A Plot is Worth a Thousand Tests: Assessing Residual Diagnostics with the Lineup Protocol

Weihao Li* Monash University
Dianne Cook† Monash University
Emi Tanaka‡ Australian National University
Susan VanderPlas§ University of Nebraska–Lincoln

ABSTRACT

Regression experts consistently recommend plotting residuals for model diagnosis, despite the availability of many numerical hypothesis test procedures designed to use residuals to assess problems with a model fit. Here we provide evidence for why this is good advice using data from a visual inference experiment. We show how conventional tests are too sensitive, which means that too often the conclusion would be that the model fit is inadequate. The experiment uses the lineup protocol which puts a residual plot in the context of null plots. This helps generate reliable and consistent reading of residual plots for better model diagnosis. The lineup protocol also detects a range of departures from good residuals simultaneously.

Index Terms: Mathematics of computing—Probability and statistics—Statistical paradigms—Regression analysis; Mathematics of computing—Probability and statistics—Statistical paradigms—Statistical graphics

1 INTRODUCTION

In linear regression analysis, studying the residuals from a model fit is a common diagnostic activity. Residuals summarise what is not captured by the model, and thus provide the capacity to identify what might be wrong. Linear regression is a well-established procedure, and there is considerable literature describing diagnostic procedures, e.g. [4], [6] and [3]. Interestingly, despite the abundance of residual-based conventional tests, ALL of these writings advise that plotting residuals is an essential tool for diagnosing regression model problems:

> *Some most useful checks allowed by data plots should be done on a routine basis for every regression.* [4]
>
> *Such formal and informal procedures are complementary, and both have a place in residual analysis.* [3]
>
> *In our experience, statistical tests on regression model residuals are not widely used. In most practical situations the residual plots are more informative than the corresponding tests. However, since residual plots do require skill and experience to interpret, the statistical tests may occasionally prove useful.* [6]

The common wisdom of experts is that plotting the residuals is indispensable for diagnosing model fits. The ubiquity of this advice is *curious*.

While examining the scatterplots of residuals, if there are any visually discoverable patterns, the model is potentially inadequate or incorrectly specified. In general, one looks for noticeable departures from the model such as non-linear pattern or heteroskedasticity. However, correctly judging whether NO pattern exists in a residual plot is a difficult task for humans. Humans will almost always see a pattern (see [5]), so the question that really needs answering is whether any pattern perceived is consistent with randomness, purely sampling variability, or noise.

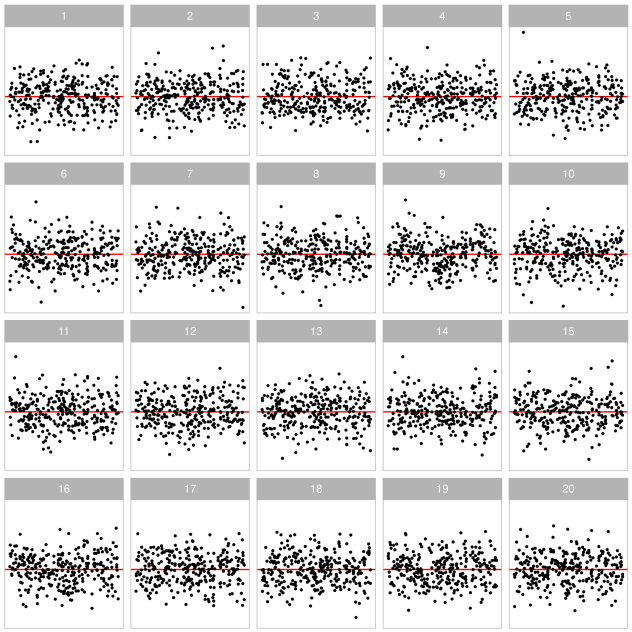

Figure 1: Visual testing is conducted using a lineup, as in the example here. The residual plot computed from the observed data (plot 2^3+1) is embedded among 19 null plots, where the residuals are simulated from a standard error model.

To address the issue of over-interpretation, we can examine the plot in the context of natural sampling variability assumed by the model, called the lineup protocol, as proposed in [2]. The protocol consists of m randomly placed plots, where one plot is the data plot, and the remaining $m-1$ plots, referred to as the *null plots*, are constructed using the same graphical procedure as the data plot but the data is replaced with null data that is generated in a manner consistent with the null hypothesis, H_0. Computing the p-value requires that the lineup be examined by a number of human judges, each asked to select the most different plot. A small p-value would result from a substantial number selecting the data plot. Under H_0, it is expected that the data plot would have no distinguishable difference from the null plots, and the human judge would only select the data plot by chance. Fig. 1 is an example of a lineup protocol. If the data plot at position 2^3+1 is identifiable, then it is evidence for the rejection of H_0. Otherwise, we fail to reject H_0.

In this study, we assessed the advice of plotting residuals in regression diagnostics by comparing conventional residual-based test-

*e-mail: weihao.li@monash.edu
†e-mail: dicook@monash.edu
‡e-mail: emi.tanaka@anu.edu.au
§e-mail: susan.vanderplas@unl.edu

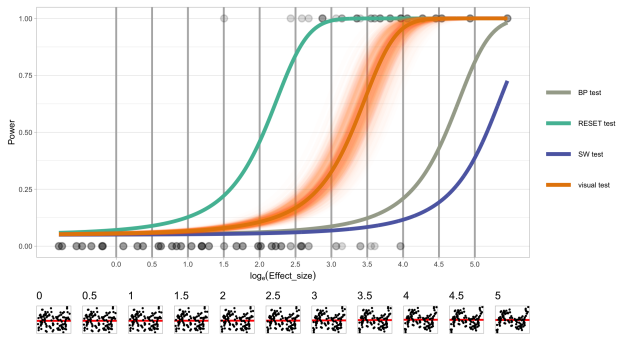

Figure 2: Comparison of power between different tests for non-linear patterns. The power curves are estimated using logistic regression, and the horizontal lines of dots represent non-reject and reject results from visual tests for each lineup. The visual test has multiple power curves estimated from bootstrap samples. The row of scatterplots at the bottom are examples of residual plots corresponding to the specific effect sizes marked by vertical lines in the main plot.

ing and visual testing in a perceptual experiment.

2 Experimental design

Our experiment was conducted over three data collection periods. Data collection period I was designed to study the ability of participants to detect the effect of a non-linear term z constructed using Hermite polynomials formulate as $y = 1 + x + z + \varepsilon$, $z \propto He_j(x)$ and $\varepsilon \sim N(0, \sigma^2 I)$. The null regression model used to fit the realizations generated is formulated as $y = \beta_0 + \beta_1 x + u$. Since $z = O(x^j)$, for $j > 1$, z is a higher order term left out of the null regression, which will result in model misspecification.

Data collection period II was designed to study the ability of participants to detect the heteroskedasticity under a simple linear regression model setting formulated as $y = 1 + x + \varepsilon$, $\varepsilon \sim N(0, 1 + (2 - |a|)(x - a)^2 bI)$. The same null regression model was used to fit the realizations. For $b \neq 0$, the variance-covariance matrix of the error term ε is correlated with the predictor x, which will lead to the presence of heteroskedasticity.

In visual p-value caculation [9], the parameter α of the dirichlet distribution usually needs to be estimated using data collected from the experiment. Thus, we recorded human responses to null lineups in data collection period III.

Overall, we collected 7974 evaluations on 1152 unique lineups performed by 443 participants throughout three data collection periods.

3 Results

Fig. 2 compares the power for the different tests for non-linear structure in the residuals, where the effect size is quantified via an approach based on Kullback-Leibler divergence, coupled with simulation. The test with the uniformly higher power is the Ramsey Regression Equation Specification Error Test (RESET) [7], one that specifically tests for non-linearity. Note that the Breusch-Pagan (BP) test [1] and Shapiro-Wilk (SW) normality test [8] have much lower power, which is expected because they are not designed to detect non-linearity. The bootstrapped power curves for the visual test are effectively a shift right from that of the RESET test. This means that the RESET test will reject at a lower effect size (less structure) than the visual test, but otherwise the performance will be similar. In other words, the RESET test is more sensitive than the visual test. This is not necessarily a good feature for the purposes of diagnosing model defects: If we scan the residual plot examples at the bottom, we might argue that the non-linearity is not sufficiently problematic until an effect size of around 3 or 3.5. The RESET test would reject closer to an effect size of 2, but the visual test would reject closer to 3.25, for a significance level of 0.05. The visual test matches the robustness of the model to (minor) violations of assumptions much better.

Fig. 1 is an example where the RESET test would reject H_0 (p-value $= 2 \times 10^{-9}$), but a visual test would not (visual p-value $=$ 0.530), when the effect size (log_e(Effect) $= 2.68$) is particularly small. Only one out of eleven participants identified the data plot for this lineup. It can be also seen from the lineup that there is no obvious non-linear patterns exhibited in the data plot. In this case, we can argue that the fitted model is a good enough approximation to the underlying data genearting process.

Similar findings can be extracted for the heteroskedasticity pattern so we will not repeat the discussion here.

4 Conclusion

Regression analysis experts suggest that residual plots are indispensable methods for assessing model fit. We conducted a perceptual experiment using visual inference to assess this oft-repeated advice. The experiment tested two primary departures from good residuals: non-linearity and heteroskedasticity.

We found that conventional residual-based statistical tests are more sensitive to weak departures than visual tests. Conventional tests often reject when departures in the form of non-linearity and heteroskedasticity are not visibly different from null residual plots.

While it might be argued that the conventional tests are correctly detecting small but real effects, this can also be seen as the conventional tests are rejecting unnecessarily. Many of these rejections happen even when downstream analysis and results would not be significantly affected by the small departures from a good fit. The results from human evaluations provide a more practical solution, which reinforces the statements from regression experts that residual plots are an indispensable method for model diagnostics.

It is important to note that residual plots need to be delivered as a lineup, embedded in a field of null plots. A residual plot may contain many visual features, but some are caused by the characteristics of the predictors and the randomness of the error, not by the violation of the model assumptions. The lineup enables a careful calibration for reading structure in residual plots.

References

[1] T. S. Breusch and A. R. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, pp. 1287–1294, 1979.

[2] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. 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.

[3] R. D. Cook and S. Weisberg. *Residuals and influence in regression*. New York: Chapman and Hall, 1982.

[4] N. R. Draper and H. Smith. *Applied regression analysis*, vol. 326. John Wiley & Sons, 1998.

[5] D. Kahneman. *Thinking, fast and slow*. macmillan, 2011.

[6] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 1982.

[7] J. B. Ramsey. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(2):350–371, 1969.

[8] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[9] S. VanderPlas, C. Röttger, D. Cook, and H. Hofmann. Statistical significance calculations for scenarios in visual inference. *Stat*, 10(1):e337, 2021.