ARTICLE TEMPLATE

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

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

[a]Department of Econometrics and Business Statistics, Monash University, Clayton, VIC, Australia; [b]Biological Data Science Institute, Australian National University, Acton, ACT, Australia; [c]Research School of Finance, Actuarial Studies and Statistics, Australian National University, Acton, ACT, Australia; [d]Department of Statistics, University of Nebraska, Lincoln, Nebraska, USA

ARTICLE HISTORY
Compiled April 14, 2023

ABSTRACT
Abstract to fill.

KEYWORDS
statistical graphics; data visualization; visual inference; hypothesis testing; reression analysis; cognitive perception; simulation; practical significance; effect size

1. Introduction

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

Diagnostics are the key to determining whether there is anything importantly wrong with a model. 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.

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

In addition, scatterplots of residuals plotted against the fitted values and each of the explanatory variables, are made to scrutinize their relationships. 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 dependency or heteroskedasticity. A non-linear dependency would suggest that the

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

model needs to have some additional non-linear terms. Heteroskedasticity suggests that the error is dependent on the predictors, and hence violates the independence assumption. However, correctly judging whether NO pattern exists in a residual plot is a difficult task for humans. We humans will almost always will see a pattern, so the question that really needs answering is whether any pattern perceived is consistent with randomness, purely sampling variability or noise. It is especially difficult to teach this to new analysts and students (Loy 2021). To answer this, there have been numerous conventional hypothesis tests made available to test for non-linear dependence (e.g. Ramsey 1969), and heteroskedasticity (e.g. Breusch and Pagan 1979).

Linear regression is a well-established procedure, and there is considerable literature describing diagnostic procedures, e.g. Draper and Smith (1998), Montgomery and Peck (1982), Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1999) and Cook and Weisberg (1982). An interesting details, is that despite the abundance of conventional tests, ALL of these writings advise that plotting residuals is an essential tool for diagnosing regression model problems. Draper and Smith (1998) and Belsley, Kuh, and Welsch (1980) explain that residual plots are usually revealing when the assumptions are violated. Cook and Weisberg (1999) thinks formal tests and graphical procedures are complementary and both have a place in residual analysis, but they focus on graphical methods rather than on formal testing. Montgomery and Peck (1982) even suggests that residual plots are more informative in most practical situations than the corresponding conventional hypothesis tests.

The common wisdom of experts is that the 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 is, is the subject of this paper.

The paper is structured as follows. The next section describes the background on the types of departures that one expects to detect, and outlines a formal statistical process for reading residual plots, called visual inference. Section 4 details the experimental design to compare the decisions made by formal hypothesis testing, and how humans would read diagnostic plots. The results are reported in Section 5. We conclude with a discussion of the presented work, and ideas for future directions.

2. Background

2.1. *Departures from good residual plots*

Graphical summaries where residuals are plotted against fitted values, or other functions of the predictors (expected to be approximately orthogonal to the residuals) are referred to as standard residual plots in Cook and Weisberg (1982). Figure 1A shows an example of an ideal residual plot where points are symmetrically distributed around the horizontal zero line (red), with no discernible patterns. There can be various types of departures from this ideal pattern. Non-linearity, heteroskedasticity and non-normality are three commonly checked departures.

Model misspecification occurs if functions of predictors that needed to accurately describe the relationship between response and predictors were omitted. Any non-linear pattern visible in the residual plot could be indicative of this problem. An example residual plot containing visual pattern of non-linearity is shown in Figure 1B. One can clearly observe the "S-shape" from the residual plot, which corresponds to cubic term that should have been included in the model.

Heteroskedasticity refers to the presence of non-constant error variance in a regres-

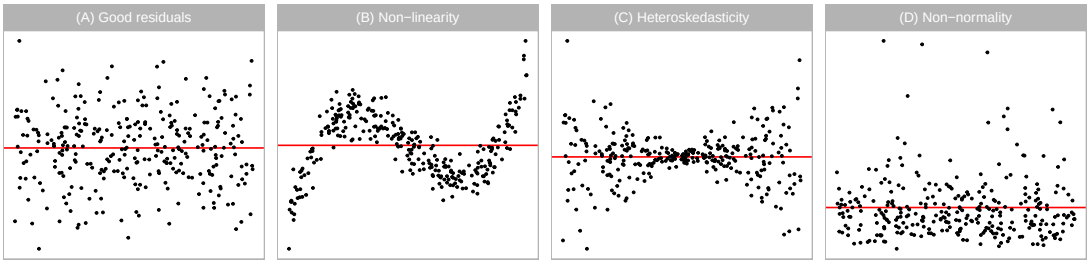

Figure 1. Example residual vs fitted value plots (red line indicates 0): (A) classically good looking residuals, (B) non-linear pattern indicates that the model has not captured a non-linear association, (C) heteroskedasticity indicating that variance around the fitted model is not uniform, and (D) non-normality 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 value plot.

sion model. It indicates that the distribution of residuals depends on the predictors, violating the independence assumption. This can be seen in a residual plot as an inconsistent spread of the residuals relative to the fitted values or predictors. An example is the "butterfly" shape shown in Figure 1C, or a "left-triangle" and "right-triangle" shape where the smallest variance occurs at one side of the horizontal axis.

Figure 1D) shows a scatterplot where the residuals have a skewed distribution, as seen by the uneven vertical spread. Unlike non-linearity and heteroskedasticity, non-normality is usually detected with a different type of residual plot, a histogram or normal probability plot. Because we focus on scatterplots, non-normality is not one of the departures examined in this paper.

2.2. *Conventionally testing for departures*

Many different hypothesis tests are available to detect specific model defects. For example, the presence of heteroskedasticity can usually be tested by applying the White test (White 1980) or the BP 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. To test specific forms of 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 predictors, or the significance of proxy variables as in the Ramsey Regression Equation Specification Error Test (RESET) (Ramsey 1969). The SW test (Shapiro and Wilk 1965) is the most widely used test of non-normality included by many of the statistical software programs. The Jarque–Bera test (Jarque and Bera 1980) is also used to directly check whether the sample skewness and kurtosis match a normal distribution.

Table 1 displays the p-values from the RESET, BP and SW tests applied to the residual plots in Figure 1. The RESET test and BP test were computed using the `resettest` and `bptest` functions from the R package `lmtest`, respectively. The SW test was computed using the `shapiro.test` from the core R package `stats`. (The the R package `skedastic` (Farrar 2020) contains a large collection of tests for heteroskedasticity.) Although, the RESET test is exact, it requires the selection of a power parameter. According to Ramsey (1969), a power of four is recommended, which we adopted in our analysis. The BP and SW tests are approximate.

We would expect the RESET test for non-linearity to reject residual plot B, the BP test for heteroskedasticity to reject the residual plot C, and SW test for non-normality to reject residual plot D, which they all do and correctly fail to reject residual plot A.

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 BP and the SW 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	BP	SW
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*

Interestingly, the BP and SW tests also reject the residual plots exhibiting structure that they weren't designed for. Cook and Weisberg (1982) discusses that most residual-based tests for particular types of departure from model assumptions are also sensitive to other types of departures, that is, correctly rejected but for the wrong reason, a phenomenon known as the "Type III error". Also, some types of departure can also have elements of other types of departure, for example, non-linearity could be viewed as heteroskedasticity. Additionally, other problesms such as outliers can trigger the rejection (Cook and Weisberg 1999).

With large sample sizes, hypothesis tests are may reject a residual plot when there is only a slight departure (XXX is there a reference for this statement?). While such rejections may be statistically correct, their sensitivity may render the results impractical. Therefore, a key goal of residual plot diagnostics is to identify potential issues that could lead to incorrect conclusions or errors in subsequent analyses. However, minor defects in the model are unlikely to have a significant impact and may be best disregarded for practical purposes. In fact, the experiment discussed in this paper specifically addresses this.

2.3. *Visual test procedure based on lineups*

The examination of data plots to infer signal or patterns (or lack thereof) is fraught with variation in the human ability to interpret and decode the information embedded in graph (Cleveland and McGill 1984). Such examination thus feels there is a subjective nature to evaluate, say, diagnostic plots to infer appropriateness of statistical models, although arguably we would expect such evaluation should be done objectively and such human-to-human variation in "reading" data plots should be minimised.

In practice, over-interpretation of a single plot is common. For instance, in Roy Chowdhury et al. (2015), some over-interpreted the separation between gene groups in a two-dimensional projection of a linear discriminant analysis when in fact there are no differences in the expression levels between the gene groups. To mitigate this over-interpretation, Buja et al. (2009) proposed a line-up protocol to assess plots in a manner analogous to the null hypothesis significance testing (NHST) framework. More specifically, 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*, have the identical graphical procedure as the data plot except the data is replaced with a data generation mechanism that is consistent with the null hypothesis, 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

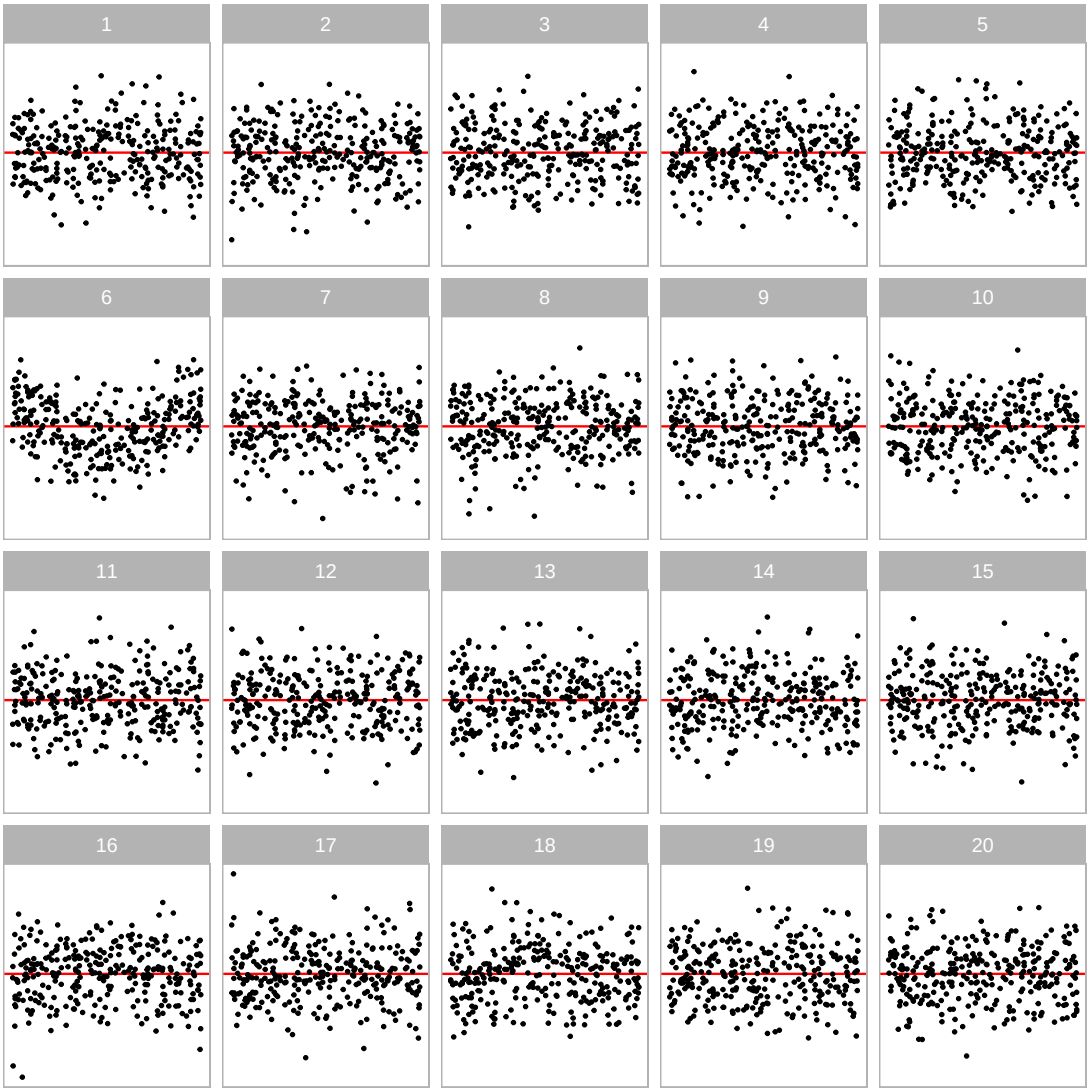

Figure 2. Visual testing is conducted using a lineup, as in the example here. The residual plot computed from the observed data (plot $2^2 + 2$, exhibiting non-linearity) is embedded among 19 null plots, where the residuals are simulated from a standard error model. Computing the p-value requires that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

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 $1/m$. This protocol requires apriori specification of H_0 (or at least a null data generating mechanism) much like the requirement of knowing the distribution of the test statistic in NHST.

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. In fact, the actual residual plot is obtained from a misspecified regression model with missing non-linear terms.

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 predictors, then rescaling it by the ratio of residual sum of square in two regressions.

The effectiveness of lineup protocol for regression analysis is validated by Majumder, Hofmann, and Cook (2013) under relatively simple settings with up to two predictors. Their results suggest that visual tests are capable of testing the significance of a single predictor 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. 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.

3. Calculation of statistical significance and test power

3.1. *What is being tested?*

In diagnosing a model fit from residuals, we are generally interested in *the regression model is correctly specified* (H_0) against the broad alternative *the regression model is misspecified* (H_a).

However, it is practically impossible to test this specific H_0 with conventional tests, which are constructed to measure specific departures. For example, the RESET test is formulated as $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$ against $H_a : \gamma_1 \neq 0$ or $\gamma_2 \neq 0$ or $\gamma_3 \neq 0$, from $y = \tau_0 + \sum_{i=1}^{p} \tau_p x_p + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + \gamma_3 \hat{y}^4 + u, \quad u \sim N(0, \sigma_u^2)$. Similarly, the BP test is designed to specifically test H_0 : *error variances are all equal* ($\zeta_i = 0$ for $i = 1, .., p$) versus the alternative H_a : *that the error variances are a multiplicative function of one or more variables* (at least one $\zeta_i \neq 0$) from $e^2 = \zeta_0 + \sum_{i=1}^{p} \zeta_i x_i + u, \; u \sim N(0, \sigma_u^2)$.

One of the potential benefits of the visual test, based on the lineup protocol, is that it works as an omnibus test, able to detect a range of departures from good residuals.

3.2. *Statistical significance*

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. Conventional hypothesis tests usually have an existing method to derive or compute p-value based on the null distribution. What we need to discuss in the following is the method to estimate p-value for a visual test.

Within the context of visual inference, by involving k independent observers, the visual 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 binomial cumulative distribution function, $f(.)$ is the binomial 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 does not 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 human responses to Rorschach lineups, which is a type of lineup that consists of plots constructed from the same null data generating mechanism, are required.

3.3. *Power of the tests*

The power of a model misspecification test is the probability that H_0 is rejected given the regression model is misspecified in a specific way. 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.

The power of a conventional hypothesis test is affected by both the true parameter $\boldsymbol{\theta}$ and the sample size n. These two can be quantified in terms of effect size E to measure the strength of the residual departures from the model assumptions. Details about the effect size is provided in Section 4.2.2 after the introduction of the simulation model used in our human subject experiment. The theoretical power of a test is sometimes not a trival solution, but it can be estimated if the data generating process is known. We use a predefined model to generate a large set of simulated data under different effect sizes, and record if the conventional test rejects H_0. The probability of the conventional test rejects H_0 is then fitted by a logistic regression formulated as

$$Pr(\text{reject } H_0 | H_1, E) = \Lambda\left(log\left(\frac{0.05}{0.95}\right) + \beta_1 E\right), \tag{6}$$

where $\Lambda(.)$ is the standard logistic function given as $\Lambda(z) = exp(z)/(1+exp(z))$. The effect size E is the only predictor and the intercept is fixed to $log(0.05/0.95)$ so that $\hat{Pr}(\text{reject } H_0 | H_1, E = 0) = 0.05$, which is the desired significance level.

The power of a visual test on the other hand, may additionally 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 (Majumder, Hofmann, and Cook 2013). To address this issue, Majumder, Hofmann, and Cook (2013) models 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 effect. Then, the estimated power of a visual test evaluated by a single subject is the predicted value obtained from the mixed effects model. However, this mixed effects model does not work with scenario where subjects are asked to select one or more most different

plots. In this scenario, having the probability of a subject j correctly picking the data plot from a lineup l is insufficient 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 (See Equation 3). Therefore, we directly estimate the probability of a lineup being rejected by assuming that individual skill has negligible effect on the variation of the power. This assumption is not necessary true, but it helps simplifying the model structure, thereby obviate a costly large-scale experiment to estimate complex covariance matrices. The same model given in Equation 6 is applied to model the power of a visual test.

To study various factors contributing to the power of both tests, 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 fitted values, type of the simulation model and the shape of the residual departures.

4. Experimental design

An experiment is conducted in three data collection periods to investigate the difference between conventional hypothesis testing and visual inference in the application of linear regression diagnostics. Two types of departures, non-linearity and heteroskedasticity, are collected during data collection periods I and II. The data collection period III was designed primarily to measure human responses to null lineups so that the parameter α in Equation 1 can be estimated. Additional lineups for both non-linearity and heteroskedasticity, using uniform fitted value distribution, were included so that the participants were evaluating some lineups with signal also. It would be too frustrating for participants to only be assigned lineups with all null plots. Overall, we collected 7974 evaluations on 1152 unique lineups performed by 443 subjects throughout three data collection periods.

4.1. *Simulating departures from good residuals*

4.1.1. Non-linearity

Data collection period 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

$$\boldsymbol{y} = 1 + \boldsymbol{x} + \boldsymbol{z} + \boldsymbol{\varepsilon}, \tag{7}$$
$$\boldsymbol{x} = g(\boldsymbol{x}_{raw}, 1), \tag{8}$$
$$\boldsymbol{z} = g(\boldsymbol{z}_{raw}, 1), \tag{9}$$
$$\boldsymbol{z}_{raw} = He_j(g(\boldsymbol{x}, 2)), \tag{10}$$

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

$$g(\boldsymbol{x}, k) = (\boldsymbol{x} - min(\boldsymbol{x}))/max(\boldsymbol{x} - min(\boldsymbol{x}))2k - k, \quad \text{for} \quad k > 0. \tag{11}$$

Table 2. Levels of the factors used in data collection periods 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
			64.00		

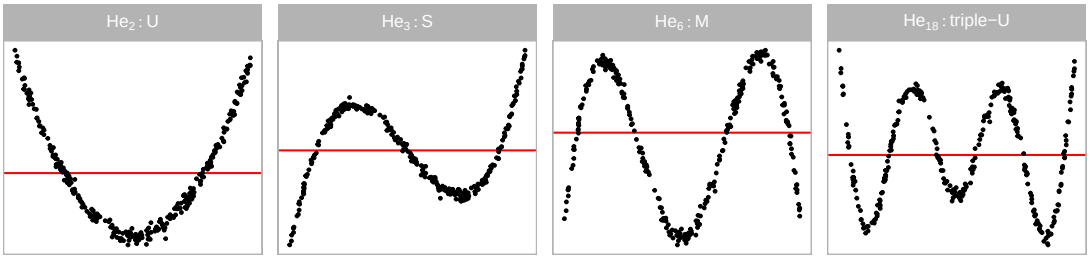

Figure 3. Polynomial forms generated for the residual plots used to assess detecting non-linearity. The four shapes are generated by varying the order of polynomial given by j in $He_j(.)$.

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

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{u}, \tag{12}$$

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 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 5 demonstrates one of the lineups used in non-linearity detection. This lineup is produced by the non-linearity model with $j = 6$. The data plot location is $2^3 - 4$. All five subjects correctly identify the data plot from this lineup.

4.1.2. Heteroskedasticity

Data collection period II is designed to study the ability of human subjects to detect the appearance of a heteroskedasticity pattern under a simple linear regression model

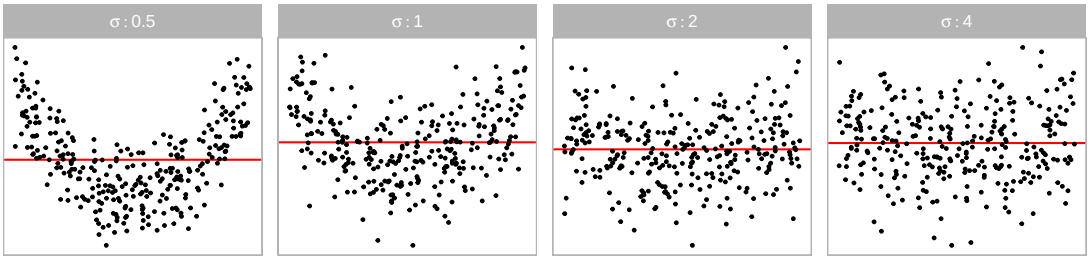

Figure 4. Examining the effect of σ on the signal strength in the non-linearity detection, for $n = 300$, uniform fitted value distribution and the "U" shape. As σ increases the signal strength decreases, to the point that the "U" is almost unrecognisable when $\sigma = 4$.

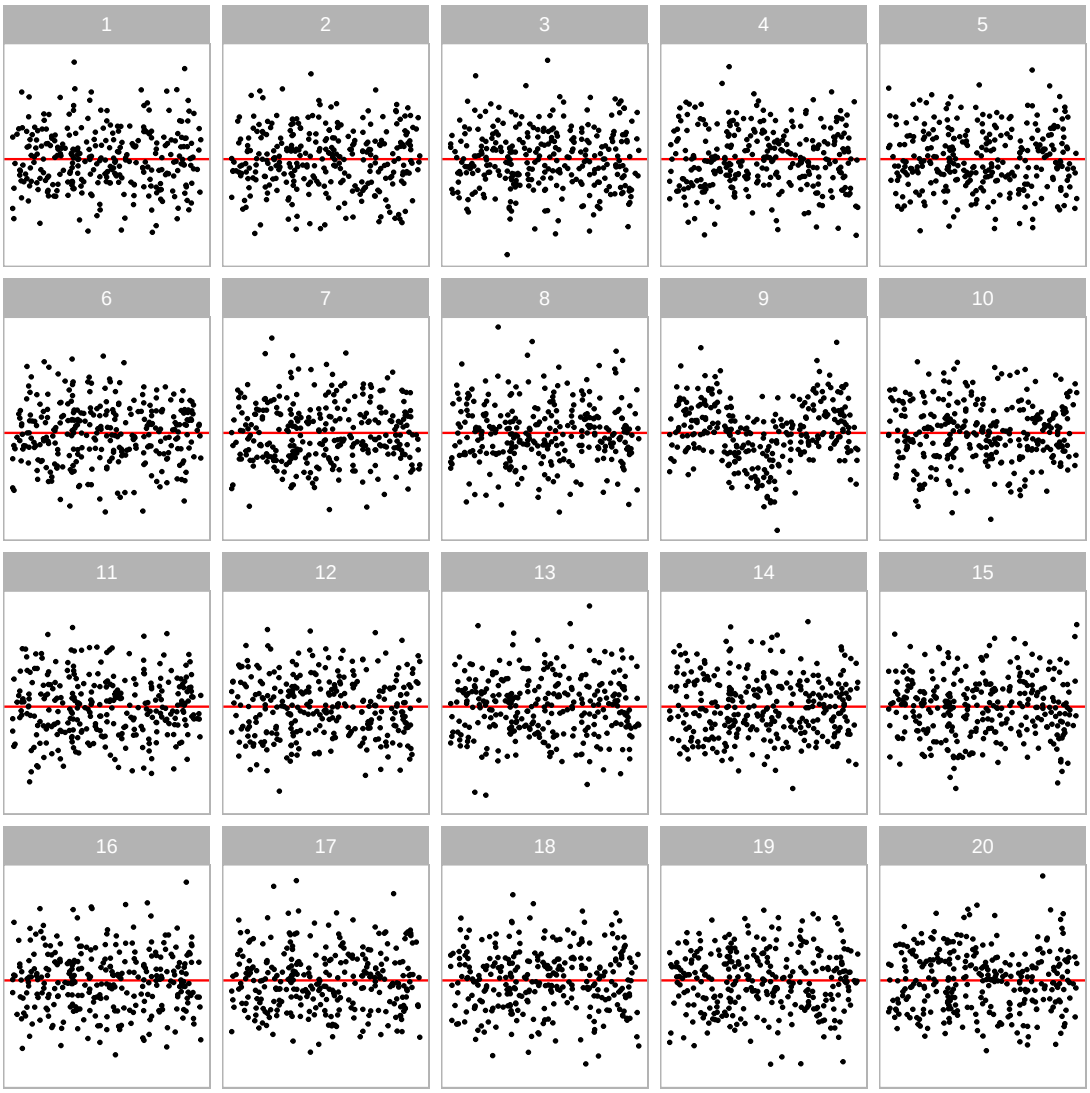

Figure 5. One of the lineups containing non-linearity patterns used in data collection period I. Can you spot the most different plot? The data plot is positioned at $2^3 + 1$.

setting:

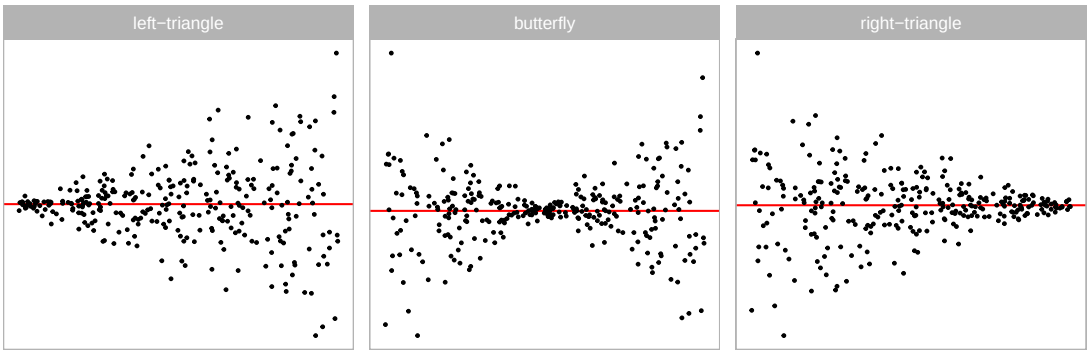

Figure 6. Heteroskedasticity forms used in the experiment. Three different shapes ($a = -1, 0, 1$) are used in the experiment to create left-triangle, "butterfly" and "right-triangle" shapes, respectively.

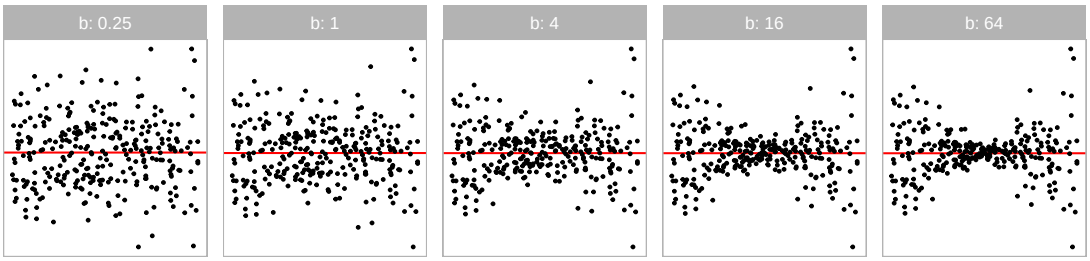

Figure 7. Five different values of b are used in heteroskedasticity simulation to control the strength of the signal. Larger values of b yield a bigger difference in variation, and stus stronger heteroskedasticity signal.

$$\boldsymbol{y} = 1 + \boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{13}$$

$$\boldsymbol{x} = g(\boldsymbol{x}_{raw}, 1), \tag{14}$$

$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, 1 + (2 - |a|)(\boldsymbol{x} - a)^2 b\boldsymbol{I}), \tag{15}$$

where $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{\varepsilon}$ are vectors of size n and $g(.)$ is the scaling function defined in Equation 11.

The null regression model used to fit the realizations generated by the above model is formulated exactly the same as Equation 12.

For $b \neq 0$, the variance-covariance matrix of the error term $\boldsymbol{\varepsilon}$ is correlated with the predictor $\boldsymbol{x}$, 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 6. The term $(2 - |a|)$ maintains the magnitude of residuals across different values of a.

An example lineup of this model used in data collection period II is shown in Figure 8 with $a = -1$. The data plot location is 2^4+2. Nine out of 11 subjects correctly identify the data plot from this lineup.

4.1.3. Factors common to both data collection periods

Fitted values are a function of the independent variables, and the distribution of the observed values affects the distribution of the fitted values. In the best case scenario the

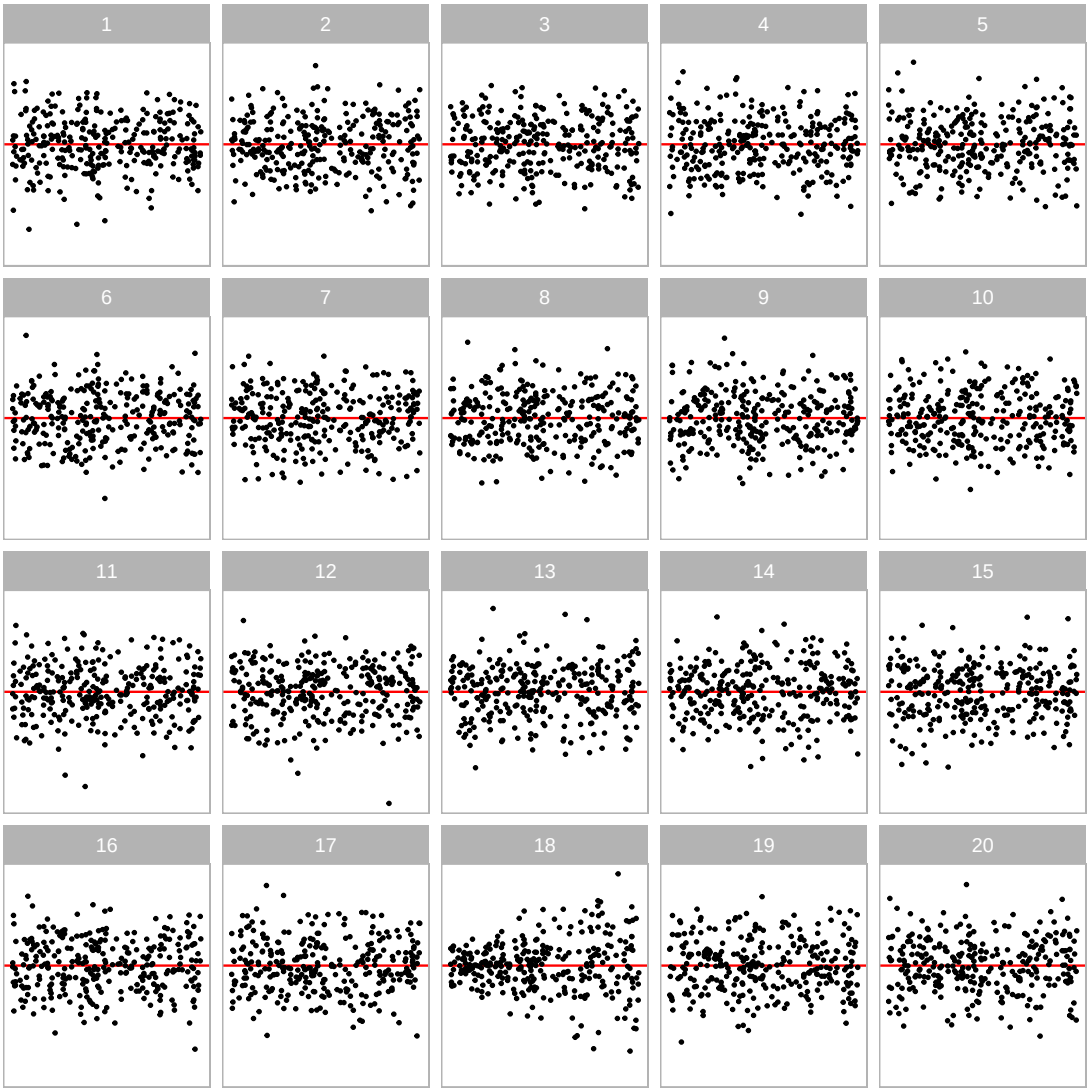

Figure 8. One of the lineups containing heteroskedasticity pattern used in data collection period II. Can you spot the most different plot? The data plot is positioned at $3^3 - 3^2$

fitted values will have a uniform distribution, which means that there is even coverage of possible observed values across all of the predictors. This is not always present in the collected data. Sometimes the fitted values are discrete because one or more predictors were measured discretely. The distribution may be relatively Gaussian, reflecting a linear combinatio of many predictors, adhering to the Central Limit Theorem. It is also common to see a skewed distribution of fitted values, if one or more of the predictors has a skewed distribution. This latter problem is usually corrected before modelling using a variable transformation. Our simulation assess this by using four different distributions to represent fitted values: (1) uniform, (2) normal, (3) skewed and (4) discrete. This is constructed by defining the raw predictor X_{raw} in four corresponding distributions: (1) $U(-1, 1)$, (2) $N(0, 0.3^2)$, (3) $lognormal(0, 0.6^2)/3$ and (4) $u\{1, 5\}$. We would expect that the best reading of residual plots occurs when the fitted values are uniformly distributed.

Three different sample sizes are used, $n = 50, 100, 300$ across the experiments. We would expect considerable variation in the signal strength in the simulated data plots

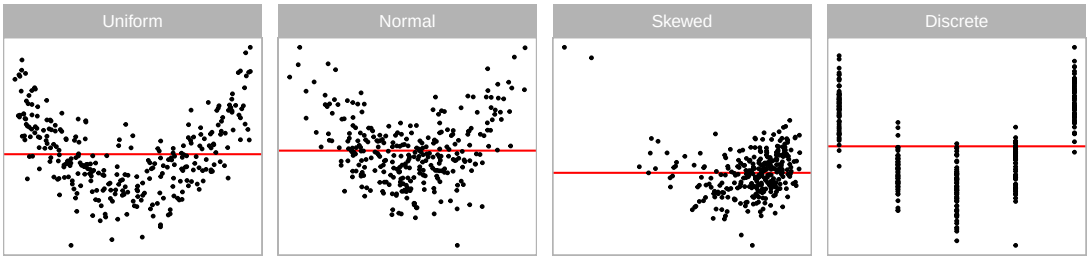

Figure 9. Variations in fitted values, that might affect perception of residual plots. Four different distributions are used.

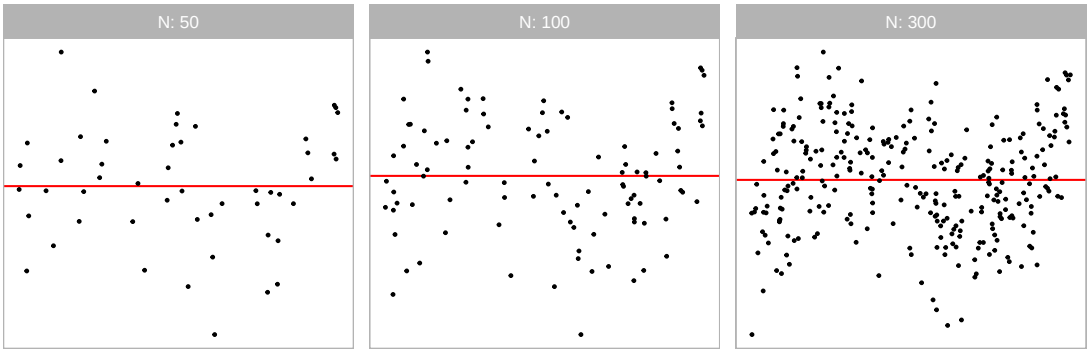

Figure 10. Examining the effect of signal strength for the three different values of n used in the experiment, for non-linear structure with fixed $\sigma = 1.5$, uniform fitted value distribution, and "S" shape. For these factor levels, only when $n = 300$ is the "S" shape clearly visible.

with smaller n. A sample size of 300 is typically enough for structure to be visible in a scatter plot reliably.

4.2. *Experimental setup*

4.2.1. 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 data collection periods I and II respectively. Figure 4 and 7 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 predictor. 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 data collection periods. It can be observed from Figure 10 that with fewer data points drawn in a residual plot, the visual pattern is more difficult to be detected.

4.2.2. Effect size

Effect size in statistics measures the strength of the signal relative to the noise. It is surprisingly difficult to quantify in general, even for simulated data as used in this

experiment.

For the non-linearity model, the key items defining effect size are sample size (n) and variance of the error term (σ^2), and so effect size would be roughly calculated as $\sqrt{n}/\sigma$. As sample size increases the effect size would increase, but as variance increases the effect size decreases. However, it is not clear how the additional parameter for the model polynomial order, k, should be incorporated. Intuitively, the large k means more complex pattern, which likely means effect size would decrease. For the purposes of our calculations we have chosen to use an approach based on Kullback-Leibler divergence (Kullback and Leibler 1951), coupled with simulation. This formulation defines effect size to be:

$$E = \frac{1}{2}\left(\boldsymbol{\mu}_z'(diag(\boldsymbol{R}\sigma^2))^{-1}\boldsymbol{\mu}_z\right)$$

where $diag(.)$ is the diagonal matrix constructed from the diagonal elements of a matrix, $\boldsymbol{R} = \boldsymbol{I}_n - \boldsymbol{H}$ is the residual operator, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the hat matrix, $\boldsymbol{\mu}_z = \boldsymbol{R}\boldsymbol{Z}\boldsymbol{\beta}_z$ is the expected values of residuals with $\boldsymbol{Z}$ be any higher order terms of $\boldsymbol{X}$ leave out by the regression equation and $\boldsymbol{\beta}_z$ be the corresponding coefficients, and $\sigma^2\boldsymbol{I}$ is the assumed covariance matrix of the error term when H_0 is true.

In the heteroskedasticity model, the key elements for measuring effect size are sample size, n, and the ratio of the biggest variance to smallest variance, b. Larger values of both would produce higher effect size. However, it is not clear how to incorporate the additional shape parameter, a. Thus the same approach is used here, where the formula can be written as:

$$E = \frac{1}{2}\left(log\frac{|diag(\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}')|}{|diag(\boldsymbol{R})|} - n + tr(diag(\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}')^{-1}diag(\boldsymbol{R}))\right)$$

where $\boldsymbol{V}$ is the actual covariance matrix of the error term. (Derivations for these equations are provided in the Appendix.)

To compute the effect size for each lineup we simulate a sufficient large number of samples from the same model, in each sample, the number of observations n is fixed. We then compute the effect size for each sample and take the average as the final value. This ensures lineups constructed with the same experimental factors will share the same effect size.

4.2.3. Subject allocation

As shown in Table 2, there are a total of $4\times4\times3\times4 = 192$ and $3\times5\times3\times4 = 180$ number of combinations of parameter values for non-linearity model and heteroskedasticity model respectively. Three replications are made for each of the combination results in $192 \times 3 = 576$ and $180 \times 3 = 540$ lineups. In addition, each lineup is designed to be evaluated by five different subjects. After attempting some pilot studies internally, 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 data collection period I and II respectively.

As mentioned in Section 3.3, α used in Equation 1 needs to be estimated using null lineups. Three replications are made for $3 \times 4 = 12$ combinations of common

factors n and fitted value distribution, results in $12 \times 3 = 36$ lineups included in data collection period 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. This generation procedure differs from the canonical Rorschach lineup procedure, which requires that all 20 plots are generated directly from the null model. However, these lineups serve the same fundamental purpose: to assess the number of visually interesting plots generated under H_0.

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 value distribution, result in $(36 \times 20 + 279 \times 3 \times 6)/(20 - 2) = 133$ subjects recruited for data collection period III.

4.2.4. Collecting results

Subjects for all three data collection periods 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 4.1, while the data of the remaining 19 null plots are generated by the residual rotation technique discussed in Section 2.3.

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 experiments 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.

5. Results

Data collection used a total of 1152 lineups, and resulted in a total of 7974 evaluations from 443 participants. Roughly half corresponded to the two models, non-linearity and heteroskedasticiy, and the three collection periods had similar numbers in each. Each participant received two of the 24 attention check lineups which were used to filter results of participants who were clearly not making an honest effort (only 11 of 454).

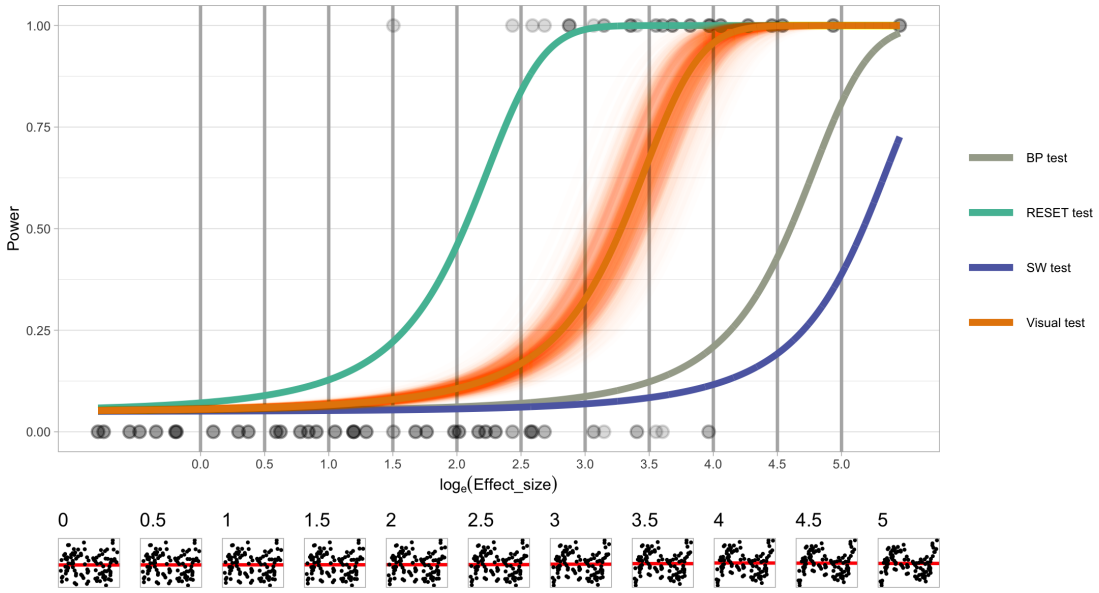

Figure 11. Comparison of power between different tests for non-linear patterns (uniform fitted values only). The power curves are estimated using logistic regression, and the horizontal lines of dots represent non-reject and reject results from visual tests for each lineup. The visual test has multiple 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.

To estimate α for calculating statistical significance (see Section 3.2) there were 720 evaluations of 36 null lineups. Neither the attention checks nor null lineups were used for the subsequent analysis. The de-identified data, `vi_survey`, is made available in the R package, `visage`.

The data was collected on lineups constructed from four different fitted value distributions: uniform, normal, skewed and discrete. More data was collected on the uniform (each evaluated by 11 participants) than the others (each evaluated by 5 participants). The analysis in Sections 5.1-5.4, uses only results from lineups generated with uniform fitted values, a total 3069 lineup evaluations. This was decided in order to compare the conventional and visual test performance for an optimal scenario. Section 5.5 examines how the results may be affected if the fitted value distribution was different.

5.1. *Power comparison of the tests*

Figures 11 and 12 present the power curves of various tests plotted against the effect size in the residuals, for non-linearity and heteroskedasticity, respectively. In each case the power of visual test is re-calculated for multiple bootstrap samples leading to the many (orange) curves. The effect size was computed at a 5% significance level and plotted on a natural logarithmic scale. To facilitate visual calibration of effect size values with the corresponding diagnostic plots, a sequence of example residual plots with increasing effect sizes is provided at the bottom of these figures. These plots serve as a visual aid to help readers understand how different effect size values translate to changes in the diagnostic plots. The horizontal lines of dots at 0 and 1 represent the non-reject or reject made by visual tests for each lineup.

Figures 11 compares the power for the different tests for non-linear structure in the residuals. The test with the uniformly higher power is the RESET test, one that specifically tests for non-linearity. The power curves for the visual test are effectively

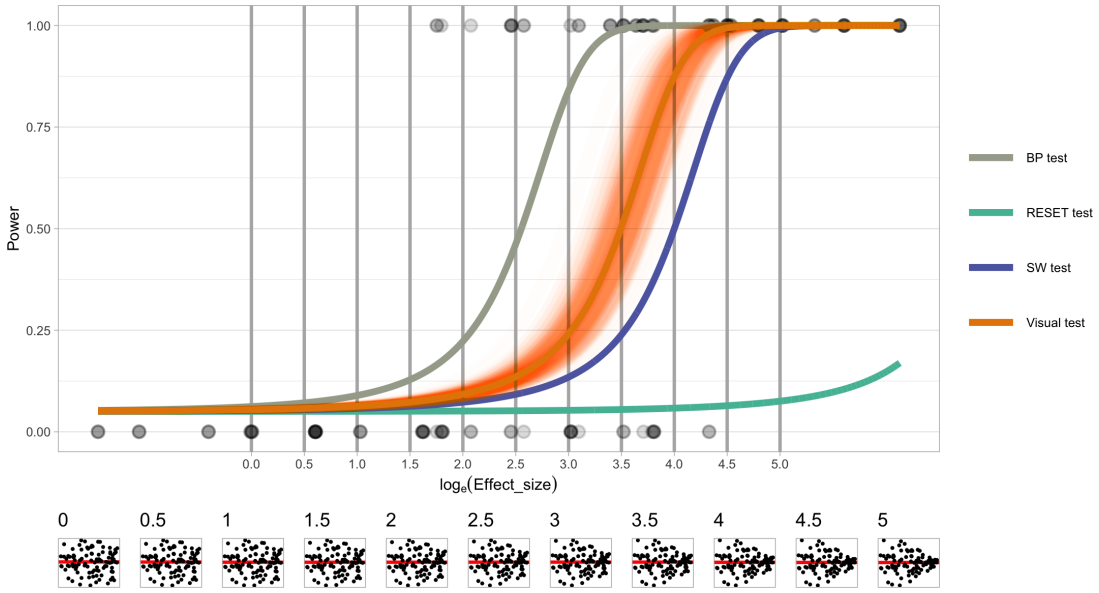

Figure 12. Comparison of power between different tests for heteroskedasticity patterns (uniform fitted values only). Main plot shows the power curves, with dots indicating non-reject and reject in visual testing of lineups. The multiple lines for the visual test arise from estimating the power on many 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.

a shift right from that of the RESET test. This means that the RESET test will reject 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 one scans 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 3.25. As expected the BP and SW tests have much lower power - they are not designed to detect non-linearity.

For the heteroskedasticity pattern, the power of BP test, designed for detecting heteroskedasticity, is uniformly higher than the other tests. The visual test power curve is a right shift. This shows a similar story to the power curves for non-linearity pattern - the conventional test is more sensitive than the visual test. From the example residual plots at the bottom we might argue that the heteroskedasticity becomes noticeably visible around an effect size of 3 or 3.5. However the BP test would reject at around effect size 2.5. Interestingly, the power curve for the SW test (for non-normality) is only sighly different to that of the visual test, suggesting that it performs reasonably for detecting heteroskedasticity, too. The power curve for the BP test suggests it is not useful for detecting heteroskeadsticity, as expected.

Overall, the results show that the conventional tests are more sensitive than the visual test. The conventional tests do have higher power for the patterns they are designed to detect, and are generally unable to detect other patterns. The visual test doesn't require specifying the pattern ahead of time, relying purely on whether the observed residual plot is detectably different from "good" residual plots. They will perform equally well regardless of the type of model defect. This aligns with the advice of experts on residual analysis, who consider residual plot analysis to be an indispensable tool for diagnosing model problems. What we gain from using a visual

test for this purpose is the removal of any subjective arguments about whether a pattern is visible or not. The lineup protocol provides the calibration for detecting patterns: that if the pattern in the data plot cannot be distinguished from patterns in good residual plots, then no discernible problem with the model exists.

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

The power comparison demonstrated that the appropriate conventional tests will reject more aggressively than visual tests, but we don't know how the decisions for each lineup would agree or disagree. Here we compare the reject or fail to reject decisions of these tests, across all the lineups. Figure 13 shows the agreement of the conventional and visual tests using a mosaic plot for both non-linearity patterns and heteroskedasticity patterns.

For both patterns the lineups resulting in a rejection by the visual test are *all* also rejected by the conventional test, except for one from the heteroskedasticity model. This reflects exactly the story from the previous section, that the conventional tests reject more aggressively than the visual test.

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. The visual test rejects 73% of the lineups that the conventional visual test reject.

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 BP test p-value = 0.056, is slightly above the significance cutoff of 0.05. This lineup was evaluated by 11 subjects, it has experimental factors $a = 0$ ("butterfly" shape), $b = 64$ (large variance ratio), $n = 50$ (small sample size), and a uniform distribution for the fitted values. It may have been the small sample size and the presence of a few outliers that may have resulted in the lack of detection by the conventional test.

5.3. *Effect of amount of non-linearity*

The order of the polynomial is a primary factor contributing to the pattern produced by the non-linearity model. Figure 15 explores the relationship between polynomial order and power of the tests. The conventional tests have higher power for lower orders of Hermite polynomials, and it drops substantially for the "triple-U" shape. 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 is constructed from the Hermite polynomials using 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 be closer to other power curves. This illustrates the sensitivity of the conventional test to the parameter choice, and highlights a limitation that it helps to know the data generating process to set the parameters for the test, which is unrealis-

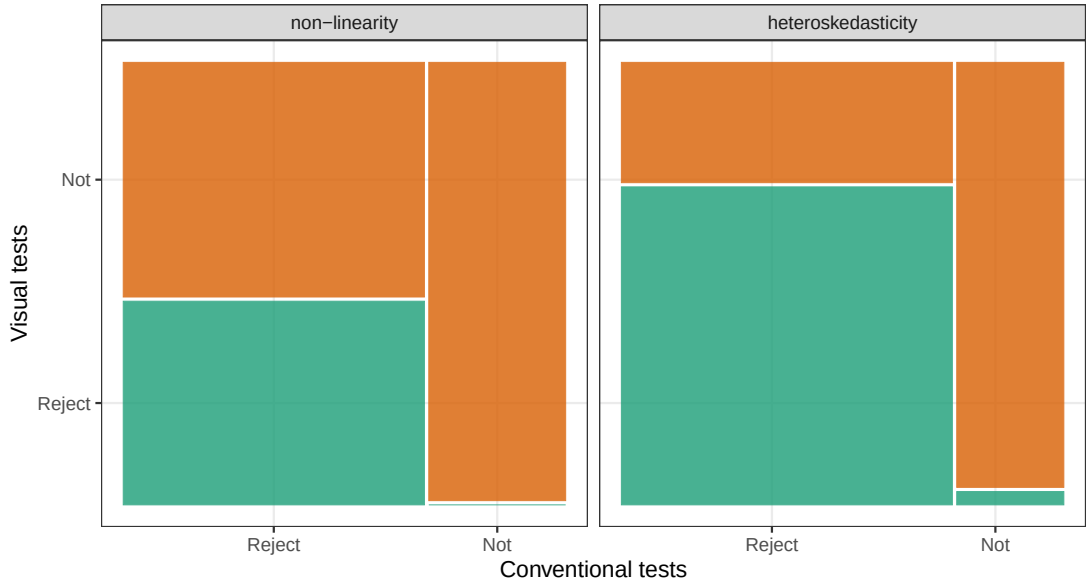

Figure 13. Rejection rate (p-value ≤ 0.05) of visual test conditional on the conventional test decision on non-linearity (left) and heteroskedasticity (right) lineups (uniform fitted values only) displayed using a mosaic plot. The visual test rejects less frequently than the conventional test, and (almost) only rejects when the conventional test does. Surprisingly, one lineup in the heteroskedasticity group is rejected by the visual test but NOT the conventional test.

tic. However, we examined this in more detail (see Appendix) and found that there is no harm for setting the parameter higher than four on the tests' operation for lower order polynomial shapes. Using a parameter value of six, instead of four, yields higher power regardless of generating process, and would be recommended.

For visual tests, we expect the "U" shape to be detected more readily, followed by the "S", "M" and "triple-U" shape. From Figure 15, it can be observed that the power curves mostly align with these expectations, except for the "M" shape, which is as easy to be detected as the "S" shape. This suggests a benefit of the visual test: knowing the shape ahead of time is *not* needed for its application.

5.4. *Effect of shape of heteroskedasticity*

Figure 16 examines the impact of the shape of the heteroskedasticity on the power of of both tests. The butterfly shape has higher power on both types of tests. The "left-triangle" and the "right-triangle" shapes are functionally identical, and this is observed for the conventional test, where the power curves are identical. Interestingly there is a difference for the visual test, where the power curve of the "left-triangle" shape is slightly higher than that of the "right-triangle" shape. This indicates a bias in perceiving heteroskedasticity depending on the direction. This would be worth investigating further.

5.5. *Effect of fitted value distributions*

In regression analysis, predictions are conditional on the observed values of the predictors, that is, the conditional mean of the dependent variable Y given the value of the independent variable X, $E(Y|X)$. This is an often forgotten element of regression

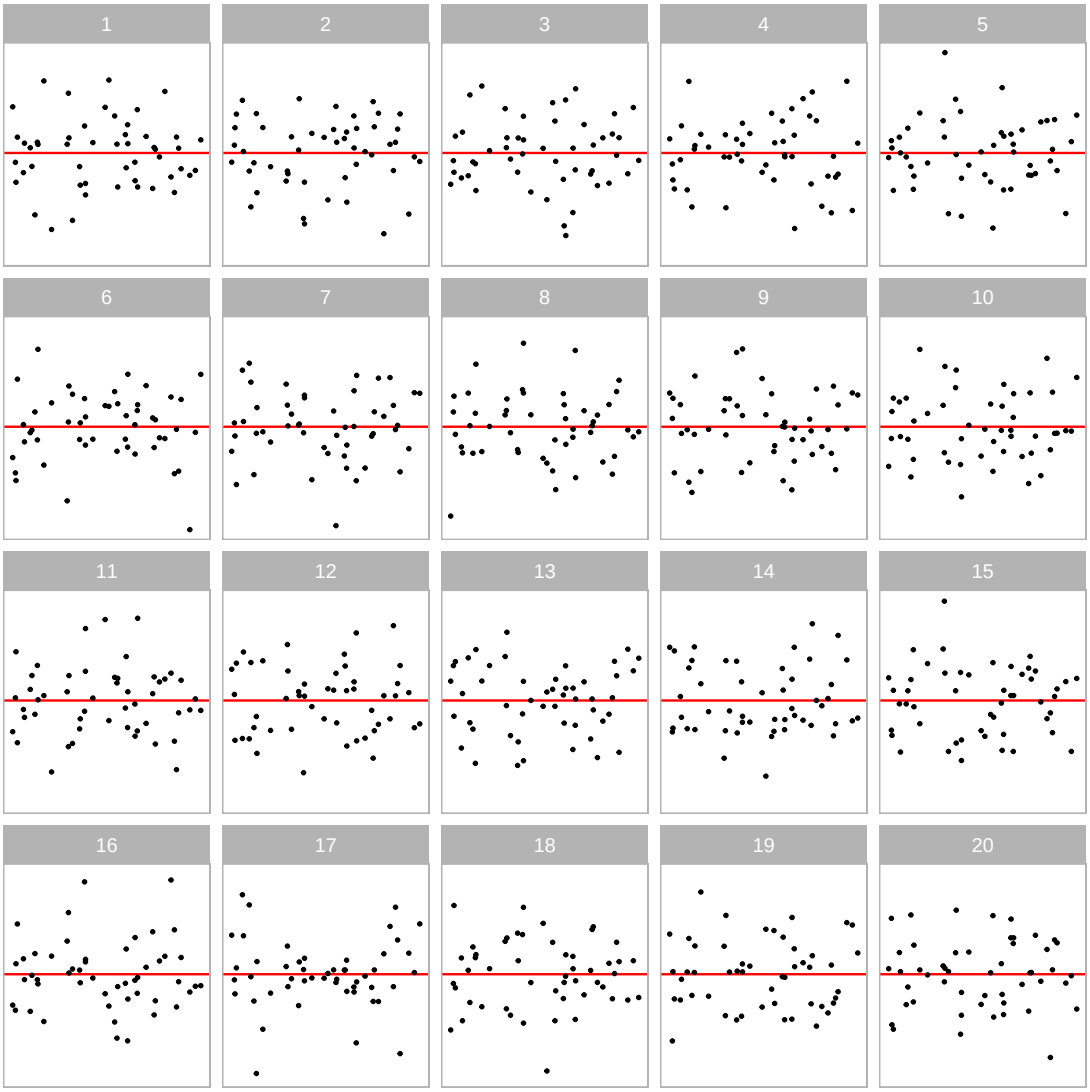

Figure 14. The single heteroskedasticity lineup that is rejected by the visual test but not by the BP test. The data plot (position 17) contains a "butterfly" shape. It has a log effect size of 3.76, and visibly displays heteroskedasticity, making it somewhat surprising that it is not detected by the BP test.

analysis but it is important. Where X is observed, particularly the distribution of the X values in the sample, or consequently $\hat{Y}$, may affect the ability to read any patterns in the residual plots. This experiment was constructed to assess this, based on four different distributions of fitted values: uniform, normal, discrete and lognormal (skewed). We would expect that if $\hat{Y}$ has a uniform distribution, this would make it easier to read the relationship with the residuals.

Figure 17 examines the impact of the fitted value distribution on the power of conventional (top row) and visual (bottom row) tests for both the non-linearity (left column) and heteroskedasticity (right column) patterns. For conventional tests, only the power curves of appropriate tests which are RESET tests and BP tests for non-linearity pattern and heteroskedasticity pattern respectively are shown. For visual tests, note that, more evaluations on lineup with uniform fitted value distribution were collected, so to have a fair comparison, we randomly sample five from the total eleven evaluations to estimate the power curves. Re-sampling produces the multiple

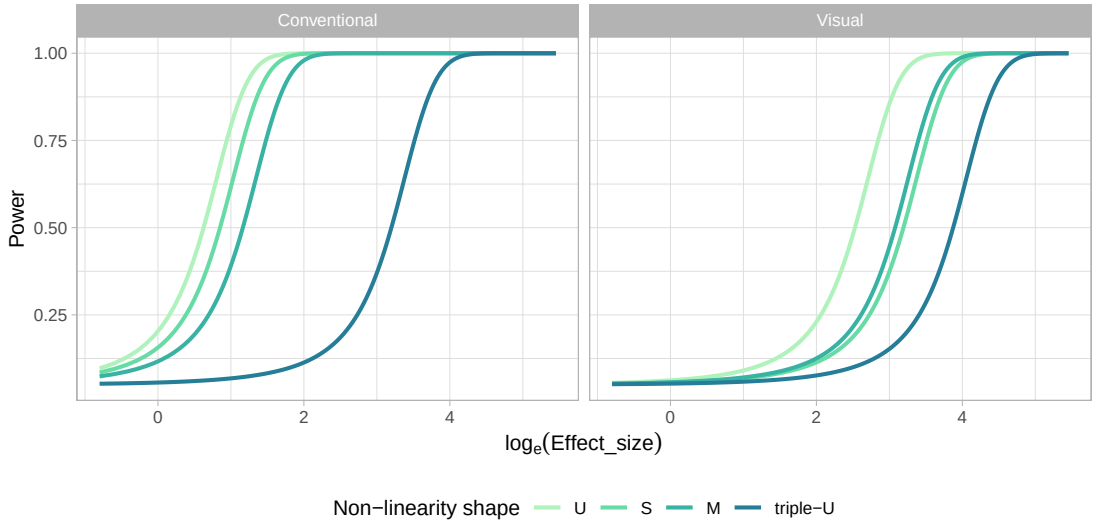

Figure 15. The effect of the order of the polynomial on the power of conventional and visual tests. Deeper colour indicates higher order. The default RESET tests under-performs significantly in detecting the "triple-U" shape. To achieve a similar power as other shapes, a higher order polynomial parameter needs to be used for the RESET test, but this higher than the recommended value.

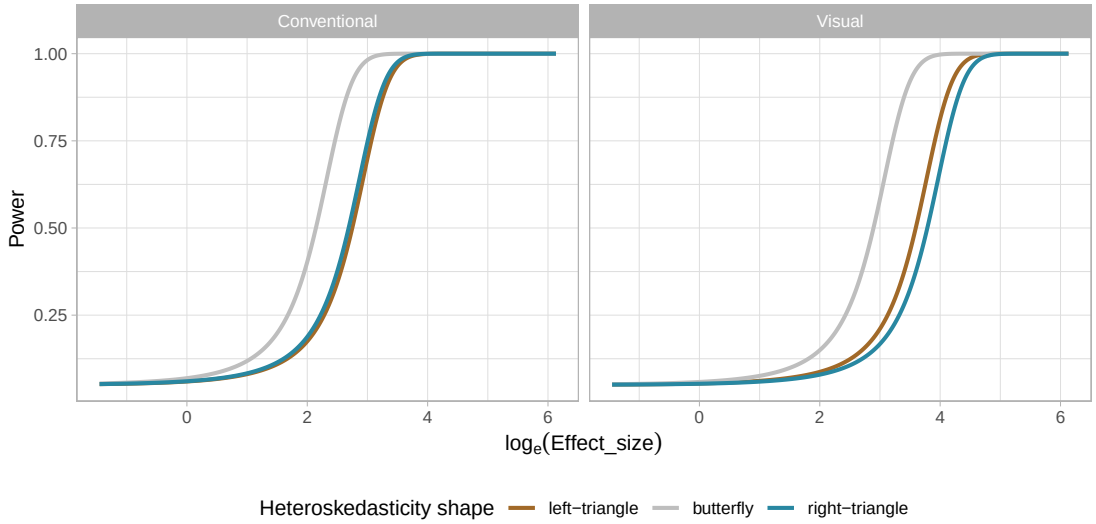

Figure 16. The effect of heteroskedasticity shape (parameter a) on the power of conventional and visual tests. The butterfly has higher power in both tests. Curiously, the visual test has slightly higher power for the "left-triangle" than the "right-triangle" shape, when it would be expected that they should be identical, which is observed in conventional testing.

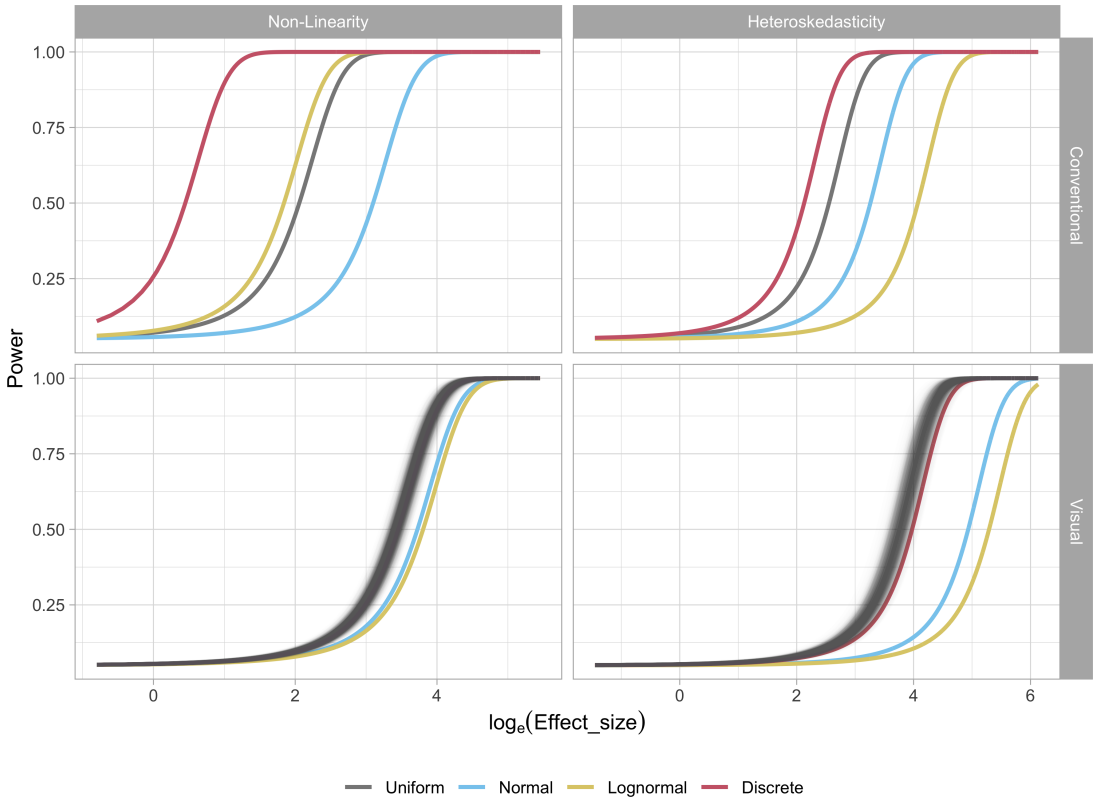

Figure 17. Comparison of power on lineups with different fitted value distributions for conventional and visual tests for non-linearity and heteroskedasticity patterns. The power curves of conventional tests for non-linearity and heteroskedasticity patterns are produced by RESET tests and BP tests, respectively. Power curves of visual tests are estimated using five evaluations on each lineup. For lineups with a uniform fitted value distribution, the five evaluations are repeatedly sampled from the total eleven evaluations to give multiple power curves (grey). Surprisingly, the fitted value distribution has produces more variability in the power of conventional tests than visual tests. Uneven distributions, normal and lognormal distributions, tend to yield lower power.

curves for the uniform, and provides an indication of the variation of the power.

Perhaps surprisingly, the visual tests have more consistent power across the different fitted value distributions. For the non-linear pattern, there is almost no power difference. For the heteroskedastic pattern, uniform and discrete have higher power than normal and lognormal. The likely reason is that these latter two have less observations in the tails where the heteroskedastic pattern needs to be detected.

The variation in power in the conventional tests is at first sight, shocking. However, it is discussed, albeit rarely, in the testing literature. See, for example, Jamshidian, Jennrich, and Liu (2007), Olvera Astivia, Gadermann, and Guhn (2019) and Zhang and Yuan (2018) which show derivations and use simulation to assess the effect of the observed distribution of the predictors on test power. The big differences in power curves that are seen in Figure 17 is echoed in the results reported in these articles.

6. Conclusions

Motivated by the advice of regression analysis experts, that residual plots as opposed to conventional tests are an indispensible methods for assessing model fit, a human subjects experiment was conducted using visual inference. The experiment tested two

primary departures from good residuals: non-linearity and heteroskedasticity.

The experiment found that conventional residual-based statistical tests are more sensitive to weak residual departures from model assumptions than visual tests as would be evaluated by humans. That is, conventional tests conclude there are problems with the model fit almost twice as often as humans would. They often reject when departures in the form of non-linearity and heteroskedasticity are not visible to a human.

One might say that this is correct behaviour, but it can be argued that the conventional tests are rejecting when it is not necessary. 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 indispensible method for model diagnostics.

Now it is important to note that residual plots need to be delivered as a lineup, where it is embedded in a field of null plots. A residual plot may contains 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 subjects with a comparison to null plots, results in a set of more accurate visual findings. This enables a careful calibration for reading structure in residual plots.

However, human evaluation of residuals is expensive. It is time-consuming, laborious and unfriendly to vision-impaired people. This is another reason why it often appears to be ignored. With the availability of sophisticated computer vision algorithms today, the goal of this work is to form the basis of providing automated residual plot reading. The findings suggest the strong demand of graphical inspection in regression diagnostics, so developing an automatic visual inference system to evaluate lineups of residual plots is valuable. We plan to build a completed open-source system in an `R` package and provide a web interface such as a website for public to interact with. Details about this system will be discussed in our next paper.

The experiment also revealed some interesting details about how residual plots are read. For the most part, the visual test performed very similarly to the appropriate conventional test only with the power curve shifted in the less sensitive direction. Unlike the conventional tests, where one needs to specifically test for non-linearity or heteroskedasticity the visual test operated effectively across the range of departures from good residuals.

As expected, if the fitted value distribution is not uniform, there is a loss of power in the visual test. Structure is hardest to detect if fitted values are lognormal. Also, complex structure are generally harder to detect, but there are outliers. Under the designed scenarios in this paper, we find the visual test to be a more robust test against the change of fitted value distributions. A surprising finding was that the direction of heteroskedasticity appears to affect the ability to visually detect it, with wedge to the right being less detectable.

Acknowledgements

These `R` packages are used for the work: `cli` (Csárdi 2022), `curl` (Ooms 2022), `dplyr` (Wickham et al. 2023), `ggplot2` (Wickham 2016), `jsonlite` (Ooms 2014), `lmtest` (Zeileis and Hothorn 2002), `mpoly` (Kahle 2013), `progress` (Csárdi and FitzJohn 2019), `tibble` (Müller and Wickham 2022), `ggmosaic` (Jeppson, Hofmann, and Cook

2021), `purrr` (Henry and Wickham 2022), `tidyr` (Wickham and Girlich 2022), `readr` (Wickham, Hester, and Bryan 2022), `stringr` (Wickham 2022), `here` (Müller 2020), `kableExtra` (Zhu 2021), `patchwork` (Pedersen 2022), `rcartocolor` (Nowosad 2018). The study website is powered by `PythonAnywhere` (PythonAnywhere LLP 2023) and the `Flask` web framework (Grinberg 2018). The `jsPsych` framework (De Leeuw 2015) is used to create behavioral experiments that run in our study website.

The article was created with R packages `rticles` (Allaire et al. 2022), `knitr` (Xie 2014) and `rmarkdown` (Xie, Dervieux, and Riederer 2020). The project's Github repository (`https://github.com/TengMCing/lineup_residual_diagnostics`) contains all materials required to reproduce this article.

Supplementary material

The supplementary material is available at `https://github.com/TengMCing/lineup_residual_diagnostics/blob/master/appendix.pdf`. It includes more details about the experimental setup, the derivation of the effect size, the effect of data collection period, and the estimate of α.

References

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

Allaire, JJ, Yihui Xie, Christophe Dervieux, R Foundation, Hadley Wickham, Journal of Statistical Software, Ramnath Vaidyanathan, et al. 2022. *rticles: Article Formats for R Markdown*. R package version 0.24, `https://CRAN.R-project.org/package=rticles`.

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.

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.

Csárdi, Gábor. 2022. *cli: Helpers for Developing Command Line Interfaces*. R package version 3.4.1, `https://CRAN.R-project.org/package=cli`.

Csárdi, Gábor, and Rich FitzJohn. 2019. *progress: Terminal Progress Bars*. R package version 1.2.2, `https://CRAN.R-project.org/package=progress`.

De Leeuw, Joshua R. 2015. "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser." *Behavior research methods* 47: 1–12.

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.

Grinberg, Miguel. 2018. *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.".

Henry, Lionel, and Hadley Wickham. 2022. *purrr: Functional Programming Tools*. R package version 0.3.5, `https://CRAN.R-project.org/package=purrr`.

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

Jamshidian, Mortaza, Robert I Jennrich, and Wei Liu. 2007. "A study of partial F tests for multiple linear regression models." *Computational statistics & data analysis* 51 (12): 6269–6284.

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.

Jeppson, Haley, Heike Hofmann, and Di Cook. 2021. *ggmosaic: Mosaic Plots in the 'ggplot2' Framework*. R package version 0.3.3, `https://CRAN.R-project.org/package=ggmosaic`.

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

Kullback, Solomon, and Richard A Leibler. 1951. "On information and sufficiency." *The annals of mathematical statistics* 22 (1): 79–86.

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

Loy, Adam. 2021. "Bringing visual inference to the classroom." *Journal of Statistics and Data Science Education* 29 (2): 171–182.

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*. John Wiley & Sons.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. R package version 1.0.1, `https://CRAN.R-project.org/package=here`.

Müller, Kirill, and Hadley Wickham. 2022. *tibble: Simple Data Frames*. R package version 3.1.8, `https://CRAN.R-project.org/package=tibble`.

Nowosad, Jakub. 2018. *'CARTOColors' Palettes*. R package version 1.0, `https://nowosad.github.io/rcartocolor`.

Olvera Astivia, Oscar L, Anne Gadermann, and Martin Guhn. 2019. "The relationship between statistical power and predictor distribution in multilevel logistic regression: a simulation-based approach." *BMC medical research methodology* 19 (1): 1–20.

Ooms, Jeroen. 2014. "The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects." *arXiv:1403.2805 [stat.CO]* `https://arxiv.org/abs/1403.2805`.

Ooms, Jeroen. 2022. *curl: A Modern and Flexible Web Client for R*. R package version 4.3.3, `https://CRAN.R-project.org/package=curl`.

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

Pedersen, Thomas Lin. 2022. *patchwork: The Composer of Plots*. R package version 1.1.2, `https://CRAN.R-project.org/package=patchwork`.

PythonAnywhere LLP. 2023. "PythonAnywhere." `https://www.pythonanywhere.com`.

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.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (4): 817–838.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. `https://ggplot2.tidyverse.org`.

Wickham, Hadley. 2022. *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.1, `https://CRAN.R-project.org/package=stringr`.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *dplyr: A Grammar of Data Manipulation*. R package version 1.1.0, `https://CRAN.R-project.org/package=dplyr`.

Wickham, Hadley, and Maximilian Girlich. 2022. *tidyr: Tidy Messy Data*. R package version 1.2.1, `https://CRAN.R-project.org/package=tidyr`.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2022. *readr: Read Rectangular Text Data*. R package version 2.1.3, `https://CRAN.R-project.org/package=readr`.

Xie, Yihui. 2014. "knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman and Hall/CRC. ISBN 978-1466561595, `http://www.crcpress.com/product/isbn/9781466561595`.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman and Hall/CRC. ISBN 9780367563837, `https://bookdown.org/yihui/rmarkdown-cookbook`.

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10.

Zhang, Zhiyong, and Ke-Hai Yuan. 2018. *Practical statistical power analysis using Webpower and R*. Isdsa Press.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.3.4, `https://CRAN.R-project.org/package=kableExtra`.