ARTICLE TEMPLATE

Why is it that statistical tests for residuals are not widely used? An explanation using visual inference.

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

[a]Department of Econometrics and Business Statistics, Monash University, Clayton, VIC, Australia; [b]Department of Statistics, University of Nebraska, Lincoln, Nebraska, USA

ARTICLE HISTORY
Compiled February 1, 2023

ABSTRACT
Abstract to fill.

KEYWORDS
data visualization; visual inference; hypothesis testing; residual plots;

Title (jokes: finding the pattern in residual tests) (why using visulization is important in regression diagnostics? an explanation using visual inference) (a picture is worth a thousand word) (a plot is worth a thousand test:)

1. Introduction

> *"Since all models are wrong the scientist must be alert to what is importantly wrong."* (Box 1976)

Diagnosing a model is the key to determining whether there is anything importantly wrong. In linear regression analysis, residuals are typically examined for model diagnostics. Residuals summarise what is not captured by the model, and thus provide the capacity to identify what might be wrong.

We can assess residuals in multiple ways. Residuals might be plotted, as a histogram or quantile-quantile plot to examine the distribution. Using the classical normal linear regression model as an example, if the distribution is symmetric and unimodal, we consider it well-behaved. But if the distribution is skewed, bimodal, multimodal, or contains outliers, there is cause for concern. One could also inspect the distribution by conducting a goodness of fit test, such as the Shapiro-Wilk Normality test (Shapiro and Wilk 1965).

More typically, residuals will be plotted, as a scatter plot against the predicted values and each of the explanatory variables to scrutinize their relationships. If there are any visually discoverable patterns, the model is potentially misspecified. In general, one looks for noticeable departures from the model like non-linear dependency or heteroskedasticity. However, correctly judging a residual plot where no pattern exists is

CONTACT Weihao Li. Email: weihao.li@monash.edu, Dianne Cook. Email: dicook@monash.edu, Emi Tanaka. Email: emi.tanaka@monash.edu, Susan VanderPlas. Email: susan.vanderplas@unl.edu

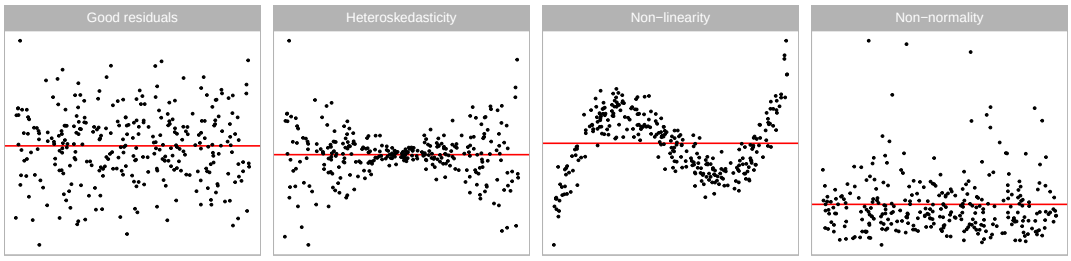

Figure 1. Example residual vs fitted plots: (A) classically good looking residuals, (B) non-linear pattern indicates that the model has not captured a non-linear association, (C) heteroskedasticity indicating that variance around the fitted model is not uniform, and (D) non-normality where the residual distribution is not symmetric around 0. The latter pattern might best be assessed using a univariate plot of the residuals, but patterns B and C need to be assessed using a residual vs fitted plot.

a painstakingly difficult task for humans (?citation). It is especially common, particularly among new data analysts, to report patterns when an experienced one might quickly conclude that there are none. It is also possible to conduct hypothesis tests for non-linear dependence (Ramsey 1969), and use a Breusch-Pagan test (Breusch and Pagan 1979) for heteroskedasticity.

Abundance of literature describe appropriate diagnostic methods for linear regression: Draper and Smith (1998), Montgomery and Peck (1982), Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1999) and Cook and Weisberg (1982). The diligent reader of these sage writings will also notice sentences that express sentiments like *based on their experience, statistical tests are not widely used in regression diagnostics since the same or even larger amount of information can be provided by diagnostic plots than the corresponding tests in most empirical studies.* A common guidance by experts is that optimal method for diagnosing model fits is by plotting the data.

The persistence of this advice to check the plots is curious, and investigating why this might be common advice is the subject of this paper. The paper is structured as follows. The next background section describes the types of departures that one expects to detect, and outlines a formal statistical process for reading residual plots, called visual inference. Section 3 details the experimental design to compare the decision made by formal hypothesis testing, and how humans would read diagnostic plots. The results are reported in Section 4. We finish with a discussion on future work, in particular how the responsibility for residual plot reading might be relieved.

2. Background

2.1. *Departures from good residual plots*

(change it to facet)

(remove backticks for some plots)

Graphical summaries in which residuals are plotted against fitted values or other functions of the regressors that are approximately orthogonal to residuals are referred to as standard residual plots in Cook and Weisberg (1982). The first panel of Figure 1 shows an ideal residual plot with residuals evenly distributed at both sides of the horizontal zero line, with no noticeable patterns.

There are various types of departures from an ideal residual plot. Non-linearity, heterskedasticity and non-normality are perhaps the three mostly checked departures.

Non-linearity is a type of model misspecification caused by failing to include higher order terms of the regressors in the regression equation. Any non-linear functional form of residuals on fitted values in the residual plot could be indicative of non-linearity. An example residual plot containing visual pattern of non-linearity is given at the second panel of Figure 1. One can clearly observe the "S-shape" from the residual plot as the cubic term is not captured by the misspecified model.

Heteroskedasticity refers to the presence of nonconstant error variance in a regression model. It is mostly due to the strict but false assumptions on the variance-covariance matrix of the error term. The usual pattern of heteroskedasticity on a residual plot is the inconsistent spread of the residuals across the horizontal axis. Visually, it sometimes results in the so-called "butterfly" shape as shown in the third panel of Figure 1, or the "left-triangle" and "right-triangle" shape where the smallest variance occurs at one side of the horizontal axis.

Compared to non-linearity and heteroskedasticity, non-normality is usually harder to detect from a residual plot since a scatter plot do not readily reveal the marginal distribution. A favourable graphical summary for this task is the quantile-quantile plot. As we mainly discuss residual plots, non-normality will not be the focus of this paper. For a consistent comparison, the residual plot of this departure is still presented in the fourth panel of Figure 1. When the number residuals below and above the horizontal axis are uneven across the local regions along the x-axis, we expect that the normality assumption is violated. For example, given a skewed error distribution, there will be fewer data points and more outliers on one side of the horizontal axis as shown in the fourth panel of Figure 1.

2.2. *Conventionally testing for departures*

Other than checking diagnostic plots, analysts may perform formal hypothesis testing for detecting model defects. Depending on the alternative hypothesis that is focused on, a variety of tests can be applied. For example, the presence of heteroskedasticity can usually be tested by applying the White test (White 1980) or the Breusch-Pagan test (Breusch and Pagan 1979), which are both derived from the Lagrange multiplier test (Silvey 1959) principle that relies on the asymptotic properties of the null distribution. For testing non-linearity, one may apply the F-test as a model structural test to examine the significance of specific polynomial and non-linear forms of the regressors, or the significance of proxy variables as in the Ramsey Regression Equation Specification Error Test (RESET) (Ramsey 1969). The Shapiro-Wilk test (Shapiro and Wilk 1965) is the most widely used test of non-normality included by many of the statistical softwares. The Jarque–Bera test (Jarque and Bera 1980) is also used to directly checks if the sample skewness and kurtosis match a normal distribution.

Example residual plots given in Figure 1 are examined by the corresponding RESET test, Breusch-Pagan test and Shapiro-Wilk test as shown in Table 1. In the example, the Breusch-Pagan test and the Shapiro–Wilk test both reject the null hypothesis H_0 for departures that they do not intend to examine. As discussed in Cook and Weisberg (1982), most residual-based tests for a particular type of departure from model assumptions are also sensitive to other types of departures. It is likely H_0 is correctly rejected but for the wrong reason, a phenomenon known as the "Type III error". Additionally, outliers will often incorrectly trigger the rejection of H_0 despite when majority of the residuals are well-behaved (Cook and Weisberg 1999). Furthermore, with a sufficiently large sample size, residual-based tests may reject H_0 due to a slight

Table 1. Statistical significance testing for departures from good residuals for plots in Figure 1. Shown are the p-values calculated for the RESET, the Breusch-Pagan and the Shapiro–Wilk tests. The good residual plot (A) is judged a good residual plot, as expected, by all tests. The non-linearity (B) is detected by all tests, as might be expected given the extreme structure.

Plot	Departures	RESET	Breusch-Pagan	Shapiro–Wilk
A	None	0.779	0.133	0.728
B	Non-linearity	*0.000*	*0.000*	*0.039*
C	Heteroskedasticity	0.658	*0.000*	*0.000*
D	Non-normality	0.863	0.736	*0.000*

departure that is of little practical significance. These can be largely avoided in diagnostic plots as experienced analysts can evaluate the acceptability of assumptions flexibly, even in the presence of outliers and slight departures.

2.3. *Visual test procedure based on lineups*

2.3.1. Lineup protocol

One may argue that reading diagnostic plots is to some extent subjective and indecisive compared to those rigorous statistical procedures as it relies on graphical perception - human ability to interpret and decode the information embedded in graph (Cleveland and McGill 1984). Further, the degree of the presence of the visual features typically can not be measured quantitatively and objectively, which may lead to over or under-interpretations of the data. For instance, people over-interpret the separation between gene groups in a two-dimensional projection from a linear discriminant analysis when in fact there are no differences in the expression levels between the gene groups and separation is not an uncommon occurrence (Roy Chowdhury et al. 2015).

Visual inference was first introduced in a 1999 Joint Statistical Meetings (JSM) talk with the title "Inference for Data Visualization" by Buja, Cook, and Swayne (1999) as an idea to address the issue of valid inference for visual discoveries of data plots. Later, Buja et al. (2009) proposed the lineup protocol as a visual test inspired by the "police lineup" or "identity parade" which is the act of asking the eyewitness to identify criminal suspect from a group of irrelevant people. The protocol consists of m randomly placed plots, where one plot is the data plot, and the remaining $m-1$ null plots have the identical graphical procedure except the data has been replaced with data consistent with H_0. Then, an observer who have not seen the data plot will be asked to point out the most different plot from the lineup. Under H_0, it is expected that the data plot would have no distinguishable difference from the null plots, and the probability that the observer correctly picks the data plot is $1/m$. If one rejects H_0 as the observer correctly picks the data plot, then the Type I error of this test is $1/m$.

Figure 2 is an example of a lineup protocol. If the data plot at position $2^2 + 2$ is identifiable, then it is evidence for the rejection of H_0 that the regression model is correctly specified. In fact, the actual residual plot is obtained from a misspecified regression model with non-linearity defect.

The effectiveness of lineup protocol for regression analysis is validated by Majumder, Hofmann, and Cook (2013) under relatively simple settings with up to two regressors.

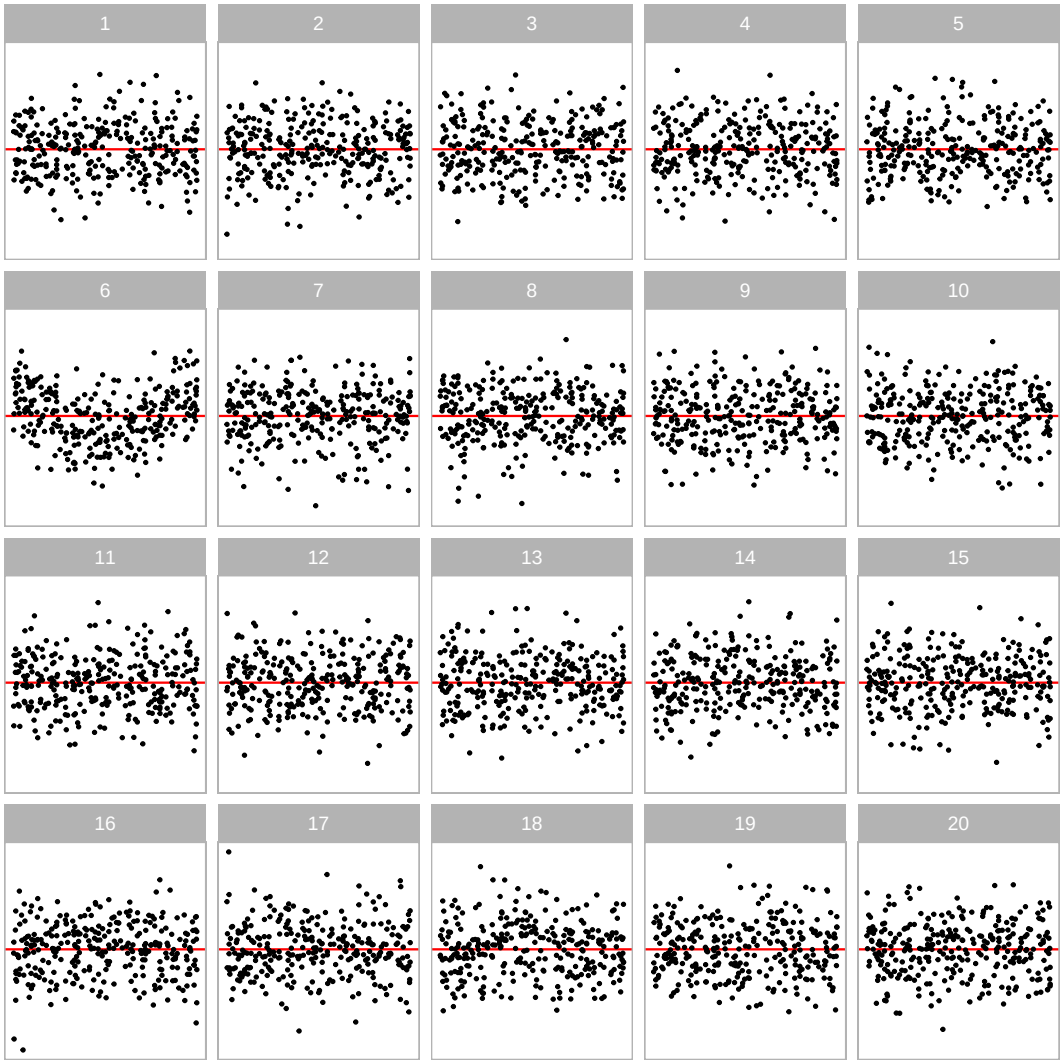

Figure 2. Visual testing is conducted using a lineup, as in the example here. The residual plot computed from the observed data (plot $2^2 + 2$, exhibiting non-linearity) is embedded among 19 null plots, where the residuals are simulated from a standard error model. Computing the p-value requires that 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 plot $2^2 + 2$.

Their results suggest that visual tests are capable of testing the significance of a single regressor with a similar power as a t-test, though they express that in general it is unnecessary to use visual inference if there exists a conventional test, and they do not expect the visual test to perform equally well as the conventional test. (contradict to our claim? expand?) In their third experiment, where there is not a conventional test, visual test outperforms the conventional test for a large margin. This is encouraging, as it promotes the use of visual inference in situations where there are no existing statistical testing procedures. Visual inference have also been integrated into diagnostic of hierarchical linear models by Loy and Hofmann (2013), Loy and Hofmann (2014) and Loy and Hofmann (2015). They use lineup protocols to judge the assumption of linearity, normality and constant error variance for both the level-1 and level-2 residuals. (expand?)

2.3.2. Sampling from the null distribution

Data used in the $m-1$ null plots needs to be simulated. In regression diagnostics, sampling data consistent with H_0 is equivalent to sampling data from the assumed model. As Buja et al. (2009) suggested, H_0 is usually a composite hypothesis controlled by nuisance parameters. Since regression models can have various forms, there is no general solution to this problem, but it sometimes can be reduced to a so called "reference distribution" by applying one of the three methods: (i) sampling from a conditional distribution given a minimal sufficient statistic under H_0, (ii) parametric bootstrap sampling with nuisance parameters estimated under H_0, and (iii) Bayesian posterior predictive sampling. The conditional distribution given a minimal sufficient statistic is the best justified reference distribution among the three (Buja et al. 2009). Essentially, null residuals can be simulated by regressing N i.i.d standard normal random draws on the regressors, then rescaling it by the ratio of residual sum of square in two regressions.

2.3.3. Calculating p-values for the visual test

In hypothesis testing, a p-value is defined as the probability of observing test results as least as extreme as the observed result given H_0 is true. Within the context of visual inference, by involving k independent observers, the p-value can be interpreted as the probability of having as many or more subjects detect the data plot than the observed result.

Let $X_j = \{0, 1\}$ be a Bernoulli random variable denoting whether subject j correctly detecting the data plot, and $X = \sum_{j=1}^{K} X_j$ be the number of observers correctly picking the data plot. Then, by imposing a relatively strong assumption on the visual test that all K evaluations are fully independent, under H_0, $X \sim \text{Binom}_{K,1/m}$. Therefore, the p-value of a lineup of size m evaluated by K observer is given as $P(X \geq x) = 1 - F(x) + f(x)$, where $F(.)$ is the cumulative distribution function, $f(.)$ is the probability mass function and x is the realization of number of observers correctly picking the data plot (Majumder, Hofmann, and Cook 2013).

As pointed out by VanderPlas et al. (2021), this basic binomial model doesn't take into account the possible dependencies in the visual test due to repeated evaluations of the same lineup. And it is inapplicable to visual test where subjects are asked to select one or more "most different" plots from the lineup. VanderPlas et al. (2021) summarises three common scenarios in visual inference: (1) K different lineups are shown to K subjects, (2) K lineups with different null plots but the same data plot

are shown to K subjects, and (3) the same lineup is shown to K subjects. Out of these three scenarios, Scenario 3 is the most common in previous studies as it puts the least constraints on the experiment design. For Scenario 3, VanderPlas et al. (2021) models the probability of a plot i being selected from a lineup as θ_i, where $\theta_i \sim Dirichlet(\alpha)$ for $i = 1, ..., m$ and $\alpha > 0$. The number of times plot i being selected in K evaluations is denoted as c_i. In case subject j makes multiple selections, $1/s_j$ will be added to c_i instead of one, where s_j is the number of plots subject j selected for $j = 1, ...K$. This ensures $\sum_i c_i = K$. Since we are only interested in the selections of the data plot i, the marginal model can be simplified to a beta-binomial model and thus the visual p-value is given as

$$P(C \geq c_i) = \sum_{x=c_i}^{K} \binom{K}{x} \frac{B(x+\alpha, K-x+(m-1)\alpha)}{B(\alpha,(m-1)\alpha)}, \quad \text{for} \quad c_i \in \mathbb{Z}_0^+ \tag{1}$$

where $B(.)$ is the beta function defined as

$$B(a,b) = \int_0^1 t^{\alpha-1}(1-t)^{b-1}dt, \quad \text{where} \quad a, b > 0. \tag{2}$$

Note that Equation 1 given in VanderPlas et al. (2021) only works with non-negative integer c_i. We extend the equation to non-negative real number c_i by applying a linear approximation

$$P(C \geq c_i) = P(C \geq \lceil c_i \rceil) + (\lceil c_i \rceil - c_i)P(C = \lfloor c_i \rfloor), \quad \text{for} \quad c_i \in \mathbb{R}_0^+, \tag{3}$$

where $P(C \geq \lceil c_i \rceil)$ is calculated using Equation 1 and $P(C = \lfloor c_i \rfloor)$ is calculated by

$$P(C = c_i) = \binom{K}{c_i} \frac{B(c_i+\alpha, K-c_i+(m-1)\alpha)}{B(\alpha,(m-1)\alpha)}, \quad \text{for} \quad c_i \in \mathbb{Z}_0^+. \tag{4}$$

Besides, the parameter α used in Equation 1 and 4 is usually unknown and hence needs to be estimated from the survey data. For low values of α, only a few plots are attractive to the observers and tend to be selected. For higher values of α, the distribution of the probability of each plot being selected is more even. VanderPlas et al. (2021) defines that a plot is c-interesting if c or more participants select the plot as the most different. Given the definition, The expected number of plots selected at least c times, $E[Z_c]$, is calculated as

$$E[Z_c(\alpha)] = \frac{m}{B(\alpha,(m-1)\alpha)} \sum_{\lceil c \rceil}^{K} \binom{K}{x} B(x+\alpha, K-x+(m-1)\alpha). \tag{5}$$

With Equation 5, α can be estimated using maximum likelihood estimation. But for precise estimate of α, additional responses to Rorschach lineups, which is a type

of lineup that consists of plots constructed from the same null data generating mechanism, are required.

2.3.4. Power of a visual test

The power of a model misspecification test is the probability that the null hypothesis is rejected given the regression model is misspecified. It is an important indicator when one is concerned about whether model assumptions have been violated. Although in practice, one might be more interested in knowing how much the residuals deviate from the model assumptions, and whether this deviation is of practical significance.

As discussed in Majumder, Hofmann, and Cook (2013), the power of a visual test may depend on the ability of the particular subject, as the skill of the individual may affect the number of observers who identify the data plot from the lineup. Previously, it is addressed by modelling the probability of a subject j correctly picking the data plot from a lineup l using a mixed-effect logistic regression, with subjects treated as random effects (Majumder, Hofmann, and Cook 2013). However, in the multiple selections scenario, having this probability is not sufficient to determine the power of a visual test because it does not provide information about the number of selections made by the subject for the calculation of the p-value.

Instead, we directly estimate the probability of a lineup being rejected by assuming that individual skill has negligible effect on the variation of the power. The assumption is to simplify the model structure, thereby obviate costly large-scale experiments to estimate complex covariance matrices. The model is a logistic regression with the natural logarithm of the effect as the only regressor formulated as

$$Pr(\text{reject } H_0|H_1, E) = \Lambda(\beta_0 + \beta_1 log(E)), \tag{6}$$

where $\Lambda(.)$ is the standard logistic function given as $\Lambda(z) = exp(z)/(1 + exp(z))$.

Effect E is derived from the Kullback-Leibler divergence (see Appendix for the details) formulated as

$$E = \frac{1}{2}\left(log\frac{|diag(\boldsymbol{R}_a\boldsymbol{V})|}{|diag(\boldsymbol{R}_a\sigma^2)|} - n + tr(diag(\boldsymbol{R}_a\boldsymbol{V})^{-1}diag(\boldsymbol{R}_a\sigma^2)) + \boldsymbol{\mu}_z^T(\boldsymbol{R}_a\boldsymbol{V})^{-1}\boldsymbol{\mu}_z\right), \tag{7}$$

$$\boldsymbol{R}_a = \boldsymbol{I}_n - \boldsymbol{H}_a, \tag{8}$$

$$\boldsymbol{H}_a = \boldsymbol{X}_a(\boldsymbol{X}_a^T\boldsymbol{X}_a)^{-1}\boldsymbol{X}_a^T, \tag{9}$$

where $diag(.)$ is the diagonal matrix constructed from the diagonal elements of a matrix, $\boldsymbol{X}_a = (\boldsymbol{1}, \boldsymbol{X})$ is the matrix of regressors including the intercept used in the regression equation, $\sigma^2\boldsymbol{I}$ is the assumed variance of the error term when H_0 is true, $\boldsymbol{V}$ is the actual variance of the error term, and $\boldsymbol{\mu}_z = \boldsymbol{R}_a\boldsymbol{Z}\boldsymbol{\beta}_z$ is the expected values of the residuals with $\boldsymbol{Z}$ be any variables leave out by the model and $\boldsymbol{\beta}_z$ be the corresponding coefficients.

To study various factors contributing to the power of the visual test, the same logistic regression model is fit on different subsets of the collated data grouped by levels of factors. These include the distribution of the regressor and the type of the simulation model (modify this according to the result).

3. Experimental design

Three experiments are conducted to investigate the difference between conventional hypothesis testing and visual inference in the application of linear regression diagnostics. Two types of departures, namely non-linearity and heteroskedasticity, are considered with the corresponding data generating process being designed for experiment I and II. The experiment III is designed for collecting human responses to null lineups such that the parameter α in Equation 1 can be estimated. Overall, we plan to collect 7974 evaluations on 1152 unique lineups performed by 443 subjects throughout three experiments.

3.1. *Simulating departures from good residuals*

3.1.1. Non-linearity

Experiment I is designed to study the ability of human subjects to detect the effect of a non-linear term $\boldsymbol{z}$ constructed using Hermite polynomials on random vector $\boldsymbol{x}$ formulated as

$$\boldsymbol{y} = 1 + \boldsymbol{x} + \boldsymbol{z} + \boldsymbol{\varepsilon}, \tag{10}$$
$$\boldsymbol{x} = g(\boldsymbol{x}_{raw}, 1), \tag{11}$$
$$\boldsymbol{z} = g(\boldsymbol{z}_{raw}, 1), \tag{12}$$
$$\boldsymbol{z}_{raw} = He_j(g(\boldsymbol{x}, 2)), \tag{13}$$

where $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{x}_{raw}$, $\boldsymbol{z}_{raw}$ are vectors of size n, $He_j(.)$ is the jth-order probabilist's Hermite polynomials, $\varepsilon \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, and $g(\boldsymbol{x}, k)$ is a scaling function to enforce the support of the random vector to be $[-k, k]^n$ defined as

$$g(\boldsymbol{x}, k) = (\boldsymbol{x} - min(\boldsymbol{x}))/max(\boldsymbol{x} - min(\boldsymbol{x}))2k - k, \quad \text{for} \quad k > 0. \tag{14}$$

According to Abramowitz and Stegun (1964), Hermite polynomials were initially defined by Laplace (1820), but named after Hermite (Hermite 1864) because of the unrecognisable form of Laplace's work. When simulating $\boldsymbol{z}_{raw}$, function `hermite` from the R package `mpoly` (Kahle 2013) is used to generate Hermite polynomials.

The null regression model used to fit the realizations generated by the above model is formulated as

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{u}, \tag{15}$$

where $\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$.

Since $z = O(x^j)$, for $j > 1$, z is a higher order term leaves out by the null regression, which will lead to model misspecification.

Visual patterns of non-linearity are simulated using four different orders of probabilist's Hermite polynomials ($j = 2, 3, 6, 18$). (A summary of the factors is given in Table 2.) The values of j is chosen so that distinct shapes of non-linearity are included in the residual plot. These include "U", "S", "M" and "Triple-U" shape as shown in

Table 2. Levels of the factors used in experiments I, II, III.

Non-linearity		Heteroskedasticity		Common	
Poly Order (j)	SD (σ)	Shape (a)	Ratio (b)	Size (n)	Distribution of fitted values
2	0.25	-1	0.25	50	Uniform
3	1.00	0	1.00	100	Normal
6	2.00	1	4.00	300	Skewed
18	4.00		16.00		Discrete uniform
			64.00		

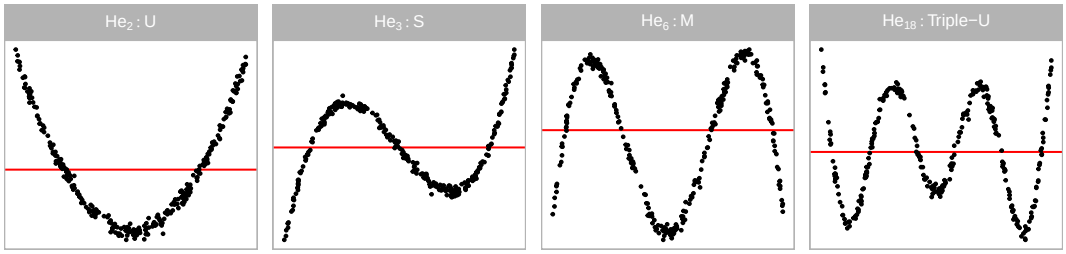

Figure 3. Polynomial forms generated for the residual plots used in experiment I. The four shapes are generated by varying the order of polynomial given by j in $He_j(.)$.

Figure 3. A greater value of j will result in a curve with more turning points. It is expected that the "U" shape will be the easiest one to detect because complex shape tends to be concealed by cluster of data points.

Figure 7 demonstrates one of the lineups used in experiment I. This lineup is produced by the non-linearity model with $j = 6$ and uniform fitted values. The data plot location is $2^3 - 4$. All five subjects correctly identify the data plot from this lineup.

3.1.2. Heteroskedasticity

Experiment II is designed to study the ability of human subjects to detect the appearance of a heteroskedasticity pattern under a simple linear regression model setting:

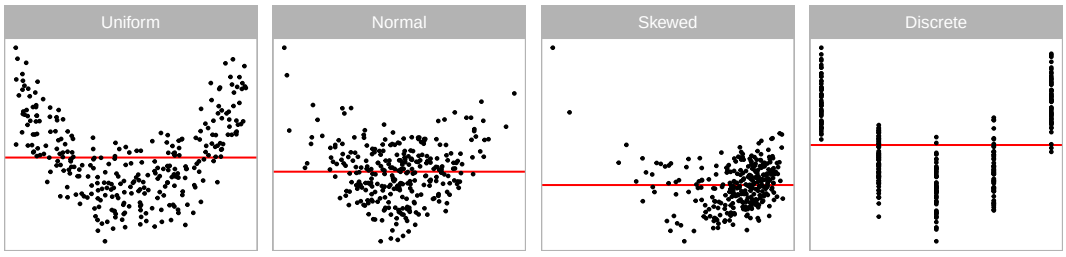

Figure 4. Variations in fitted values, that might affect perception of residual plots. Four different distributions are used.

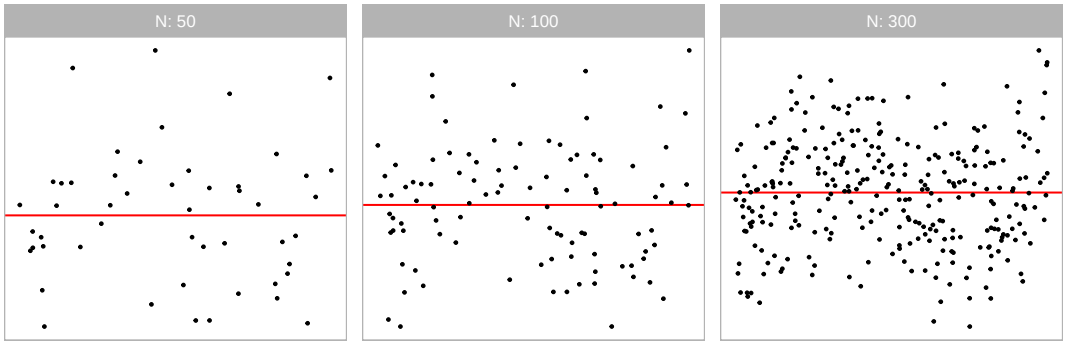

Figure 5. Number of observations n have an impact on the strength of the signal. Three different values of n are used in experiment I, II and III.

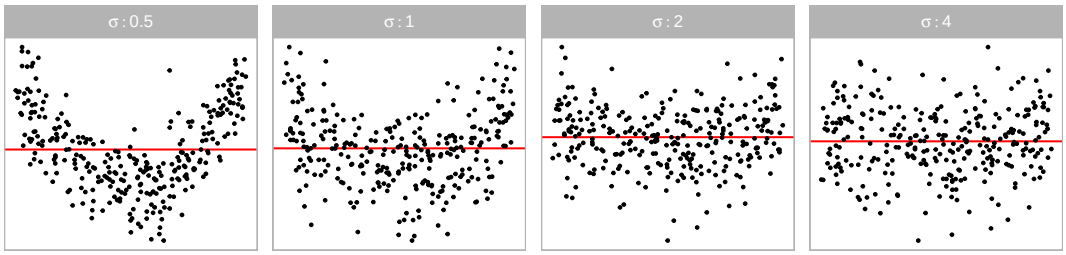

Figure 6. For the non-lienearity model, the standard deviation of the error term σ affects the strength of the signal. Four different values of σ are used in the experiment I.

$$\boldsymbol{y} = 1 + \boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{16}$$

$$\boldsymbol{x} = g(\boldsymbol{x}_{raw}, 1), \tag{17}$$

$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, 1 + (2 - |a|)(\boldsymbol{x} - a)^2 b\boldsymbol{I}), \tag{18}$$

where $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{\varepsilon}$ are vectors of size n and $g(.)$ is the scaling function defined in Equation 14.

The null regression model used to fit the realizations generated by the above model is formulated exactly the same as Equation 15.

For $b \neq 0$, the variance-covariance matrix of the error term $\boldsymbol{\varepsilon}$ is correlated with the regressor $\boldsymbol{x}$, which will lead to the presence of heteroskedasticity. Visual patterns of heteroskedasticity are simulated using three different shapes (a = -1, 0, 1). (A summary of the factors can be found in Table 2.)

Since $supp(X) = [-1, 1]$, choosing a to be -1, 0 and 1 can generate "left-triangle", "butterfly" and "right-triangle" shape as displayed in Figure 8. The term $(2 - |a|)$ maintains the magnitude of residuals across different values of a.

An example lineup of this model used in Experiment II is shown in Figure 10 with $a = -1$ and $X_{raw} \sim U(-1, 1)$. The data plot location is $2^4 + 2$. Nine out of 11 subjects correctly identify the data plot from this lineup.

3.1.3. Factors common to both experiments

XXX DC will clean this up

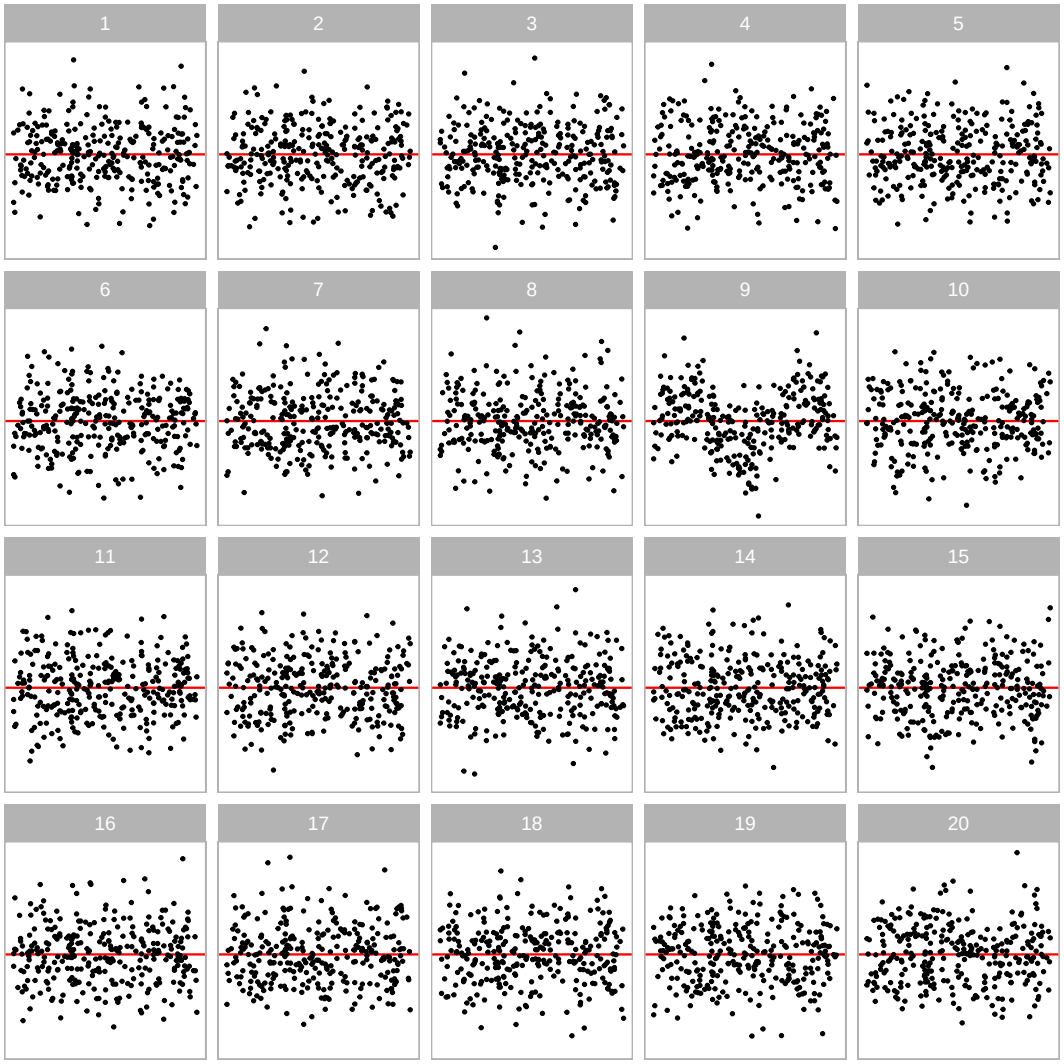

Figure 7. One of the lineups containing non-linearity patterns used in experiment I. Can you spot the most different plot? The data plot is positioned at $2^3 + 1$.

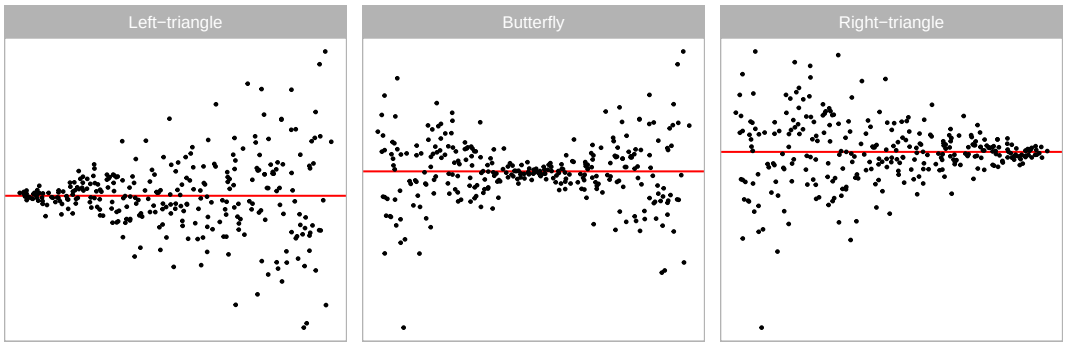

Figure 8. Heteroskedasticity forms used in experiment II. Three different shapes (a = -1, 0, 1) are used in the experiment to create left-triangle, butterfly and right-triangle shapes.

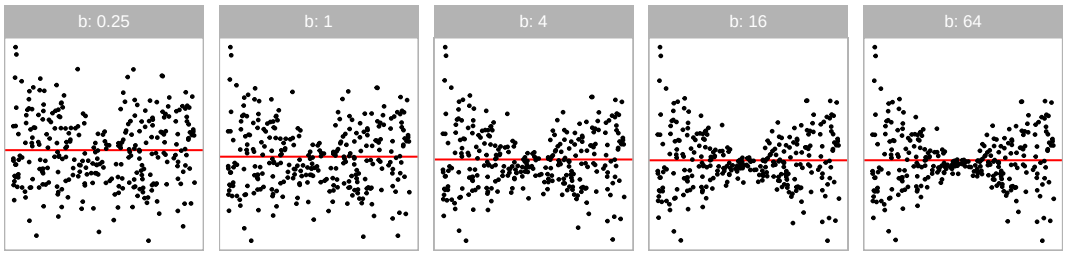

Figure 9. Five different values of b are used in experiment II to control the strength of the signal.

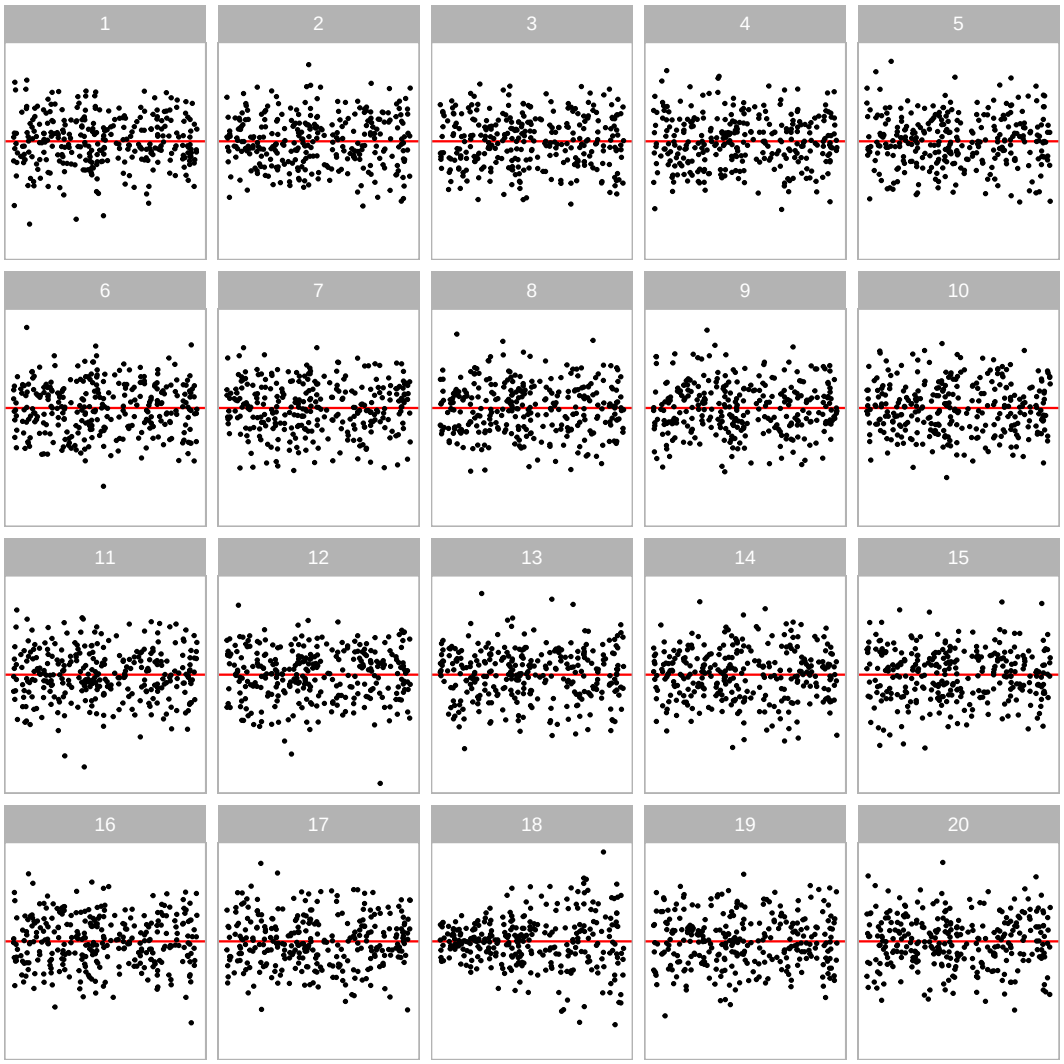

Figure 10. One of the lineups containing heteroskedasticity pattern used in experiment II. Can you spot the most different plot? The data plot is positioned at $3^3 - 3^2$

Four distributions of fitted values are also considered by using X_{raw} follows different distributions: (1) $U(-1, 1)$, (2) $N(0, 0.3^2)$, (3) $lognormal(0, 0.6^2)/3$ and (4) $u\{1, 5\}$.

Different distributions of fitted values help enriching the pool of visual patterns as illustrated in Figure 4. The uniform and the normal distribution are symmetric and commonly assumed in statistical models. The adjusted log-normal distribution provides skewed density, while the discrete uniform distribution provides discreteness in residual plot.

3.2. *Experimental setup*

3.2.1. Controlling the strength of the signal

As summarised in Table 2, three additional parameters n, σ and b are used to control the strength of the signal so that different difficulty levels of lineups are generated, and therefore, the estimated power curve will be smooth and continuous. Parameter $\sigma \in \{0.5, 1, 2, 4\}$ and $b \in \{0.25, 1, 4, 16, 64\}$ are used in experiment I and II respectively. Figure 6 and 9 demonstrate the impact of these two parameters. A large value of σ will increase the variation of the error of the non-linearity model and decrease the visibility of the visual pattern. The parameter b controls the standard deviation of the error across the support of the regressor. Given $x \neq a$, a larger value of b will lead to a larger ratio of the variance at x to the variance at $x - a = 0$, making the visual pattern more obvious.

Three different sample sizes are used (n = 50, 100, 300) in all three experiments. It can be observed from Figure 5 that with fewer data points drawn in a residual plot, the visual pattern is more difficult to be detected.

3.2.2. Subject allocation

Three replications are made for each of the parameter values shown in Table 2 resulting in $(4 \times 4 \times 4 \times 3 + 4 \times 3 \times 5 \times 3) \times 3 = 1116$ different lineups. In addition, each lineup is designed to be evaluated by five different subjects. After attempting some pilot studies internally in our research group, we decide to present a block of 20 lineups to every subject. And to ensure the quality of the survey data, two lineups with obvious visual patterns are included as attention checks. Thus, $576 \times 5/(20 - 2) = 160$ and $540 \times 5/(20 - 2) = 150$ subjects are recruited to satisfy the design of the experiment I and experiment II respectively.

As mentioned in Section 2.3.3, α used in Equation 1 needs to be estimated using null lineups. Hence, three replications of 36 lineups with all combinations of n and fitted value distributions are included in experiment III. In these lineups, the data of the data plot is generated from a model with zero effect size, while the data of the 19 null plots are generated using the same simulation method discussed in Section 2.3.2. This generation procedure differs from the canonical Rorschach lineup procedure, which requires that all 20 plots are generated from the null hypothesis. However, these lineups serve the same fundamental purpose: to assess the number of visually interesting plots generated under the null hypothesis.

To account for the fact that our simulation method for these lineups is not the Rorschach procedure, we use the method suggested in VanderPlas et al. (2021) for typical lineups containing a data plot to estimate α. We have included a sensitivity analysis in the Appendix to examine the impact of the variance of the α estimate on our findings.

All lineups consist of only null plots are planned to be evaluated by 20 subjects. However, presenting only these lineups to subjects are considered to be bad practices as subjects will lose interest quickly. Therefore, we plan to collect 6 more evaluations on the 279 lineups with uniform fitted values, result in $(36 \times 20 + (4 \times 4 \times 3 + 3 \times 5 \times 3) \times 3 \times 6)/(20 - 2) = 133$ subjects recruited for experiment III.

3.2.3. Collecting results

Subjects for all three experiments are recruited from an crowdsourcing platform called Prolific (Palan and Schitter 2018). Prescreening procedure is applied during the recruitment, subjects are required to be fluent in English, with 98% minimum approval rate and 10 minimum submissions in other studies.

During the experiment, every subject is presented with a block of 20 lineups. A lineup consists of a randomly placed data plot and 19 null plots, which are all residual plots drawn with raw residuals on the y-axis and fitted values on the x-axis. An additional horizontal red line is added at $y = 0$ as a helping line.

The data of the data plot is simulated from one of two models described in Section 3.1, while the data of the remaining 19 null plots are generated by the residual rotation technique discussed in Section 2.3.2.

In every lineup evaluation, the subject is asked to select one or more plots that are most different from others, provide a reason for their selections, and evaluate how different they think the selected plots are from others. If there is no noticeable difference between plots in a lineup, subjects are permitted to select zero plots without providing the reason. No subject are shown the same lineup twice. Information about preferred pronoun, age group, education, and previous experience in visual experiment are also collected. A subject's submission is only accepted if the data plot is identified for at least one attention check. Data of rejected submissions are discarded automatically to maintain the overall data quality.

4. Results

4.1. *Overview*

There are 2880, 2700 and 1674 lineups evaluation made by 160, 150 and 133 subjects recruited for experiment I, II and III respectively. In the total of 7974 lineup evaluations, 3744 use lineups produced by the non-linearity model, and 4230 use lineups produced by the heteroskedasticity model. Besides, there are 886 attention checks and 720 evaluations on null lineups needed for the estimate of α not included in the analysis. The collated dataset is provided in `vi_survey` of the `visage R` package.

In the following analysis, lineups with uniform fitted values will be the focus. Visual patterns are more likely to be revealed under a uniform distribution. Additionally, we have collected extra evaluations on these lineups, which will result in more reliable analysis. Analysis of lineups with other fitted value distributions can be found in Section **??**.

4.2. *Power comparison of different tests*

Figure 11 shows the estimated power of visual test on lineups produced by the non-linearity model with uniform fitted values, against the natural logarithm of the effect

$log_e(E)$, with a 5% significance level. At the bottom of the figure 11, there are a sequence of example residual plots with increasing levels of $log_e(E)$. Readers can evaluate them from left to right and determine at which level the departure from a good residual plot becomes detectable.

As discussed in Section 2.2, many conventionally tests are available for detecting residual departures. Implementation-wise, the built-in R package `stats` provides some commonly used residual-based tests, such as Shapiro-Wilk test. A more comprehensive collection of regression diagnostics tests can be found in the R package `lmtest` (Zeileis and Hothorn 2002). In terms of heteroskedasticity diagnostics, the R package `skedastic` (Farrar 2020) collects and implements 25 existing conventional tests published since 1961.

To compare the power of visual test and conventional test, we pick RESET test (`resettest`) and Breusch-Pagan test (`bptest`) from the R package `lmtest`, and Shapiro-Wilk test (`shapiro.test`) from the built-in R package `stats`. Among them, RESET test is the only exact and appropriate test in this scenario. Both the Breusch-Pagan test and the Shapiro-Wilk test are approximate and inappropriate tests. Their estimated power is shown in Figure 11. To set up the RESET test, we include different powers of fitted values as proxies. According to Ramsey (1969), there are no general rules for the power of the fitted values needed by the RESET test, but it finds power up to four is usually sufficient. Thus, we follow this guideline to conduct the RESET test. For the Breusch-Pagan test, the choice of regressors in the auxiliary regression is left to the user (Breusch and Pagan 1979). But as Waldman (1983) suggested, it is a good choice for the set of auxiliary regressors in the Breusch-Pagan test be the same as the White test. Thus, we include both $\boldsymbol{x}$ and $\boldsymbol{x}^2$ in the auxiliary regression.

Figure 12 is similar to Figure 11, but shows corresponding information on lineups produced by the heteroskedasticity model. In this scenario, the visual test is compared to an approximate test - Breusch-Pagan test, and two other inappropriate tests - RESET test and Shapiro-Wilk test.

For non-linearity patterns, the power curve of RESET test climbs aggressively from 15% to 69% as $log_e(E)$ increases from 0 to 2, while power of other tests respond inactively to the change of effect, showing that RESET test is way more sensitive to the type of model defects that being considered. Meanwhile, no noticeable visual features can be spotted from the example residual plots.

In terms of heteroskedasticity patterns, the power of Breusch-Pagan test is also almost always greater than the power of visual test. For $0 \leq log_e(E) \leq 2$, where the power curve of the visual test remains at a low level, the Breusch-Pagan test still have a decent amount of chance of rejecting H_0. Similarly, the visual feature is nearly unobservable from the example residual plots.

The power of visual test arises steadily as $log_e(E)$ increases from 2 to 5 for both non-linearity patterns and heteroskedasticity patterns, suggesting that the effect starts to make significant impact on the degree of the presence of the designed visual features. This can also be observed from the example residual plots that when $log_e(E) = 2.5$, a weak "S-shape" and a weak "triangle" shape are presented in Figure 11 and Figure 12 respectively. The visual pattern becomes much clearer as $log_e(E)$ increases. At $log_e(E) \approx 6$, the power reaches almost 100%.

The power of all inappropriate tests except for RESET test shows improvement as the effect increases but at a lower rate than the visual test in both scenarios. This coincides the point made by Cook and Weisberg (1982) that residual-based tests for a specific type of model defect may be sensitive to other types of model defects. The power curve of RESET test remains at around 5% in Figure 12 since there are no

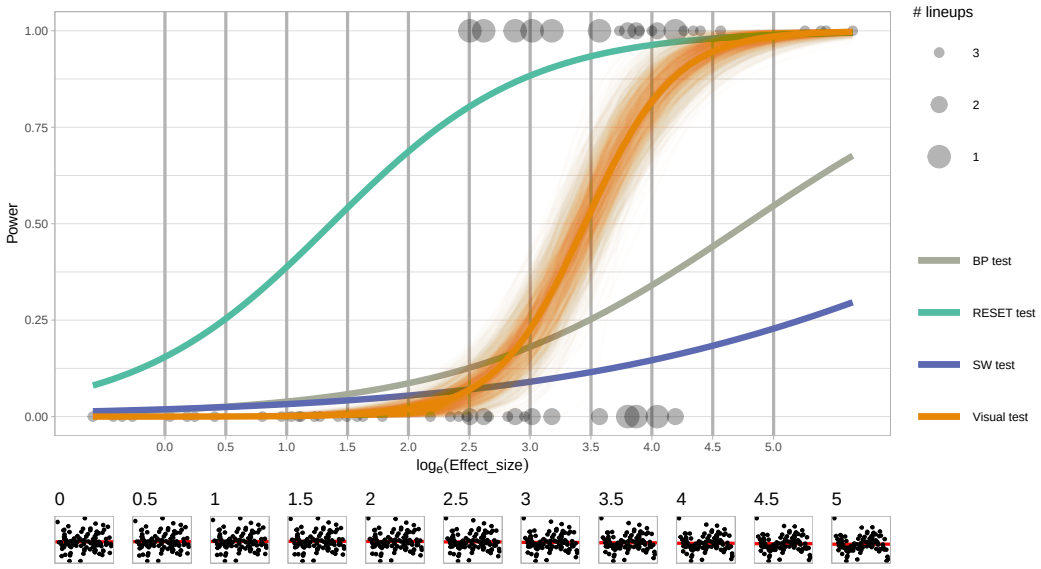

Figure 11. Comparison of power between different tests for non-linear patterns (uniform fitted values only). Main plot shows the power curves estimated using logistic regression, with dots indicating human evaluations of lineups. Surrounding lines of the visual test show the estimated power of 1000 bootstrap samples. Small row of plots shows typical residual plots corresponding to specific effect sizes, marked by dashed lines in main plot. Where would you draw the line of too much non-linearity in the residuals? For the RESET test this is around log effect size 1.5, but for the visual test it is around 3.5.

non-linear terms leave out in the heteroskedasticity model and H_0 of the test is always satisfied.

Overall, the power comparison suggests that conventional tests differs significantly from visual tests in two regression diagnostics scenarios designed by us. Visual test have much higher tolerance of the residual departures than the conventional test. Since fail to reject H_0 in a visual test usually means that there are no obvious visual discoveries found in the residual plot, analysts and the general public as the consumers of the output may not be convinced of the existence of significant residual departures in spite of the rejection of H_0 given by the conventional test. Even if the rejection is accepted, the model violation may be considered as impactless due to the fact that they are not clearly visible. Besides, the sensitivity of the conventional test could also distract and discourage analysts from finding simple but good linear approximation to the data. The rejection of H_0 because of human acceptable and negligible residual departures is not practically meaningful and useful. This may limit the popularity of conventional tests in residual diagnostics among analysts.

4.3. *Comparison of test decisions based on p-values*

The power comparison illustrates that appropriate conventional tests will reject H_0 more aggressively than visual tests. In this section, we explore how often they agree with each other by investigating the rejections for the two model designs based on p-values for each lineup.

Figure 13 provides a mosaic plot showing the rejection rate of visual tests and conventional tests for both non-linearity patterns and heteroskedasticity patterns.

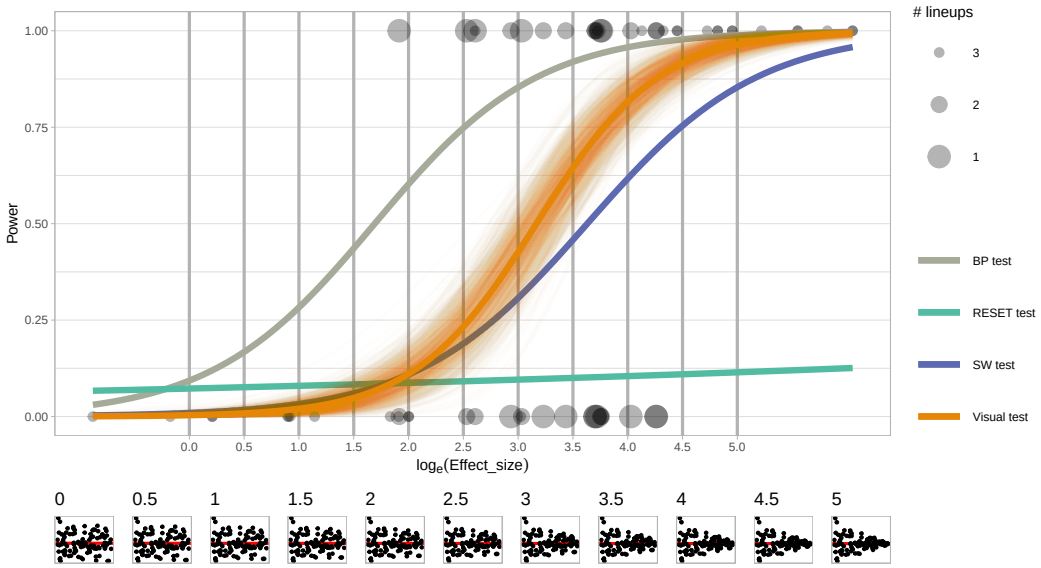

Figure 12. Comparison of power between different tests for heteroskedasticity patterns (uniform fitted values only). Main plot shows the power curves, with dots indicating human evaluations of lineups. Surrounding lines of the visual test show the estimated power of 1000 bootstrap samples. Small row of plots shows typical residual plots corresponding to specific effect sizes, marked by dashed lines in main plot. Where would you draw the line of too much heteroskedasticity in the residuals? For the BP test this is around log effect size 1.5, but for the visual test it is around 3.

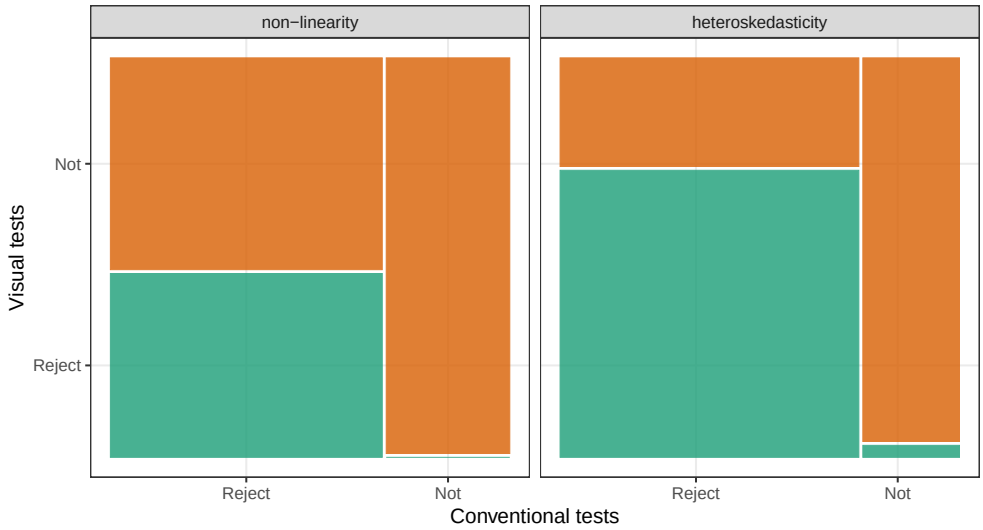

Figure 13. Rejection rate (p-value < 0.05) of visual test conditional on the conventional test decision on non-linearity (left) and heteroskedasticity (right) lineups (uniform fitted values only) displayed using a mosaic plot. The visual test rejects less frequently than the conventional test. We would generally expect that the visual test would only reject when the conventional test does. Surprisingly, there is one lineup in the heteroskedasticity study where this is not the case.

For lineups containing non-linearity patterns, conventional tests reject 69% and visual tests reject 32% of the time. Of the lineups rejected by the conventional test, 46% are rejected by the visual test, that is, approximately half as many as the conventional test. There are no lineups that are rejected by the visual test but not by the conventional test.

In terms of lineups containing heteroskedasticity patterns, 76% are rejected by conventional tests, while 56% are rejected by visual tests. When the conventional test rejects a lineup, there is a great chance (73%) the visual test will also reject it.

Surprisingly, the visual test rejects 1 of the 33 (3%) of lineups where the conventional test does not reject. Figure 14 shows this lineup. The data plot in position seventeen displays a relatively strong heteroskedasticity pattern, and has a strong effect size ($log_e E = 4.02$). This is reflected by the visual test p-value = 0.026, but the Breusch-Pagan test p-value = 0.056, slightly above the significance cutoff of 0.05. This lineup was evaluated by 11 subjects, it has experimental factors $n = 50$ (small sample size), $\sigma = 1$, and a uniform distribution for the fitted values. It must be the small sample size that may have resulted in the lack of detection.

4.4. *Effect of shape of non-linearity*

A primary factor contributes to the non-linearity model is the shape of the non-linearity. According to Figure 15, conventional tests have almost identical power in testing the "U", the "S" and the "M" shape, except for testing the "Triple-U" shape, which have significant lower power. To understand why this is, one needs to return to the way the RESET test is applied. It requires a parameter indicating degree of fitted values to test for, and the recommendation is to generically use four (Ramsey (1969)). However, the "Triple-U" shape constructed from the Hermite polynomials use power up to 18. If the RESET test had been applied using a higher power no less than six, the power curve of "Triple-U" shape will have little difference to other shapes. The recommendation of the polynomial power for the RESET should be revised, perhaps. This illustrates the sensitivity of conventional testing to the parameters, and it also points to a limitation that one needs to know the data structure in order to set the parameters for the test.

For visual tests, we expect the "U" shape will be the easiest one to be detected by subjects followed by the "S", "M" and "Triple-U" shape. From Figure 15, it can be observed that the power curves are mostly aligned with our expectation, except for the "M" shape, which is as easy to be detected as the "U" shape. This implies the readability of the shape do not strictly follow the degree of the polynomials.

4.5. *Effect of shape of heteroskedasticity*

We also investigate the impact of different heteroskedasticity shapes on power of conventional tests and visual tests. In theory, the "Left-triangle" and the "Right-triangle" shapes are functionally identical from the point of view of a Breusch-Pagan test. As shown in Figure 16, this is indeed the case where little difference between the power curves can be perceived. Similarly, visual tests will have the same power of detecting these two shapes if they are equally likely to be identified. However, it can be observed from Figure 16 that the power curve for the "Left-triangle" shape is constantly higher than the one for the "Right-triangle" shape, indicating a potential favour of orientation by human, which is worth to be explored in future studies.

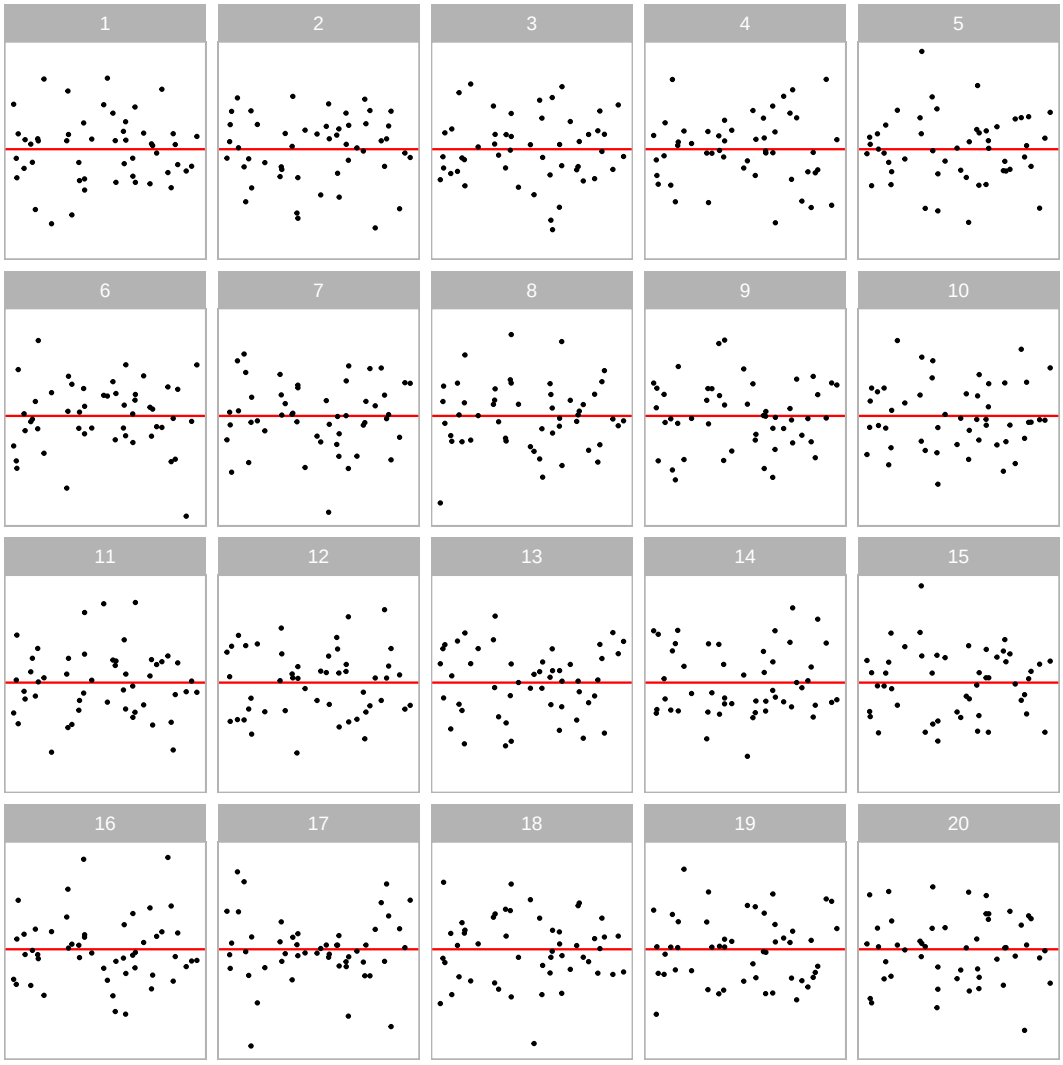

Figure 14. The single heteroskedasticity lineup that is rejected by the visual test but not by the BP test. The data plot at panel 17 contains a "Butterfly" shape. It has effect size = 4.02, somewhat surprising that it is not detected by the BP test.

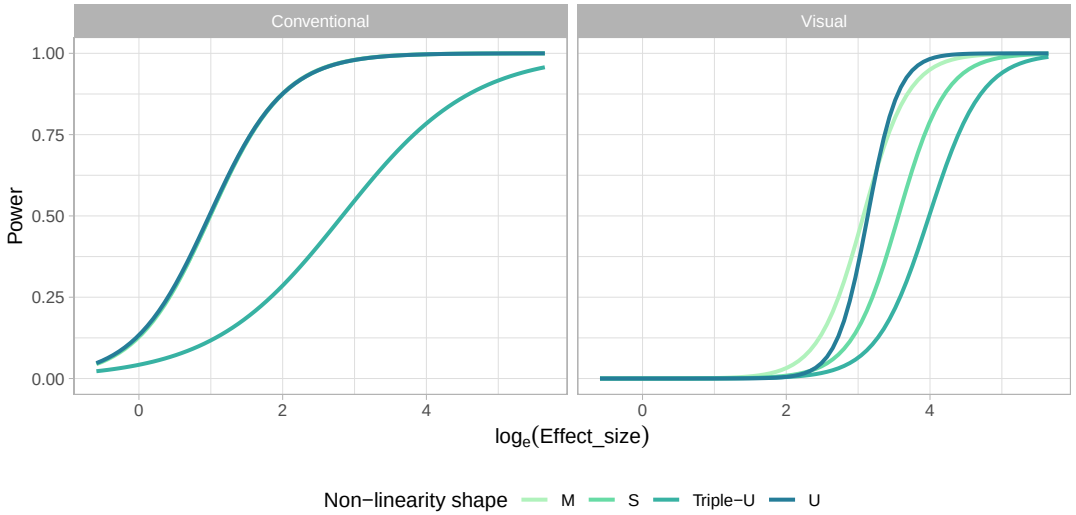

Figure 15. Power of conventional tests and visual tests on lineups containing four different non-linearity shapes. Power curves with higher order of non-linearity are drawn with deeper colours. The default RESET tests under-perform significantly in detecting "Tripe-U" shape. To achieve a similar power as other shapes, a higher order polynomial parameter needs to be used for the RESET test. But this means the order needs to be known prior to testing.

The chance of detecting the "Butterfly" shape is roughly the same as the chance of detecting the "Left-triangle" shape in both tests. The only noticeable difference from Figure 16 is the Breusch-Pagan test has slightly higher power for detecting this shape compared to other two shapes when the effect size is low.

4.6. *Effect of fitted value distributions*

The prediction in a regression model is $E(Y|X)$, that is, it is conditional on observed values of the predictors. The distribution of X, or consequently $\hat{Y}$, may however affect the ability to read any patterns in the residuals. The four distributions of fitted values, used in this experiment, were designed to examine this.

Figure 17 illustrates the change of power of visual tests and conventional tests for different fitted distributions used in the non-linearity model and the heteroskedasticity model. We focus on $log(E) > 2$ where the visual patterns start to become recognisable.

For non-linearity patterns, we do not observe significant difference in power of visual test between distributions, except for the discrete uniform distribution. It could be due to the fact that humans have difficulties in recognising visual patterns from discreteness, which makes the shape disconnected and incomplete (ref?). In terms of conventional tests, the discrete uniform distribution is instead the one with the greatest power, followed by the lognormal distribution, the uniform distribution and the normal distribution.

For heteroskedasticity patterns, the power of visual test on lineups with uniform distribution has the greatest power. This is as expected since other three distributions could reduce the chance of revealing the underlying visual pattern because of uneven data points. Although the power of visual test under the discrete uniform distribution is the lowest in the case of non-linearity patterns, it has a relatively great power this time. Considering for heteroskedasticity, the visual patterns are usually detected by connecting the maximum and minimum residuals separately at different fitted values.

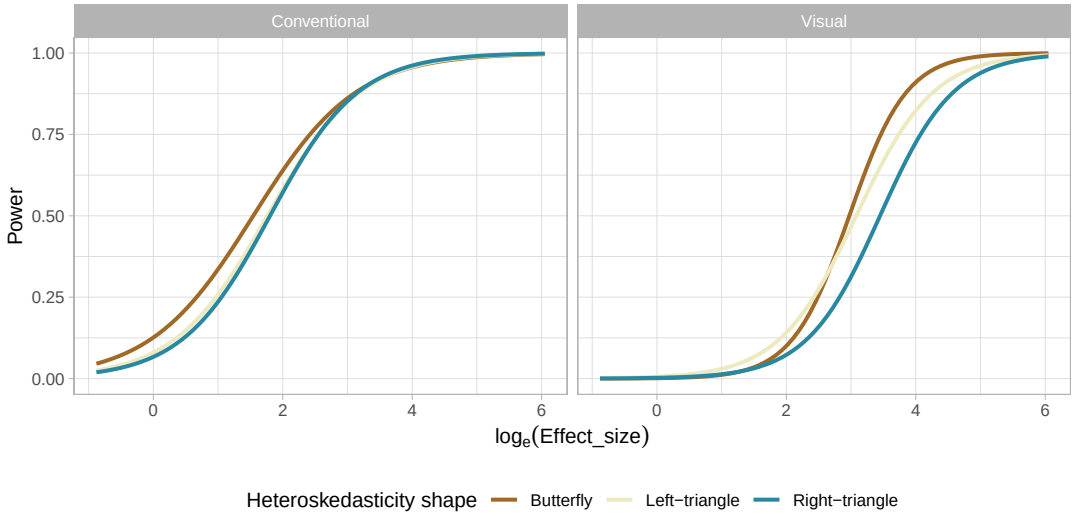

Figure 16. Power of conventional tests and visual tests on lineups produced by the heteroskedasticity model with three different shapes controlled by the parameter a.

It will not be greatly affected by the discrete uniform distribution compared to the uniform distribution. Visual tests have low power on lineups with the normal distribution and the lognormal distribution. This suggests the use of transformation of fitted values in reading residual plots. In the case of conventional tests, the discrete uniform distribution consistently perform well, followed by the uniform distribution, the normal distribution and the lognormal distribution.

In regards of the type of the simulation model, the power of visual tests under different fitted value distributions are always lower than the corresponding conventional tests, which is consistent with the result reported in the previous sections.

5. Conclusion

This paper has demonstrated that conventional residual-based statistical tests are more sensitive to departures from assumptions for residuals than visual tests evaluated by humans in the diagnostics of linear regression models. Three experiments are conducted with two simulation models designed for non-linearity and heteroskedasticity model defects being applied. Methods for obtaining the p-value of visual tests proposed by VanderPlas et al. (2021) are extended to allowed for multiple selections in a lineup protocol.

Residual plots rejected by visual tests are rarely not rejected by conventional tests. Throughout the three human subject experiments we conducted, departures of residuals that are of little practical sense to humans will often be alarmed by conventional tests, supporting the common guidance given by experts to utilize graphical representations of data in model diagnostics.

Conventional tests relying on
such as RESET test

1. summarise paper content
2. briefly describe what was learned
3. how does this apply more broadly than simple regression

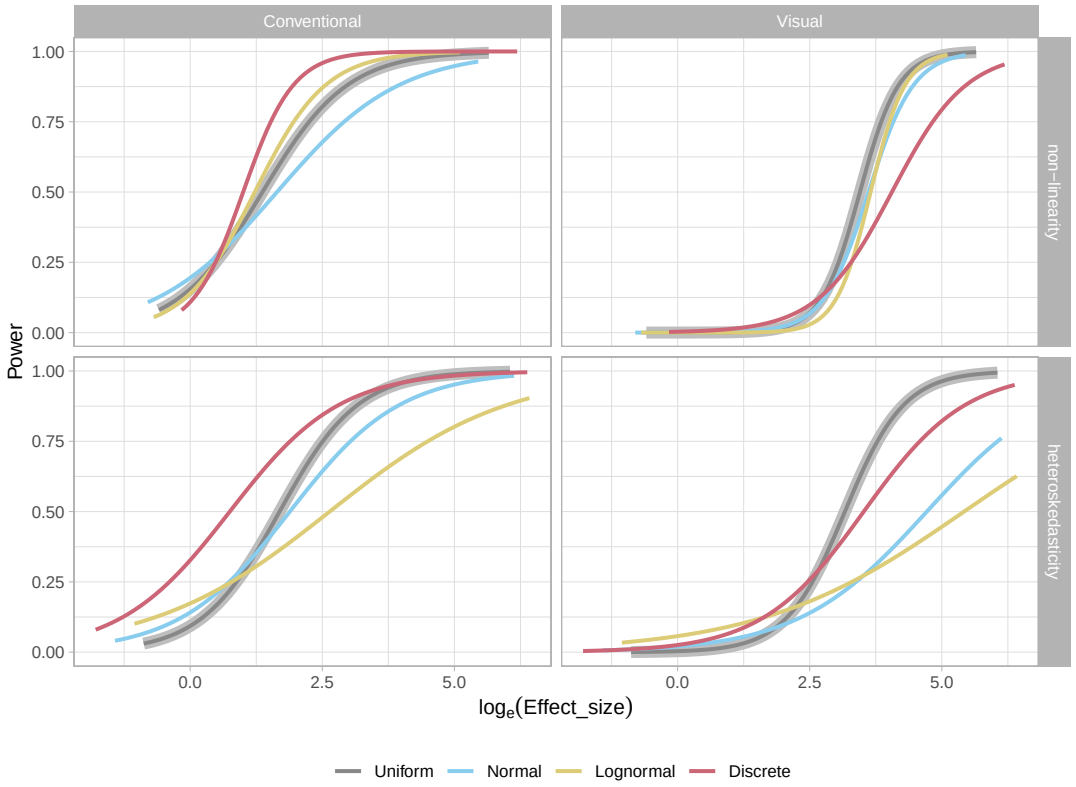

Figure 17. Comparison of power on lineups with different fitted value distributions, for conventional tests and visual tests. Each power curve of the uniform distribution is surrounded by a grey line to indicate it is the baseline in its group. For both types of tests power does change. The discrete fitted value distribution substantially lowers the power for non-linearity patterns in residual plots. For heteroskedasticity patterns, normal and lognormal distributions produce lower power.

4. where does this lead? Computer vision residual plot reading

6. Acknowledgement

1. software tools
2. point to github repo (create a new one)
3. summarise the supplementary material

References

Abramowitz, Milton, and Irene A Stegun. 1964. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Vol. 55. US Government printing office.

Belsley, David A, Edwin Kuh, and Roy E Welsch. 1980. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.

Box, George EP. 1976. "Science and statistics." *Journal of the American Statistical Association* 71 (356): 791–799.

Breusch, T. S., and A. R. Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* 47 (5): 1287–1294.

Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne, and Hadley Wickham. 2009. "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.

Buja, Andreas, Dianne Cook, and D Swayne. 1999. "Inference for data visualization." In *Joint Statistics Meetings, August*, .

Cleveland, William S., and Robert McGill. 1984. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods." *Journal of the American Statistical Association* 79 (387): 531–554.

Cook, R Dennis, and Sanford Weisberg. 1982. *Residuals and influence in regression*. New York: Chapman and Hall.

Cook, R Dennis, and Sanford Weisberg. 1999. *Applied regression including computing and graphics*. John Wiley & Sons.

Draper, Norman R, and Harry Smith. 1998. *Applied regression analysis*. Vol. 326. John Wiley & Sons.

Farrar, Thomas J. 2020. *skedastic: Heteroskedasticity Diagnostics for Linear Regression Models*. Bellville, South Africa. R Package Version 1.0.0.

Hermite, M. 1864. *Sur un nouveau développement en série des fonctions*. Imprimerie de Gauthier-Villars.

Jarque, Carlos M, and Anil K Bera. 1980. "Efficient tests for normality, homoscedasticity and serial independence of regression residuals." *Economics letters* 6 (3): 255–259.

Kahle, David. 2013. "mpoly: Multivariate Polynomials in R." *The R Journal* 5 (1): 162–170.

Laplace, Pierre-Simon. 1820. *Théorie analytique des probabilités*. Vol. 7. Courcier.

Loy, Adam, and Heike Hofmann. 2013. "Diagnostic tools for hierarchical linear models." *Wiley Interdisciplinary Reviews: Computational Statistics* 5 (1): 48–61.

Loy, Adam, and Heike Hofmann. 2014. "HLMdiag: A suite of diagnostics for hierarchical linear models in R." *Journal of Statistical Software* 56: 1–28.

Loy, Adam, and Heike Hofmann. 2015. "Are you normal? the problem of confounded residual structures in hierarchical linear models." *Journal of Computational and Graphical Statistics* 24 (4): 1191–1209.

Majumder, Mahbubul, Heike Hofmann, and Dianne Cook. 2013. "Validation of Visual Statistical Inference, Applied to Linear Models." *Journal of the American Statistical Association* 108 (503): 942–956.

Montgomery, DC, and EA Peck. 1982. *Introduction to linear regression analysis*.

Palan, Stefan, and Christian Schitter. 2018. "Prolific. ac—A subject pool for online experiments." *Journal of Behavioral and Experimental Finance* 17: 22–27.

Ramsey, J. B. 1969. "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis." *Journal of the Royal Statistical Society. Series B (Methodological)* 31 (2): 350–371.

Roy Chowdhury, Niladri, Dianne Cook, Heike Hofmann, Mahbubul Majumder, Eun-Kyung Lee, and Amy L. Toth. 2015. "Using visual statistical inference to better understand random class separations in high dimension, low sample size data." *Computational Statistics* 30 (2): 293–316.

Shapiro, Samuel Sanford, and Martin B Wilk. 1965. "An analysis of variance test for normality

(complete samples)." *Biometrika* 52 (3/4): 591–611.
Silvey, Samuel D. 1959. "The Lagrangian multiplier test." *The Annals of Mathematical Statistics* 30 (2): 389–407.
VanderPlas, Susan, Christian Röttger, Dianne Cook, and Heike Hofmann. 2021. "Statistical significance calculations for scenarios in visual inference." *Stat* 10 (1): e337.
Waldman, Donald M. 1983. "A note on algebraic equivalence of White's test and a variation of the Godfrey/Breusch-Pagan test for heteroscedasticity." *Economics Letters* 13 (2-3): 197–200.
White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (4): 817–838.
Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10.