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

Weihao Li[a], Dianne Cook[a], Emi Tanaka[b], Susan VanderPlas[c]

[a]Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia; [b]Biological Data Science Institute, Australian National University, Acton, ACT, Australia; [c]Department of Statistics, University of Nebraska, Nebraska, USA

Lineup Protocol

The lineup protocol (Buja et al., 2009) is a visual statistical test which consists:

- one randomly placed data plot;
- m - 1 null plots plotted with data generated from the null model.

Observers will be asked to select the most different plot from the lineup. Under the null hypothesis, it is expected that the data plot would have no distinguishable difference from the null plots, and successfully identify the data plot provides evidence against the null hypothesis.

Why using Lineup Protocol for Residual Diagnostics?

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. These irrelevant visual features have a chance to be filtered out by participants with a comparison to null plots, resulting in more accurate reading. The lineup enables a careful calibration for reading structure in residual plots.

Experimental Design

We designed a human subject experiment to assess residual plots with lineup protocol. The experiment was conducted over three data collection periods.

Data Collection Period I: Non-linearity

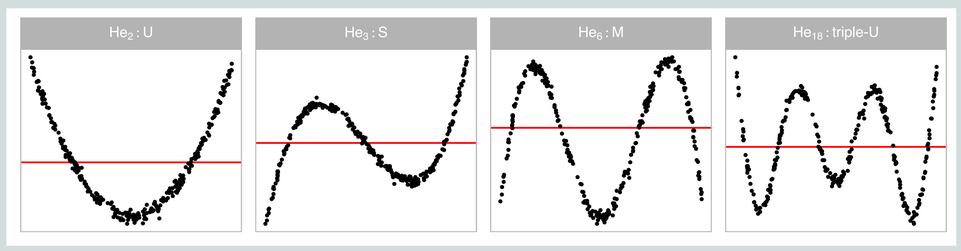


Figure 1: Polynomial forms generated for the residual plots used to assess detecting non-linearity

Data generating process: $\boldsymbol{y} = 1 + \boldsymbol{x} + \boldsymbol{z} + \boldsymbol{\varepsilon},\ \boldsymbol{z} \propto He_j(\boldsymbol{x})$ and $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.
Null regression model: $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{u}$.

Data Collection Period II: Heteroskedasticity

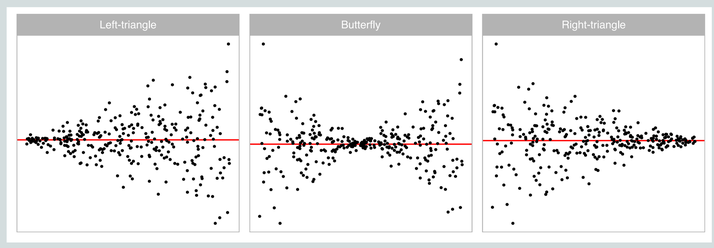


Figure 2: Heteroskedasticity forms used in the experiment. Three different shapes (a = −1, 0, 1) are used in the experiment to create left-triangle, ”butterfly” and ”right-triangle” shapes, respectively.

Data generating process: $\boldsymbol{y} = 1 + \boldsymbol{x} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, 1 + (2 - |a|)(\boldsymbol{x} - a)^2 b \boldsymbol{I})$.
Null regression model: $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{u}$.

Data Collection Period III: Null lineups

In visual p-value caculation (Vanderplas et al., 2021), 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.

Results

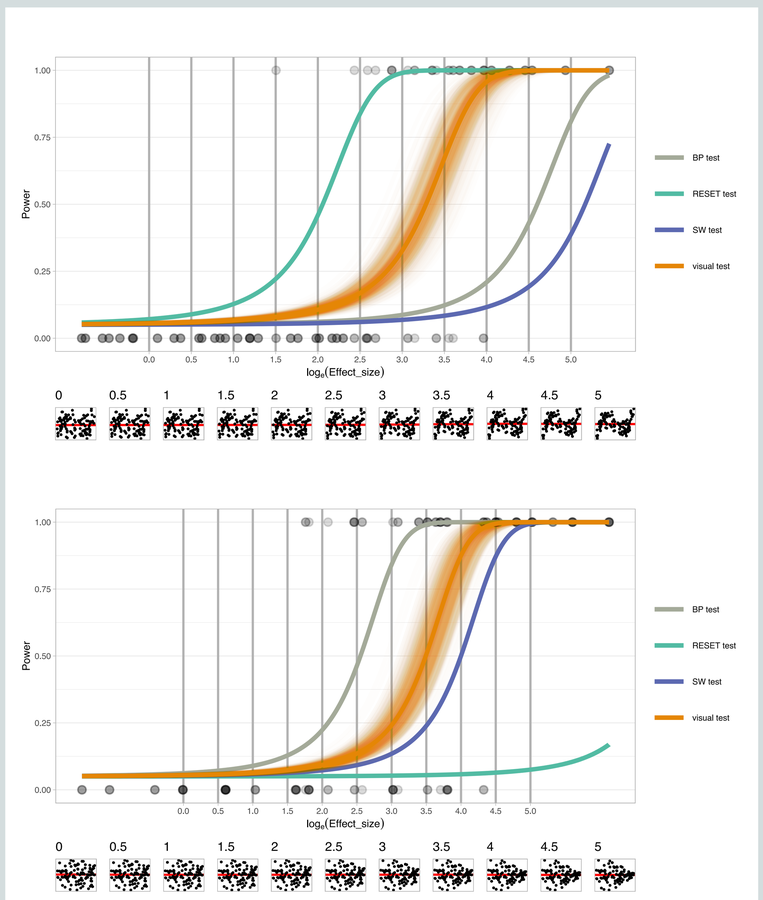


Figure 3: Comparison of power between different tests for non-linear and heteroskedasticity patterns. The row of scatterplots at the bottom are examples of residual plots at different effect sizes.

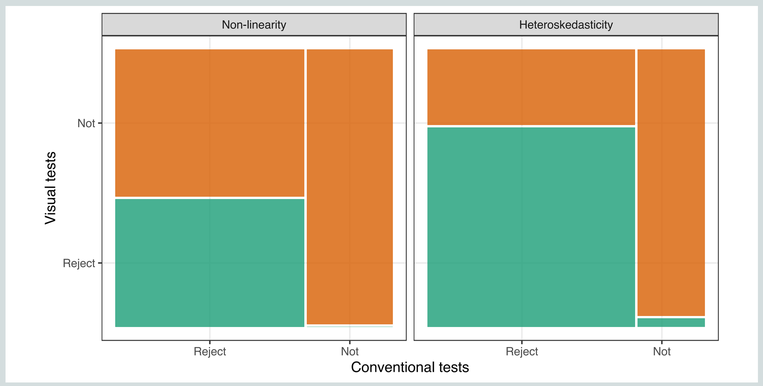


Figure 4: Rejection rate (p-value ≤ 0.05) of visual test conditional on the conventional test decision on non-linearity (left) and heteroskedasticity (right) lineups displayed using a mosaic plot.

Conventional tests are MORE sensitive than visual tests:

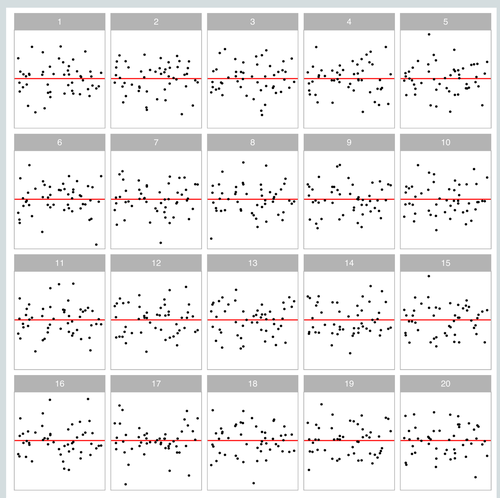
- Ramsey Regression Equation Specification Error Test (RESET) is more sensitive in detecting non-linearity.
- Breusch-Pagan (BP) test is more sensitive in detecting heteroskedasticity.

The visual test matches the robustness of the model to (minor) violations of assumptions much better:

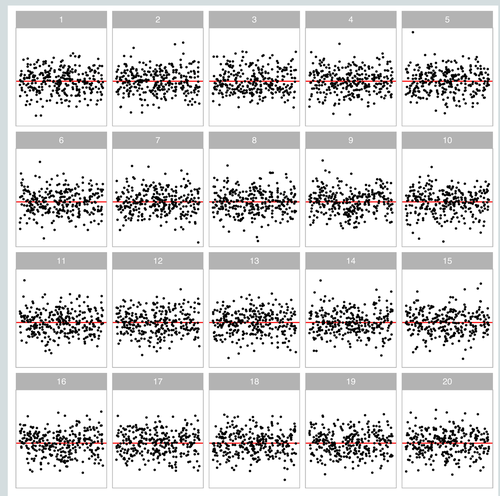
If we scan the residual plot examples at the bottom, we might argue that the non-linearity is not sufficiently problematic until an effect size of around 3 or 3.5. However, RESET test would reject closer to an effect size of 2.

A conventional test concludes there are problems with the model fit almost twice as often as a human.

Please select the most different plot.



This heteroskedasticity lineup is rejected by the visual test but not by the BP test. The data plot (position 17) contains a “butterfly” shape. It visibly displays heteroskedasticity, making it somewhat surprising that it is not detected by the BP test.



This is an example where the RESET test would reject H0 (p-value = 2×10^{-9}), but a visual test would not (visual p-value = 0.530), when the effect size is particularly small.

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 generating process.

Conclusion

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. The lineup protocol also detects a range of departures from good residuals simultaneously.

References

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.

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

Contact: patrick.li@monash.edu
Weihao (Patrick) Li