fitted model using Bayesian posterior predictive sampling. On the left, data are binned by x_1 values and compared to null distributions. On the right, the true and null data sets are visualized with smoothers on the same plot. The true data has a black curve, and null data sets have colored curves.

will always be greater than or equal to the variance of the posterior predictive distribution approximated with parametric bootstrap sampling [11]. Logistic regression is not a particularly complicated model, especially when there are few explanatory variables, and thus the parametric bootstrap is a good approximation of the posterior predictive distribution.

3.2 Logistic Regression - Visual Diagnostics

Gelman (2004) [2] argues that the most basic visual diagnostic for a model is a display of the entire data set, compared against reference distributions of data simulated from the fitted model. This is difficult in the case of binary logistic regression, as data take response values of 0 or 1 and are clustered on top of one another and difficult to interpret. There are a couple ways to address this issue. We could bin observations according to an explanatory variable x_1 in the same manner that we used for binned residual plots and empirical logit plots in Section 2. Then, instead of plotting y against x_1, we would be plotting a binned average of y (equivalently the binned probability of success) against the average value for x_1. Alternatively, we could plot the raw observations and use a smoother to approximate the probability of success at a given value of x_1. For either of these approaches, we would then simulate null data from the model using either parametric bootstrap sampling or Bayesian posterior predictive sampling in order to create reference distributions. These reference distributions are what the data should like under the given model.

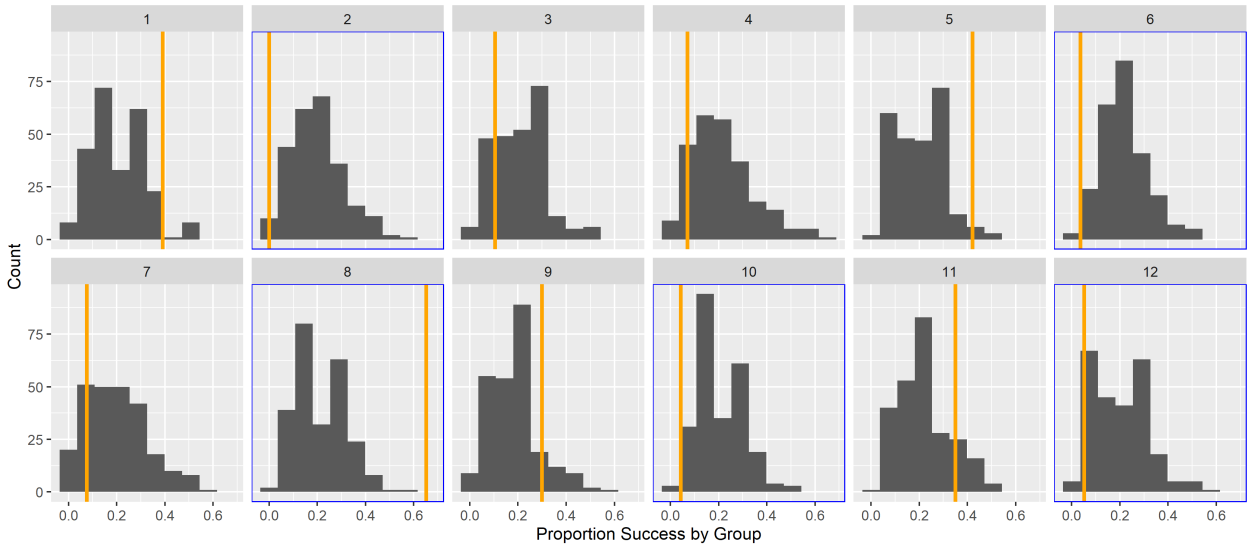
Figure 4 shows both of these approaches for a logistic model that is missing a significant squared term for the explanatory variable x_1. Note that in the binned data visualization shown on the left, it makes sense to display the true plot and reference plots side by side, in a manner reminiscent of the

lineup protocol. Unlike the lineup protocol, the plot made using the true data is clearly identified. Instead of trying to pick the true plot out from the nulls, the null plots are used as references for the true plot. In the display, it appears that the true plot has a larger average probability of success as the median value of x_1 increases compared to the null plots. This is indicative of the missing squared term from the original model. The smoother approach is shown on the right of Figure 4. Because lines are easier to distinguish from one another than points, we are able to utilize color and plot the true and null data sets in the same display. The true plot has a bold black smoother while null plots have thin smoothers of various colors. Note that for upper values of x_1, the observed probability of success is larger than in the null data sets, again hinting at the missing squared term for x_1. It should be noted that although the smoother approach is visually more difficult to interpret, it is able to compare the true data set to 30 null sets at once while the binned approach only compares the true data set to 3 null sets.

3.3 Logistic Regression - Quantitative Tests

The above visualizations are essentially looking at how the proportion of successes for y changes in relation to an explanatory variable x_1. While visualizations can be helpful in understanding how the true data is different from what is expected given the model, they do not provide an indication of if this difference is statistically significant. One general method for assessing statistical significance is the lineup protocol. However, if an analyst has already started exploratory data analysis, they are likely familiar with the data and are no longer an unbiased observer. When the lineup protocol is not an option, posterior predictive checks are an alternative way to assess significance [2] [11]. A posterior predictive check calculates some collection of test statistics for the observed data set. Then, those same test statistics are computed for null data simulated from the fitted model (typically using posterior predictive sampling). These simulated test statistics create theoretical distributions to which the observed test statistics can be compared. We then calculate a simulation-based p-value by comparing the observed test statistics to the null distributions of test statistics.

It is important to choose a collection of test statistics according to a null and alternative hypothesis. For example, suppose that we believe that data are not independent; rather, we believe that data are correlated within some grouping variable. Here, the null hypothesis is that data are independent, and the alternative is that data are correlated by groups. For these hypotheses, a reasonable collection of test statistics would be to bin the data by the supposed grouping variable and calculate the proportion of successes within each bin. If data are independent within groups, the groups should not have any effect on the proportion of success. This example is illustrated in Figure 5. Each panel represents one section of the grouping variable. We use the simulated null data sets to calculate the collection of test statistics (the proportion of successes for each group). These distributions of test statistics are expressed as histograms, and the true proportions of success for each group are plotted as vertical lines [2]. The table below Figure 5 shows the p-value by group. Because we are interested in both exceptionally high and exceptionally low values, any p-values less than 0.025 or greater than 0.975 are deemed significant and highlighted in blue. This is equivalent to a two-sided confidence interval. Note that 5 of the 12 groups had significant p-values. If the model had no violations, we would expect these p-values to be uniformly distributed. Assuming that these group p-values are independent, we calculate an overall p-value for our independence assumption using a binomial distribution. Assuming that data are independent, p-values are uniformly distributed on (0,1) and hence each has probability of 0.05 for being significant. So we have $P(X \leq 5)$ with



Group	1	2	3	4	5	6	7	8	9	10	11	12
P-value	0.91	0.00	0.06	0.04	0.96	0.01	0.08	1.00	0.83	0.01	0.82	0.02

Figure 5: A visualization of the collection of test statistics used to determine if data are independent or correlated among groups. Histograms are the distributions of null test statistics for a given group with vertical lines indicating the observed test statistics. The table gives the p-values for each group, with significant groups being highlighted in blue.

$X \sim \text{Binom}(12, p = 0.05)$ is equal to 0.000173. Thus, we reject the null hypothesis that data are independent in favor of the alternative that data are correlated among groups.

If we believe that the data violate a specific model assumption, this type of Bayesian posterior predictive check gives useful insight into the data, even if the model is fit in the Frequentist setting. However, this insight is completely dependent on the types of test statistics that we calculate. If we had binned the data by the value of x_1 instead of using the underlying group structure, we would not have found any significant model violation. This points back to the issues with quantitative testing discussed in Section 1. Quantitative testing is only as powerful as the test statistics that we choose to calculate. The lineup protocol can be seen as compromise between visual and quantitative diagnostics. On the one hand, the lineup protocol provides a quantitative p-value that we can interpret as significance, but at the same time, the reason for choosing a plot provides insight as to the how an assumption is violated.

The remainder of the article focuses on the effectiveness of the lineup protocol in binary logistic regression. In Section 5, we compare the power of the lineup protocol to the power of the Pearson Goodness-of-fit test, using a power study introduced in Section 4. Then, in Section 6, we compare the power of lineup designs that display data using different types of diagnostic plots.

4 Lineup Protocol Power Study for Binary Logistic Regression

The power study considers the effectiveness of the lineup protocol for different sample sizes, different types of model violations, and using different types of diagnostic plots. As mentioned in

Section 2, many classical diagnostics exist for logistic regression. However, there are a handful of situations where these diagnostic plots can be difficult to interpret and where quantitative tests begin to lose reliability. One such situation is when there is a small sample size relative to the number of parameters being estimated in the logistic model. Especially in binary logistic regression, small sample sizes can lead to difficulties when binning observations. Smaller sample sizes can cause either an insufficient number of observations per group so that the binned averages are noisy, or too few groups overall so that it is difficult to detect trends between bins. We are interested in instances where the lineup protocol might outperform classical quantitative tests, and thus the study focuses on data with small sample sizes. Specifically, we consider data that have sample sizes $n \in \{150, 250\}$.

Recall from Section 1 that the two assumptions of binary logistic regression are log-odds linearity and independence of observations. Thus, we will create lineups for data that violate these assumptions to varying degrees. The study investigates six unique types of model violation detailed below. All models are fit under the assumption that a single explanatory variable x_1 is linearly related to the log-odds of success for the response variable y and that all observations are independent. Thus the fitted model is given by

$$P(y_i = 1) = \text{logit}^{-1}(\alpha + \beta_1 x_{1i}). \tag{9}$$

Two of the model violations used in the study address non-linearity of log-odds due to a missing squared term for the variable x_1, three address correlation of observations due to a group random-intercept term not captured by the model, and one serves as a control where data follow all assumptions of binary logistic regression and there are no model violations.

Violation 0	**No violation:** Observations are independent and x_1 is linearly related to the log-odds of y. The model is correctly specified.
Violation 1	**Severe Non-linearity:** Observations are independent, but the log-odds of y are correlated with $(x_1)^2$ rather than x_1. The true β coefficient for the $(x_1)^2$ term in the model is large.
Violation 2	**Mild Non-linearity:** The situation is identical to Violation 1, but the true β coefficient for the $(x_1)^2$ term in the model is half that of in Violation 1.
Violation 3	**Severe Group Effects:** The variable x_1 is linearly related to the log-odds of y, but observations are not independent. Observations are randomly assigned to $\lfloor\sqrt{n}\rfloor$ groups with replacement. Each group has a random intercept that is normally distributed around zero with variance σ^2.
Violation 4	**Mild Group Effects:** The situation is identical to Violation 3, but the random intercepts are normally distributed around zero with variance $\sigma^2/2$.
Violation 5	**Severe Group Effects, Few Observations per Group:** The situation is identical to Violation 3, but there are $\lfloor\sqrt{n}\rfloor * 2$ total groups.

Finally, the simulation study creates lineups using three types of visual diagnostics. The plots are shown left to right in Figure 6. The first visual diagnostic is a simple binned residual plot (BR), the second is a binned residual plot that includes 95% error bounds (BR95), and the third is an empirical logit plot (EL). Lineups are created using different visual diagnostics to assess if lineups made with certain plot types perform better with certain model violations. For example, in exploratory data analysis, the empirical logit plot's primary purpose is to diagnose linearity violations for an explanatory variable. Thus, we might expect lineups that use empirical logit plots to be more effective at detecting linearity violations than independence violations. We include

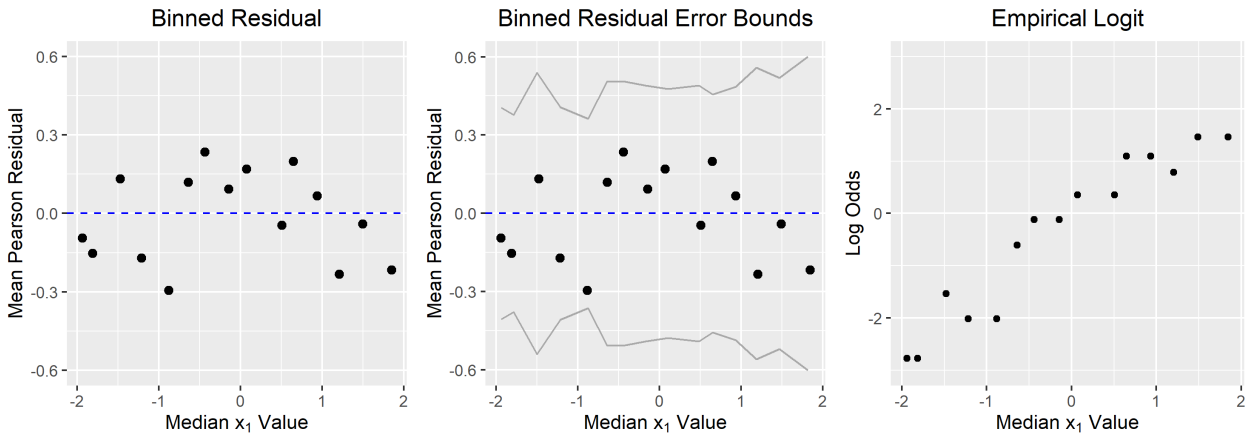


Figure 6: Examples of the three diagnostic plots used in the simulation study. The examples use data that follow all assumptions of logistic regression.

binned residual plots with and without a 95% error bar because the approximation method used to calculated the error bars is less reliable at smaller sample sizes. If error bars are sporadic at small sample sizes, they might distract the viewer from the focus of the plot - the binned residual. Note that when we assess violations of independence due to group correlation, we bin observations by the suspected group variable instead of binning by the explanatory variable value. Because the alternative hypothesis is that data are correlated in groups, we need to provide a visual that clearly differentiates between groups. In this case, we can think of bins for data being given explicitly, just as they are in binomial logistic regression. All lineups for the study are created using the R package `nulllabor` as recommended by Buja et. al. (2009) [1].

The simulation study considers all possible combinations of sample size and violation type. We draw 2 samples from each combination of the 2 sample sizes and 6 violations for a total of 24 samples. These samples act as the true data in the lineups. For each true data sample, 19 null data sets are simulated using parametric bootstrap sampling according to the fitted model described above. We simulate 2 sets of null data for each sample for a total of 48 lineup data sets. Using the 48 lineup data sets, we create a lineup using each of the 3 visual diagnostics. This comes out to a total of $48 * 3 = 144$ total lineups.

We expect participants to perform better on lineups in which data severely violate an assumption of the model, rather than data which mildly violate a model assumption. Formalizing this, we expect participants to pick the true plot more frequently in violation 1 than violation 2, and more frequently in violation 3 than violation 4. Violation 5 is unique because it examines the trade-off between observations per bin and total number of bins. We expect that in violation 5, the reduced number of observations per bins will cause the average binned residuals to be more sporadic in both the true and null plots and cause participants to have a harder time choosing the true plot compared to violation 3. Finally, we expect that participants will not be able to pick out the true data in violation 0 because the model is correctly specified. Additionally, we expect that increasing the sample size will always increase performance on a lineup because there are more total bins as well as more observations per bin, reducing the overall variability.

Using the online survey recruitment platform, Prolific [9], 361 independent viewers were recruited

and asked to view 12 lineups each. Participants randomly evaluated lineups such that each participant saw i) a true sample no more than once, ii) 2 lineups from each violation type, iii) 6 lineups from each sample size, and iv) 4 lineups using each type of visual diagnostic. For each lineup, participants were asked 1) if they believe that a plot is different from the others, 2) which plot is the most different from the others, and 3) the reason for their choice (from a list of options).

The practice of recruiting participants to view lineups remotely is well documented, typically through the online recruitment platform Amazon MTurk [5] [6] [7]. The current study presents Prolific as a more diverse and ethical alternative to MTurk. In Peer et al. (2017) [8], data collected using Prolific was shown to be of equal quality to data collected with MTurk. Additionally, Prolific requires an adequate wage for its workers, takes a smaller commission this wage, and has a larger active participant pool than MTurk [9]. To ensure that participants were completing the survey to the best of their ability, screening questions were inserted randomly through the survey and any responses submitted in under 3 minutes were removed from the study. Additionally, all participants had at least a high school diploma, had participated in at least 10 other studies on Prolific, and had an approval rate over 75% for these previous studies.

4.1 Preliminary Results

Figure 7 shows the proportion of viewers that correctly chose the true plot for each lineup. Each point represents one of the 144 total lineups, and each line connects lineups generated with the same true data set and same null data sets. Panels correspond to the situation in which data were simulated, a combination of sample size and violation type. Within each panel, lines of the same color and points of the same shape contain the same true data set. For each true set of data, the study considered two null data sets, and thus there are two lines of each color within a panel. Finally, a filled point indicates that the null hypothesis (that the data are independent or that the data are log-odds linear) was rejected with a p-value less than $\alpha = 0.05$.

4.2 Effect of Simulation Parameters

In our discussion, we call a lineup significant if it has a significant p-value less than 0.05 and resulted in the rejection of the null hypothesis. Note that generally, lineups that share the same true data are closer together, especially when they use the same visual diagnostic. The proportion of viewers to choose the correct plot for a lineup varies largely depending on the type of visual diagnostic used; however, if a lineup is significant, other lineups with the same true and null data that use different visual diagnostics are also frequently significant. In regards to sample size, for all violations with $n = 250$, it was the case that either both true samples have at least one significant lineup, or neither true sample has a significant lineup. With this larger sample size, participants consistently performed worse when the violations were mild, as expected. Compare this to smaller sample sizes. When $n = 150$, it is frequently the case that only one of the two samples had a significant lineup. Additionally, at the smaller sample size, both of the mild violations had one sample that contained significant lineups, while violation 3 had no significant lineups even though it was a severe violation. This raises questions about the reliability of the lineup protocol at small sample sizes. Why was it the case that the lineup protocol was able to detect mild violations when $n = 150$, but not when $n = 250$? Perhaps the larger sample size was not large enough of a difference for there to be a significant change in the difficulty of lineups with small and large sample sizes.

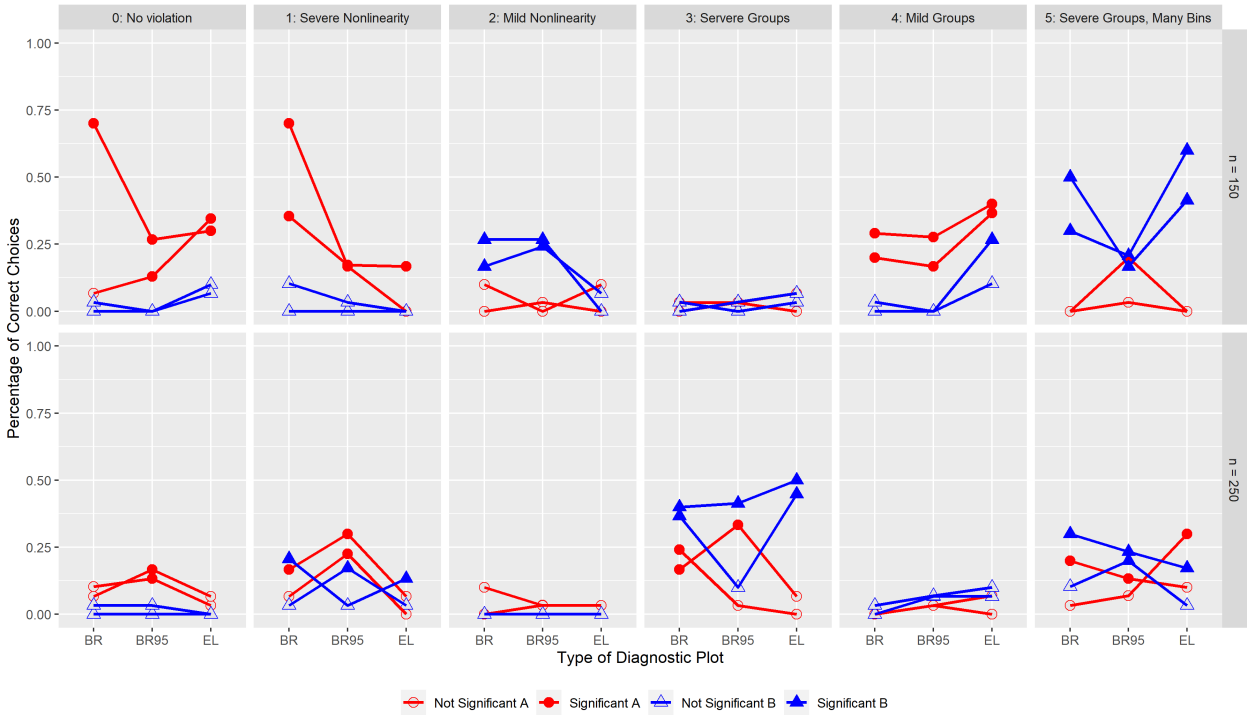


Figure 7: A faceted plot showing proportion of viewers to choose the true plot. Points represent an individual lineup, with lines connecting lineups that share both true and null data sets. Within a panel, color and shape of points indicate which lineups share a true data set. A filled point indicates the rejection of the null hypothesis. BR = binned residual, BR95 = binned residual with 95% error bounds, EL = empirical logit.

4.3 Strong Linearity within Residual Plots

It was observed repeatedly in the study that participants were drawn to panels where a small number of points in plot appeared on a strongly defined line. We provide two examples of such behavior that caused large numbers of participants to choose panels that did not violate the null hypothesis. Consider the lineup in Figure A1 of Appendix A, which uses binned residual plots without error bounds, has a sample size of 150, and has no model violation. Contrary to the fact that there was no model violation, the true plot was correctly identified by 21 of 30 viewers (shown in panel $(3^2 + 2)$). Note that the true plot has a large number of points in the middle of the plot that appear extremely linear. Of the 21 participants who chose the plot, 14 referred to a curve or line when describing their reason for choosing the plot and 2 participants used the word "linear" explicitly. This tendency of participants to choose plots with subsets of linear points could help explain the large Type-I error that was observed for the lineup protocol in the sample used to make Figure A1. A second lineup that demonstrates this tendency is shown in Figure A2 of Appendix A with the true plot displayed in panel $(2 * (4 + 2))$ which was correctly identified by 0 of 29 viewers. Here the data had $n = 250$ with mild non-linearity, and the lineup was made with binned residuals without error bounds. The most commonly selected plot in this lineup is displayed in position 13

(selected by 6 of 29 viewers), and again has a strong linear appearance in the center of the plot. One of the participants who chose plot 13 gave the reason, "points are close together on top left quadrant," referencing the position and closeness of the points on the line. Because we did not expect participants to choose plots that exhibited strong subsets of linear points, there was not an option in the drop-down list that referenced linearity or lines. When providing a reason for their choice, participants had to select "other" from the drop-down and manually enter their thoughts on linearity. In future studies, it would be interesting to include a linearity option in the drop-down to more accurately assess the thought process of participants. The tendency of viewers to select strongly linear plots was not observed in any of the lineup studies we use as references and thus could be a potential area of future research [5] [6] [7].

5 Power: Lineup Protocol vs. Goodness-of-fit

We say that a diagnostic test is more powerful than another test if given the same sample, one diagnostic is able to reject the null hypothesis more frequently when there is a model violation. Table 1 gives a comparison of the number of lineups that were rejected using the lineup protocol and the simulated Pearson Goodness-of-fit test. Recall that within each type of violation, there were 4 unique samples - each sample creating 6 lineups (2 nulls and 3 plot types). Because the Goodness-of-fit test depends only on the sample, it either rejects the null hypothesis for all or none of the 6 lineups from the same sample. This is reflected in the table by the rejections for the Goodness-of-fit test increasing in increments of 6.

Table 1: Instances the null was rejected for individual violations by diagnostic test.

	Viol. 0	Viol. 1	Viol. 2	Viol. 3	Viol. 4	Viol. 5
Lineup Protocol	7/24	11/24	4/24	8/24	7/24	14/24
Pearson GoF	0/24	24/24	12/24	6/24	0/24	0/24

Table 2 contains the same data as Table 1, but combines rejections in terms of the type of violation (control, log-linearity, or independence). Note that the lineup protocol had 7 false rejections of the null hypothesis whereas the Pearson Goodness-of-fit test had no false rejections, providing evidence that the lineup protocol might have a larger Type-I error rate than the Goodness-of-fit test. In the study, the Pearson Goodness-of-fit test was more powerful at detecting log-linearity violations than the lineup protocol, rejecting the null for 36 of 48 lineups (6 of 8 samples) at a rate more than twice that of the lineup protocol (15 of 48 lineups). However, the lineup protocol was more powerful at detecting independence violations, rejecting 29 of 72 independence lineups while the Pearson Goodness-of-fit test rejected only 6 of 72 lineups (1 of 12 samples).

Table 2: Instances the null was rejected for type of violation by diagnostic test.

	No Violation	Log-Linearity	Independence
Lineup Protocol	7/24	15/48	29/72
Pearson GoF	0/24	36/48	6/72

Figure A4 in Appendix A shows a lineup created using a sample in which the Pearson Goodness-of-fit test was able to detect a violation of log-linearity when the lineup protocol could not. The

true plot is shown in position $(2^0 + 2)$ and was selected by only 1 of 30 viewers. This results in a p-value of 0.3389 using the lineup protocol while the Goodness-of-fit test has a p-value of 0.00491. The lineup illustrating the reverse situation is shown in Figure A5 in Appendix A. The lineup has $n = 150$ and severe group effects with many groups, the violation type in which the lineup protocol most outperformed the Pearson Goodness-of-fit test. The true plot is in position $(2 * \sqrt{25})$ and was picked by 18 of 30 viewers. For this sample, all lineups, regardless of null or plot type, resulted in the rejection of the null hypothesis while the Pearson Goodness-of-fit test has a p-value of 0.374. In creating the violation types, we had assumed that adding more groups would increase the difficulty of the lineup overall, but this was not the case. We observed that increasing the number of groups also increased the power of the lineup protocol for severe group effects. If these results can be generalized to other types of model violations, it could be more effective to use diagnostic plots with more bins than is generally recommended in classical inference.

6 Power: Type of Diagnostic Plot

In addition to comparing the power of different diagnostic tests, the study compares the power of different lineup protocol designs. A lineup design is more powerful than another if given the same true and null data sets, one design results in the rejection of the null hypothesis more frequently when there is a model violation. A design having high power implies that it presents data to the viewer more effectively and clearly than a design with low power. Take for example the two types of binned residual plots used in the simulation study. Adding theoretical error bars to a graph increases the total information that is displayed in the plot. But is the additional information displayed in error bounds useful for making inferences? Or is it the case that error bounds clutter the plot and make it more difficult for a viewer to interpret the underlying data? Table 3 compares how frequently lineups resulted in the rejection of the null hypothesis based on the type of diagnostic plot that they used. Note there does not appear to be a large difference in the rejection rates for individual violations between lineups that use different diagnostic plots.

Table 3: Instances the null was rejected for individual violations by type of plot.

	Viol. 0	Viol. 1	Viol. 2	Viol. 3	Viol. 4	Viol. 5
Binned Residual	1/8	4/8	2/8	4/8	2/8	4/8
Binned Residual 95% Bounds	4/8	5/8	2/8	2/8	2/8	6/8
Empirical Logit	2/8	1/8	0/8	2/8	3/8	4/8

Again, grouping by the assumption that is violated allows us to make stronger claims about the power of the plot types. Table 4 shows combined rejection rates for each type of diagnostic plot. Note that all types of plot have instances in which the null hypothesis was wrongly rejected. However, the binned residual plot with error bounds has the largest Type-I error, rejecting half of all lineup with no violation. In terms of log-linearity model violations, both types of binned residual plots outperform the empirical logit plot. For independence violations, all types of diagnostic plots have similar rejection rates.

To illustrate the concept of one lineup design using data more effectively than another, we include two lineups that use the same true and null data sets in Figures A5 and A6 in Appendix A. Both lineups have severe non-linearity violations with a sample size of 250. Figure A5 uses binned

Table 4: Instances the null was rejected for type of violation by type of plot.

	No Violation	Log-Linearity	Independence
Binned Residual	1/8	6/16	10/24
Binned Residual 95% Bounds	4/8	7/16	10/24
Empirical Logit	2/8	1/16	9/24

residual plots with error bounds and displays the true plot in position $(11 - \sqrt{4})$, picked by 9 of 30 participants. Figure A6 uses empirical logit plots and displays the true data in position $(\sqrt{49} + 12)$ which was picked by only 2 of 30 participants. Here, the null hypothesis is rejected by the binned residual lineup (p-value $= 9.51 * 10^{-6}$) but not by the empirical logit lineup (p-value $= 0.258$). The study provides evidence that this example can be generalized, and that binned residual lineups are more powerful than empirical logit lineups for violations of log-odds linearity. This result is surprising, as in classical diagnostics, the empirical logit plot is designed specifically to assess log-odds linearity. We conjecture that it could be the case that a residual plot is easier for an untrained viewer to interpret than an empirical logit plot, leading to a higher power in visual inference. This problem could be lessened with the addition of a smoother on the empirical logit plot. A smoother has the potential to highlight the overall trend between the binned explanatory variable and the log-odds (the intended purpose of the plot). In a lineup, graphics must be interpretable to a viewer without instruction or knowledge of the model. This standard is not often present for classical visual diagnostics and could begin to explain discrepancies between the effectiveness of diagnostics in the two settings.

7 Discussion and Conclusion

The study provides evidence that the lineup protocol and Pearson Goodness-of-fit test have unique strengths in binary logistic regression. The Goodness-of-fit test is more powerful at detecting violations of log-odds linearity while the lineup protocol is more powerful at detecting violations of independence due to group correlation. There is also evidence that the Pearson Goodness-of-fit test has a lower Type-I error than the lineup protocol. The study indicates that binned residual plots are more effective at detecting violations of log-odds linearity than empirical logit plots. One drawback of the lineup protocol is the large amount of money required to pay viewers to evaluate the lineups online. If cost was not an issue, there are many follow up questions that we would have attempted to answer in the current article.

- For violations of independence, it would be worthwhile to investigate a more gradual transition between "severe" and "mild" random-effects. Additionally, the study provided evidence that the lineup protocol was better at detecting group effects when there were more groups. What range for the number of groups gives the highest power using the lineup protocol?

- Using approximation to compute the theoretical error bounds for binned residual plots was an ineffective method for displaying the error bounds. Frequently, the error bounds cluttered the graph and made binned residuals difficult to interpret. Gelman (2000) [3] notes that error bounds might need to be simulated for low sample sizes, which we did not consider in the current study. We recommend that future studies simulate theoretical error bounds for binned

residual plots instead of using approximation.

- The current study only compares the Pearson Goodness-of-fit test and the lineup protocol. Posterior predictive checks as introduced in Section 3.3 are an alternative simulation-based diagnostic that produce a p-value and can be used to assess significance. Future research might compare the power of various posterior predictive checks, the lineup protocol, and the Goodness-of-fit test.

- The study simulates all null data sets using parametric bootstrap sampling. The article also mentions Bayesian posterior predictive sampling in Section 3.1, which is an alternative way to simulate null data sets. In general, null samples created using posterior predictive sampling are more varied than samples created with a parametric bootstrap because posterior predictive sampling never estimates the parameters of the model. Is the power of the lineup protocol significantly different using the two approaches of simulating null data?

- In addition to asking participants to choose the plot they believed to be most different, the study recorded i) if a participant believed one plot was different from the others and ii) their reason for selecting the "most different" plot. We did not have time to deeply examine the study data relating to either of these questions. These responses should be examined in extensions of this work.

While conducting the study, we were surprised at the rapid pace that quality data could be collected from participants online. A survey containing 12 lineups was frequently completed by 30 participant in under 3 hours using the online survey recruitment platform Prolific, providing evidence that the lineup protocol could be used in more casual model checking settings. When checking a binary logistic model, we recommend that the analyst begins with classical visual diagnostics such as the empirical logit plot to confirm log-odds linearity for explanatory variables. After fitting the model and refining explanatory variables, the analyst should first view the binned residual plot in the context of a lineup. This way, the analyst has a numeric indicator of whether or not predicted data from the model is similar to the observed data. Because the analyst has already seen some aspects of the true data, they are not an unbiased viewer, and the p-value is not valid in inference. However, the lineup still gives valuable insight into the structure of the observed data and fitted model. From here, the analyst should refine the model as needed through classical and simulation-based diagnostics such as the Pearson Goodness-of-fit and posterior predictive checks. If there are questions about independence of observations, it could be useful to evaluate lineups using participants from Prolific to help make decisions about using alternatives models, such as binomial logistic regression or a hierarchical logistic model. One of the exciting parts of statistical analysis is the creativity needed to investigate data. There is no one correct way to fit and refine a model. However, the lineup protocol and other simulation-based diagnostics are powerful tools that should be incorporated in thorough analyses.

Acknowledgments

All data in this study was collected with approval from Carleton College's internal review board IRB 20-21 030 moranj2. I hugely appreciate the funding given by Carleton's Mathematics and Statistics Department as well as the Towsley Endowment. I would like to extend special thanks to Andy Poppick and Adam Loy for their help answering questions and quick responses throughout the comps process. I include my code to simulate all logistic data in the two .rmd files, `Nonlinearity.rmd` and `Independence.rmd`.

References

[1] Andreas Buja, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F Swayne, and Hadley Wickham. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4361–4383, 2009.

[2] Andrew Gelman. Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, 13(4):755–779, 2004.

[3] Andrew Gelman, Yuri Goegebeur, Francis Tuerlinckx, and Iven Van Mechelen. Diagnostic checks for discrete data regression models using posterior predictive simulations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(2):247–268, 2000.

[4] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[5] Heike Hofmann, Lendie Follett, Mahbubul Majumder, and Dianne Cook. Graphical tests for power comparison of competing designs. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2441–2448, 2012.

[6] Adam Loy, Lendie Follett, and Heike Hofmann. Variations of q–q plots: the power of our eyes! *The American Statistician*, 70(2):202–214, 2016.

[7] Adam Loy, Heike Hofmann, and Dianne Cook. Model choice and diagnostics for linear mixed-effects models using statistics on street corners. *Journal of Computational and Graphical Statistics*, 26(3):478–492, 2017.

[8] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

[9] Prolific. `https://www.prolific.co/`, 2014. Accessed: December 2020.

[10] F. Ramsey and D. Schafer. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Cengage Learning, 2012.

[11] Brian J Reich and Sujit K Ghosh. *Bayesian statistical methods*. CRC Press, 2019.

8 Appendix A: Lineups from Simulation Study

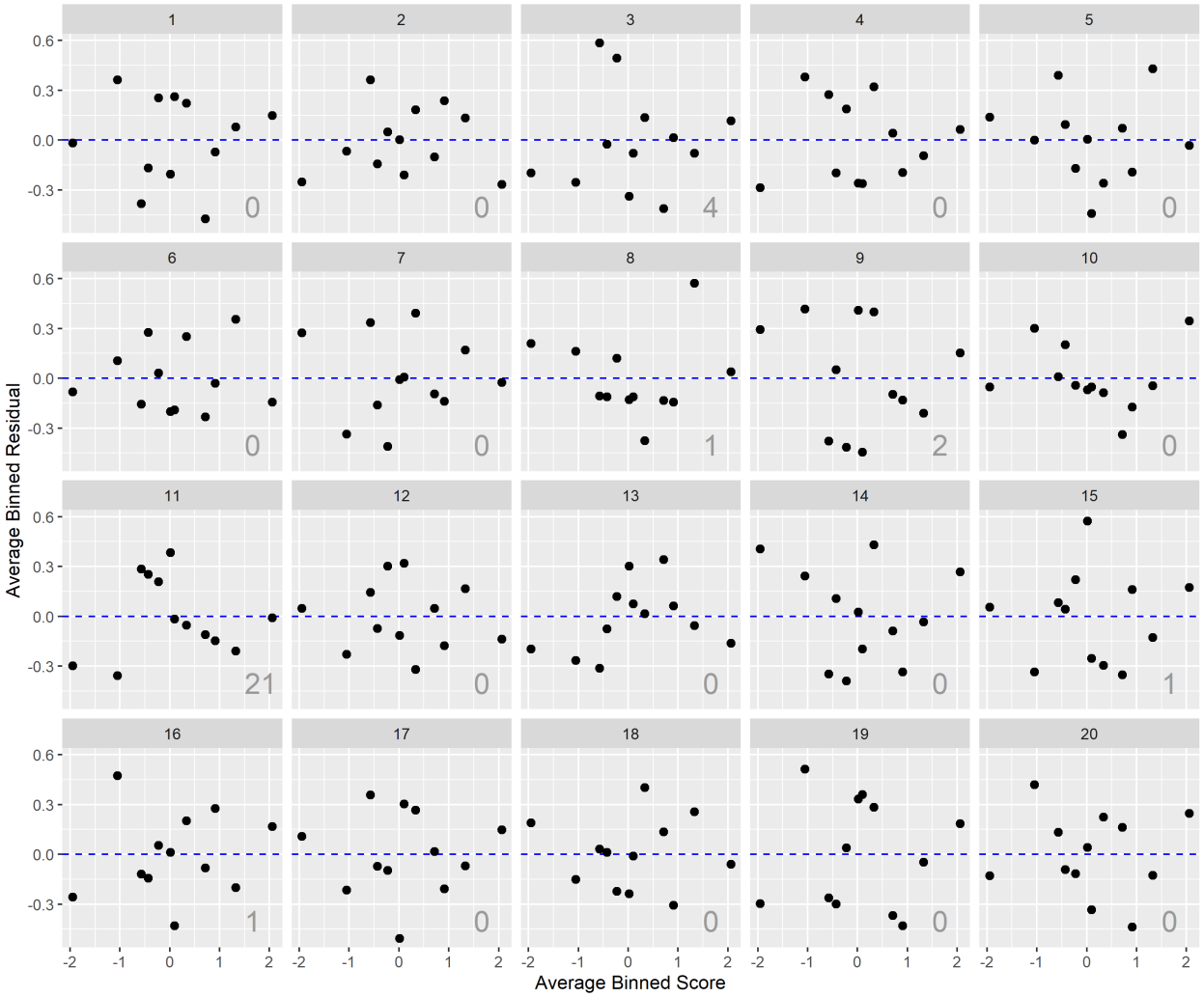


Figure A1: A lineup using binned residual plots created from data with no model violation and a sample size of 150. The true plot is in position $(3^2 + 2)$. The number of viewers to choose a plot is shown by panel, out of 30. Note how points in the true panel appear to lie on a line, even though there is no model violation.

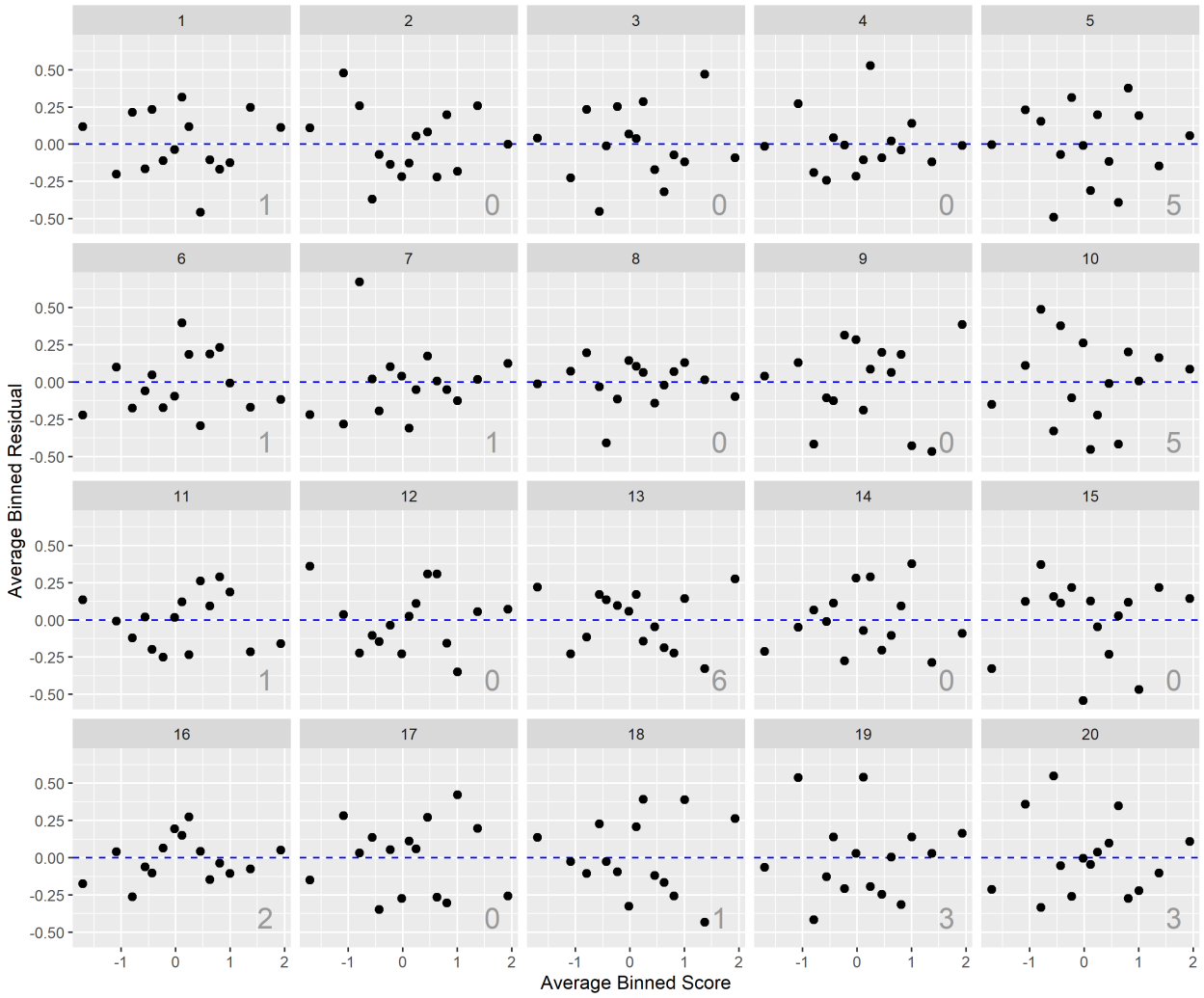


Figure A2: A lineup using binned residual plots created from data with mild non-linearity and a sample size of 250. The true plot is in position $(2 * (4 + 2))$. The number of viewers to choose a plot is shown by panel, out of 29. Note that the null plot in position 13 was selected 6 of 29 times, likely because its points appear to lie on a line.

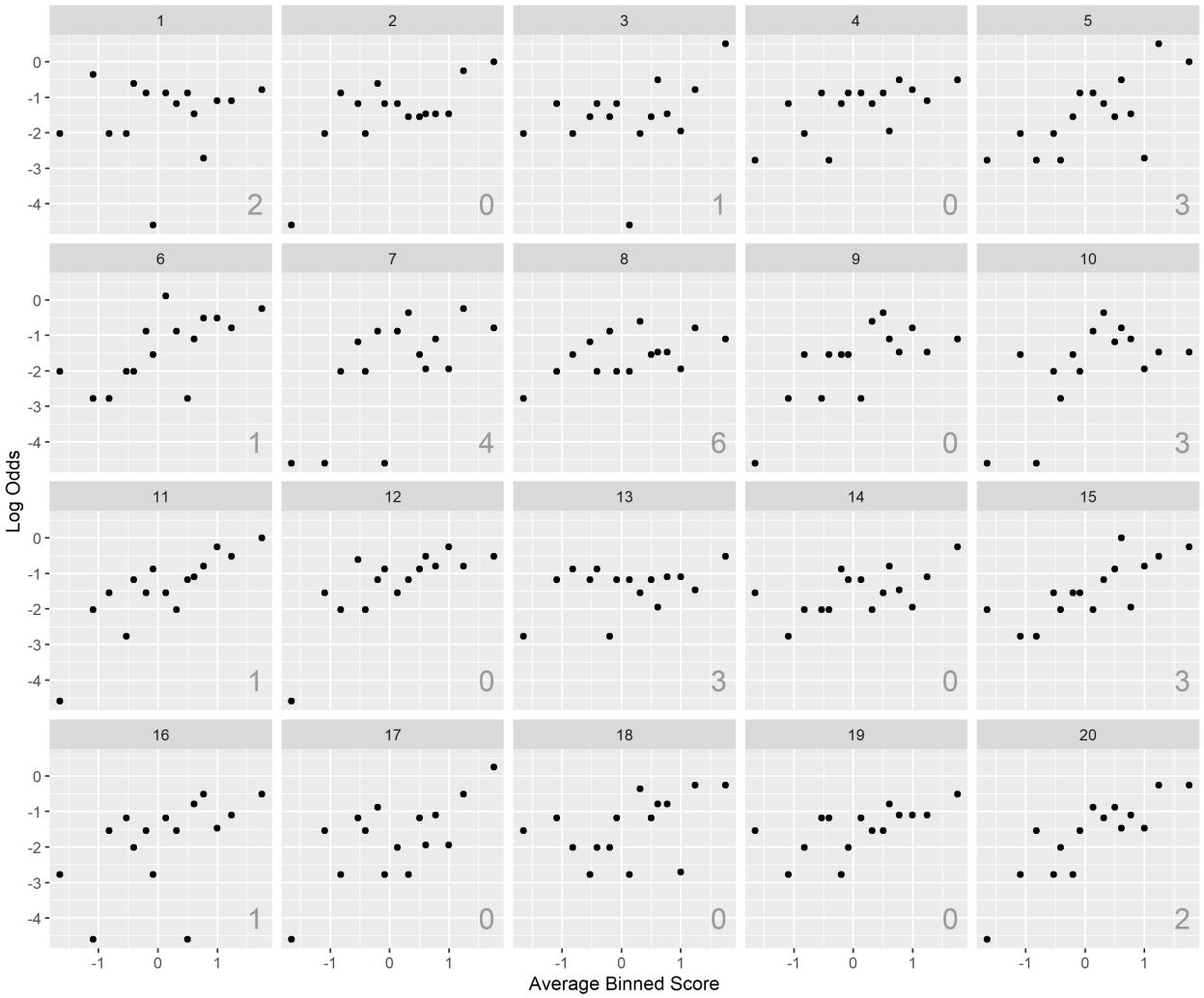


Figure A3: A lineup using empirical logit plots created from data with mild non-linearity and a sample size of 250. The true plot is in position $(2^0 + 2)$, selected by 1 of 30 viewers, giving a p-value of 0.3389. The Pearson Goodness-of-fit test successfully detected the non-linearity for the sample with a p-value of 0.00491.

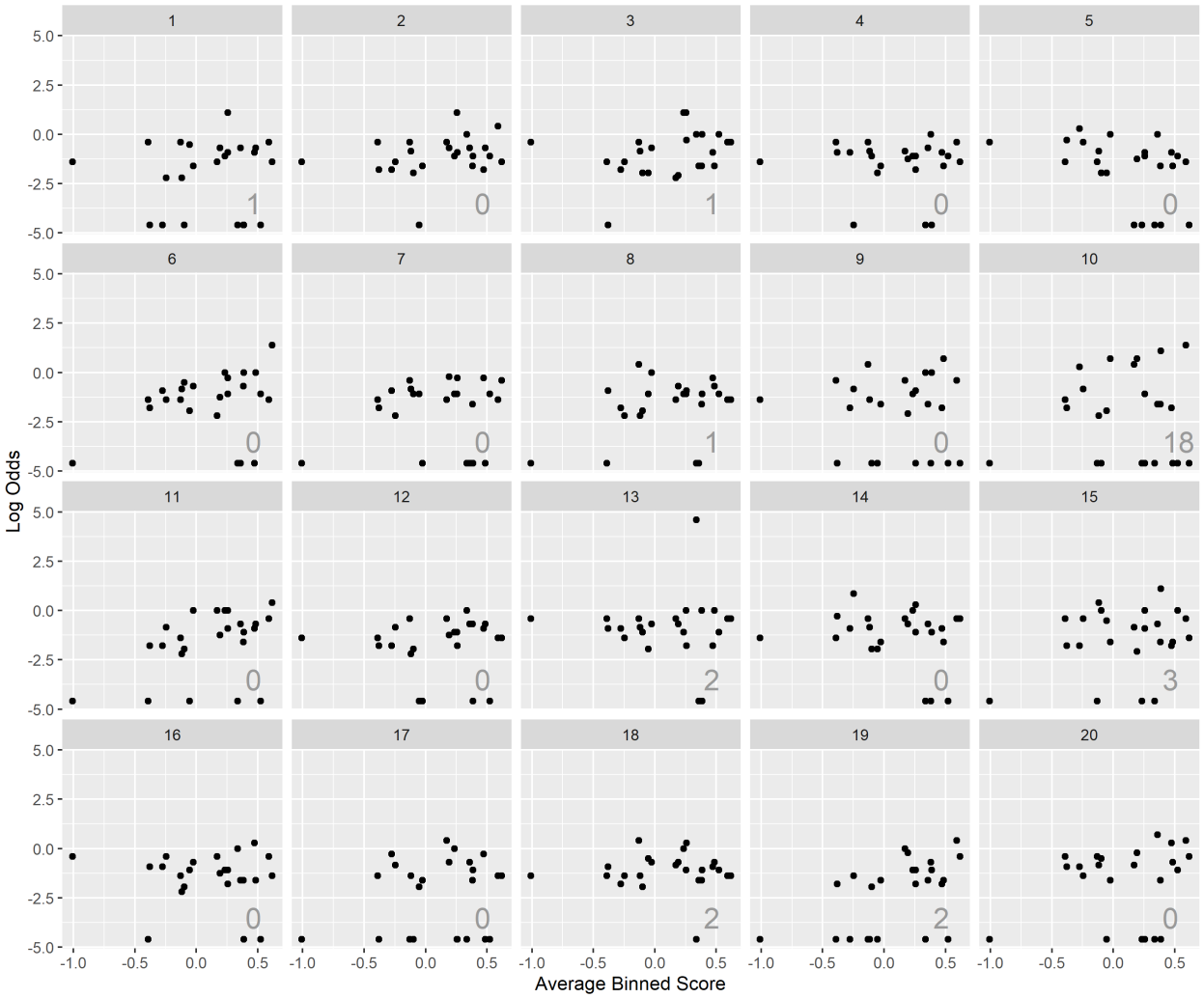


Figure A4: A lineup using empirical logit plots created from data with severe group-effects with many groups and a sample size of 150. The true plot is in position $(2 * \sqrt{25})$, selected by 18 of 30 viewers, giving a p-value of $1.782 * 10^{-16}$. The Pearson Goodness-of-fit test did not reject the null hypothesis for the sample with a p-value of 0.374.

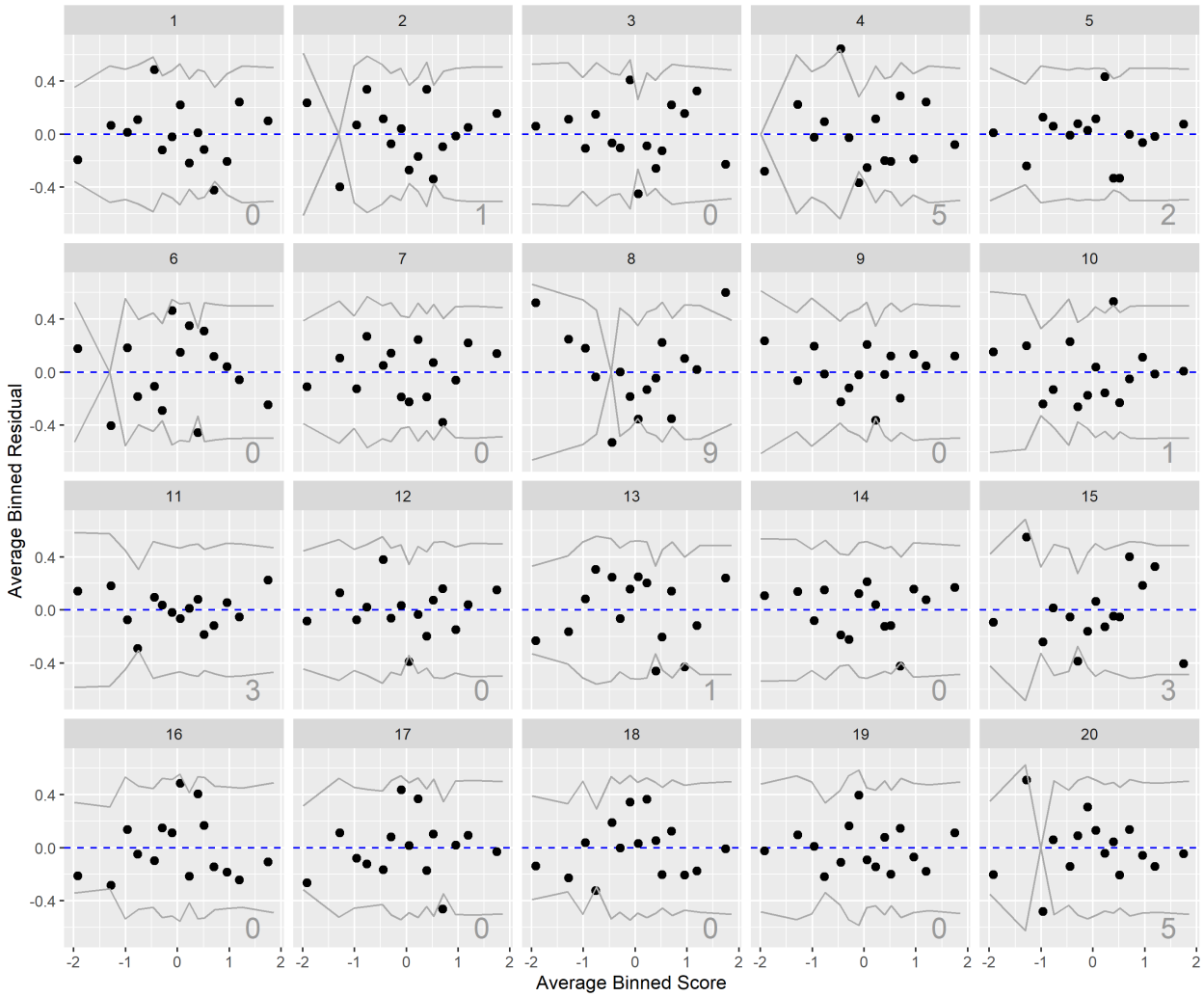


Figure A5: A lineup using binned residual plots with error bounds created from data with severe non-linearity and a sample size of 250. The true plot is in position $(11 - \sqrt{4})$, picked by 9 of 30 participants and giving a p-value of $9.51 * 10^{-6}$. The lineup uses the same true and null data as Figure A6, but does so more effectively and rejects the null whereas Figure A6 does not.

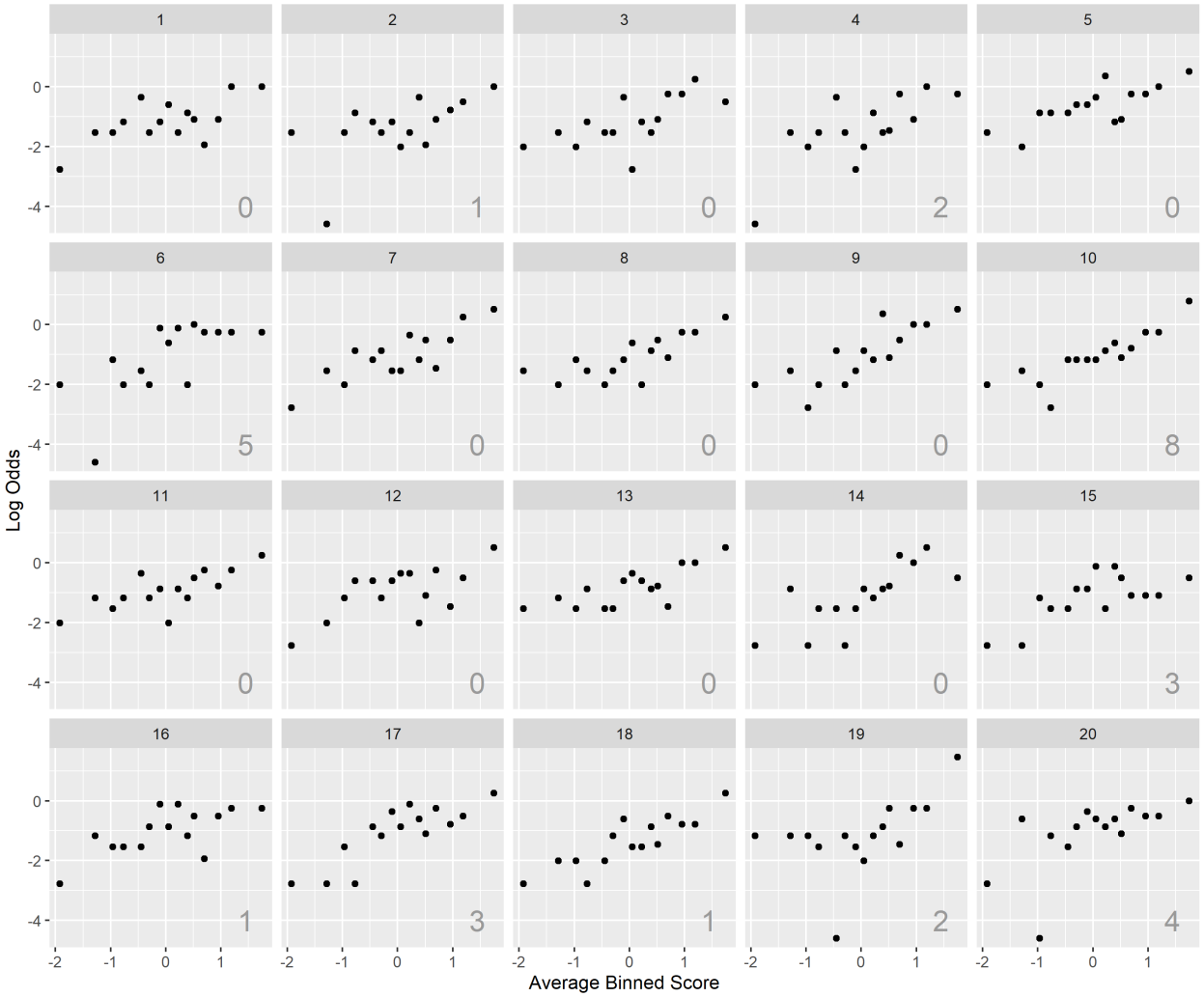


Figure A6: A lineup using empirical logit plots created from data with severe non-linearity and a sample size of 250. The true data is displayed in position $(\sqrt{49} + 12)$ which was picked by only 2 of 30 participants, giving a p-value of 0.258. The lineup uses the same true and null data as Figure A5, but does not reject the null hypothesis. This indicates that Figure A5 displays the data more effectively.