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], 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 April 3, 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, residuals from the model fit are commonly used. 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 may 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 to be well behaved. However, if the distribution is skewed, bimodal, multimodal, or contains outliers, there is a 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 can be a painstakingly difficult task for humans. It is especially common, particularly among students, to misinterpret patterns that are random noise and random deviation from a model (Loy 2021). It is also possible to conduct hypothesis tests for non-linear

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

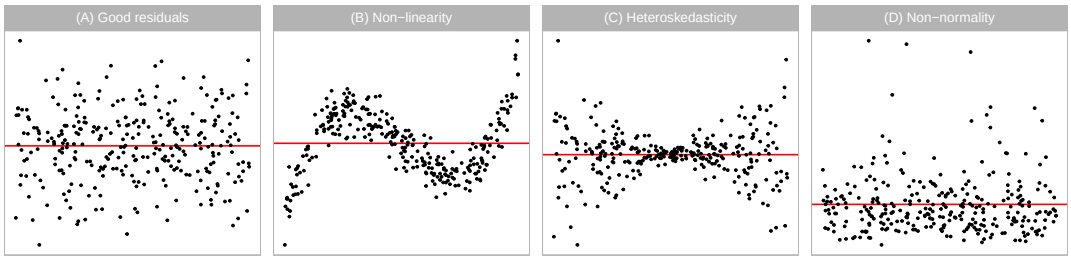

Figure 1. Example residual vs fitted value plots: (A) classically good looking residuals, (B) non-linear pattern indicates that the model has not captured a non-linear association, (C) heteroskedasticity indicating that variance around the fitted model is not uniform, and (D) non-normality where the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

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, e.g. Draper and Smith (1998), Montgomery and Peck (1982), Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1999) and Cook and Weisberg (1982). All these writings consider plotting residuals to be a standard technique that should be examined routinely in all regression modelling problems. In addition, Draper and Smith (1998) and Belsley, Kuh, and Welsch (1980) believe 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, as they are easier to use. Montgomery and Peck (1982) even suggests that residual plots are more informative in most practical situations than the corresponding formal tests, and statistical tests on regression model residuals are not widely used based on their experience.

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 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 future work, in particular, how the responsibility for residual plot reading might be relieved.

2. Background

2.1. *Departures from good residual plots*

Graphical summaries in which residuals are plotted against fitted values or other functions of the predictors that are approximately orthogonal to the residuals are referred to as standard residual plots in Cook and Weisberg (1982). Figure 1A shows an ideal residual plot where the residuals are 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, heteroskedasticity 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 predictors 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 shown in Figure 1B. 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 Figure 1C, 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 Figure 1D. 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 Figure 1D.

2.2. *Conventionally testing for departures*

Other than checking diagnostic plots, analysts may perform formal hypothesis testing to detect model defects. A variety of tests can be applied; each testing for a specific violation of the null hypothesis. 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. 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 Shapiro-Wilk 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.

We apply the RESET test, Breusch-Pagan test and Shapiro-Wilk test to the residual plots shown in Figure 1; the results of this is shown in Table 1. The Breusch-Pagan test and the Shapiro-Wilk test both reject H_0 for residual plots that display non-linearity and heteroscedasticity, even though these are not the original intention of the test. 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

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*

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 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*

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, that the model is correctly specified. 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. 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

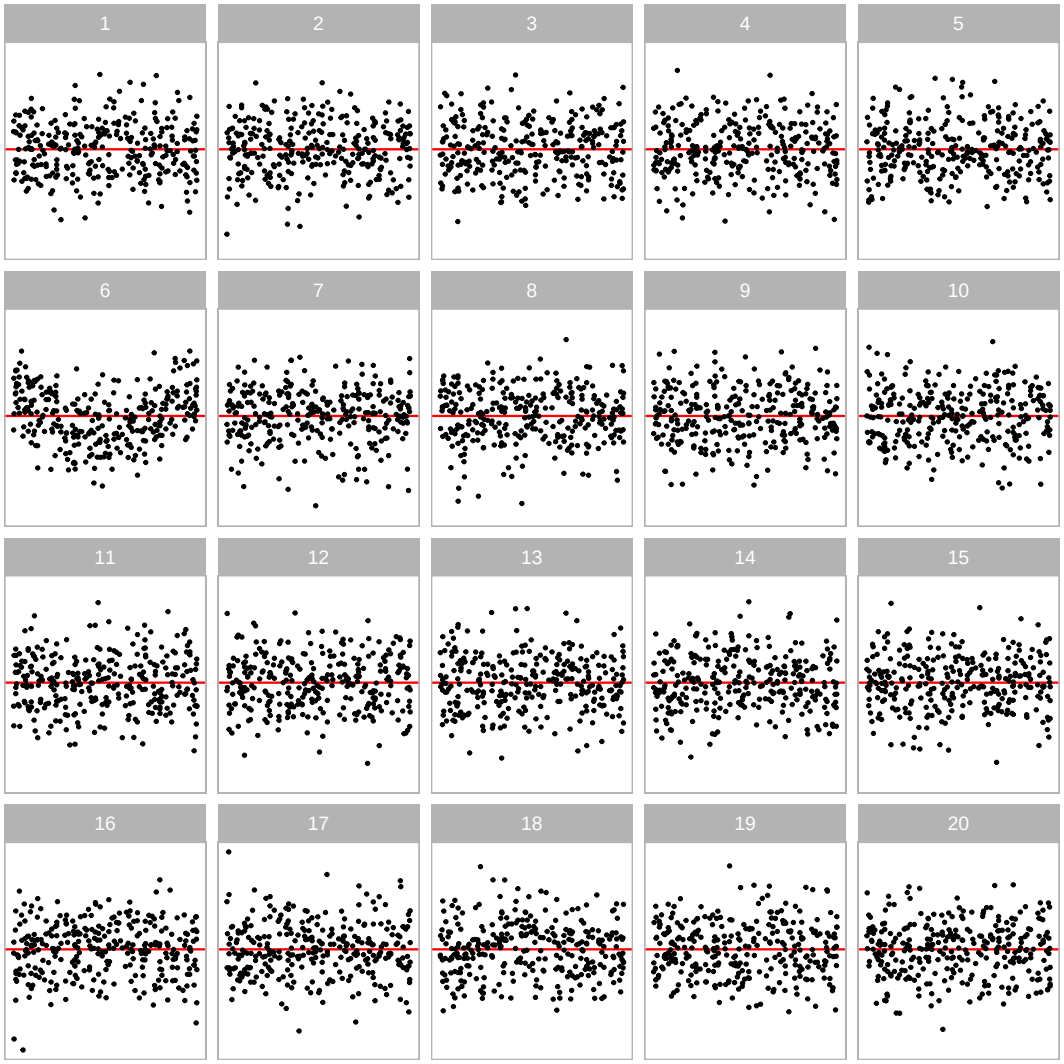

Figure 2. Visual testing is conducted using a lineup, as in the example here. The residual plot computed from the observed data (plot $2^2 + 2$, exhibiting non-linearity) is embedded among 19 null plots, where the residuals are simulated from a standard error model. Computing the p-value requires that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

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 H_0 : *The regression model is correctly specified* against the broad alternative H_a : *The regression model is misspecified.*

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 $y = \tau_0 + \sum_{i=1}^{p} \tau_p x_p + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + \gamma_3 \hat{y}^4 + u,\ \ u \sim N(0, \sigma_u^2)$. Similarly, the Breusch-Pagan 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 should be 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 mix-effect model. However, this mix-effect 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

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 uniform
			64.00		

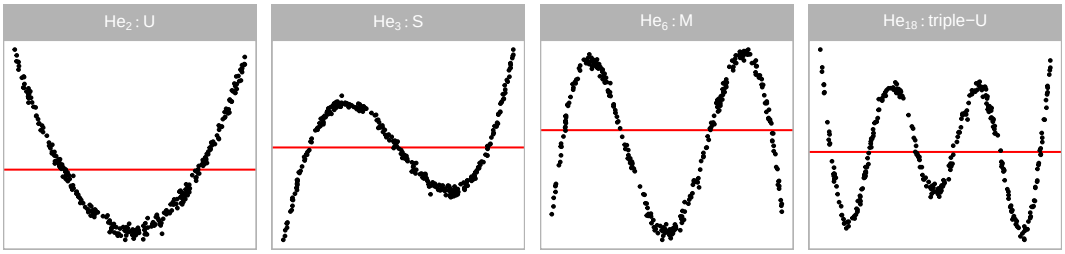

Figure 3. Polynomial forms generated for the residual plots used to assess detecting non-linearity. The four shapes are generated by varying the order of polynomial given by j in $He_j(.)$.

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

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.

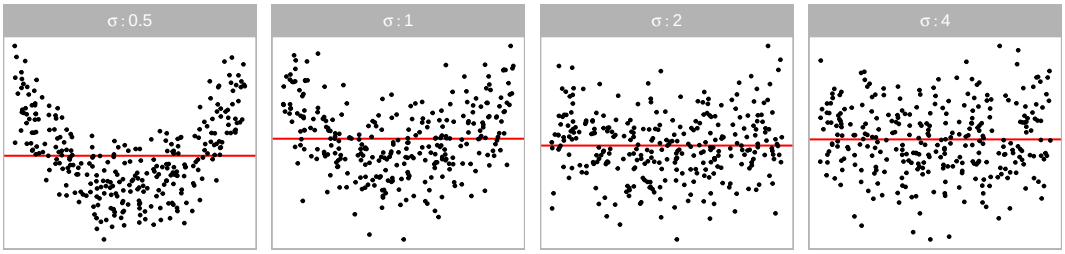

Figure 4. Examining the effect of σ on the signal strength in the non-linearity detection, for $n = 300$, uniform fitted value distribution and the "U" shape. As σ increases the signal strength decreases, to the point that the "U" is almost unrecognisable when $\sigma = 4$.

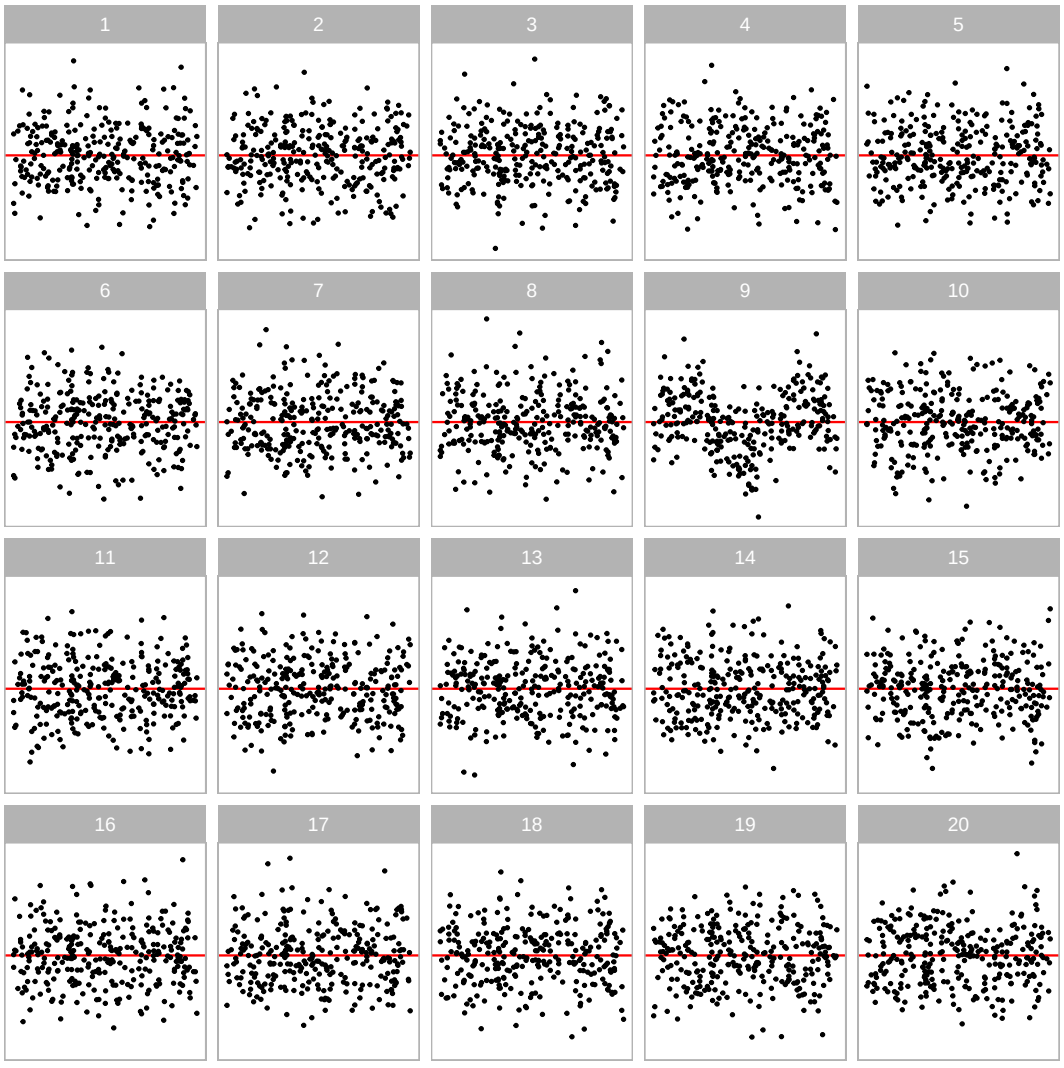

Figure 5. One of the lineups containing non-linearity patterns used in data collection period I. Can you spot the most different plot? The data plot is positioned at $2^3 + 1$.

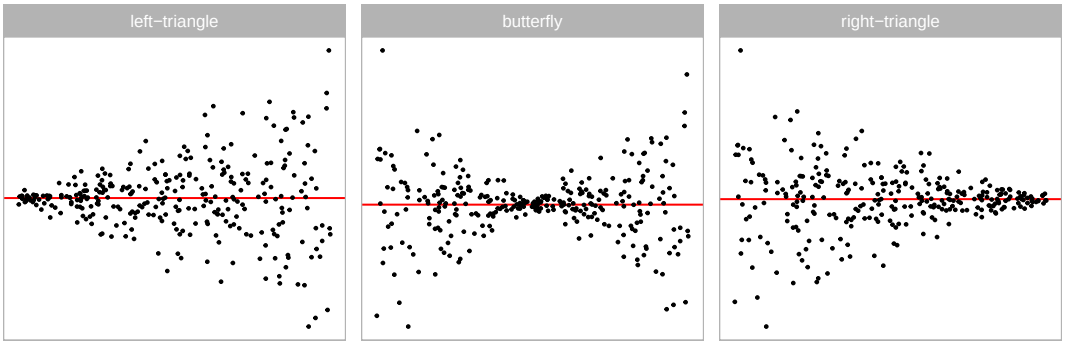

Figure 6. Heteroskedasticity forms used in the experiment. Three different shapes ($a = -1, 0, 1$) are used in the experiment to create left-triangle, "butterfly" and "right-triangle" shapes, respectively.

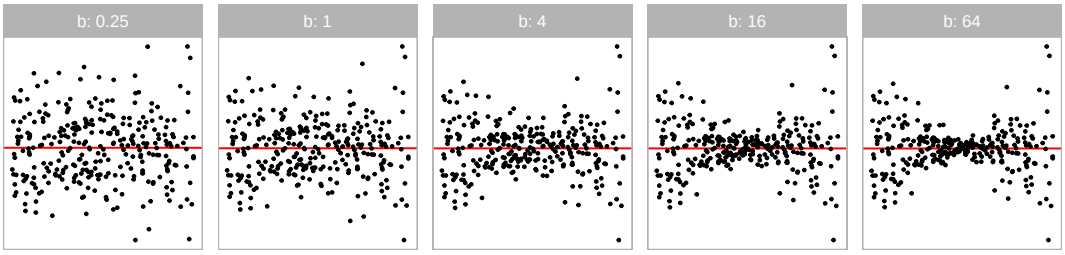

Figure 7. Five different values of b are used in heteroskedasticity simulation to control the strength of the signal. Larger values of b yield a bigger difference in variation, and stus stronger heteroskedasticity signal.

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 setting:

$$\boldsymbol{y} = 1 + \boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{13}$$
$$\boldsymbol{x} = g(\boldsymbol{x}_{raw}, 1), \tag{14}$$
$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, 1 + (2 - |a|)(\boldsymbol{x} - a)^2 b\boldsymbol{I}), \tag{15}$$

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.

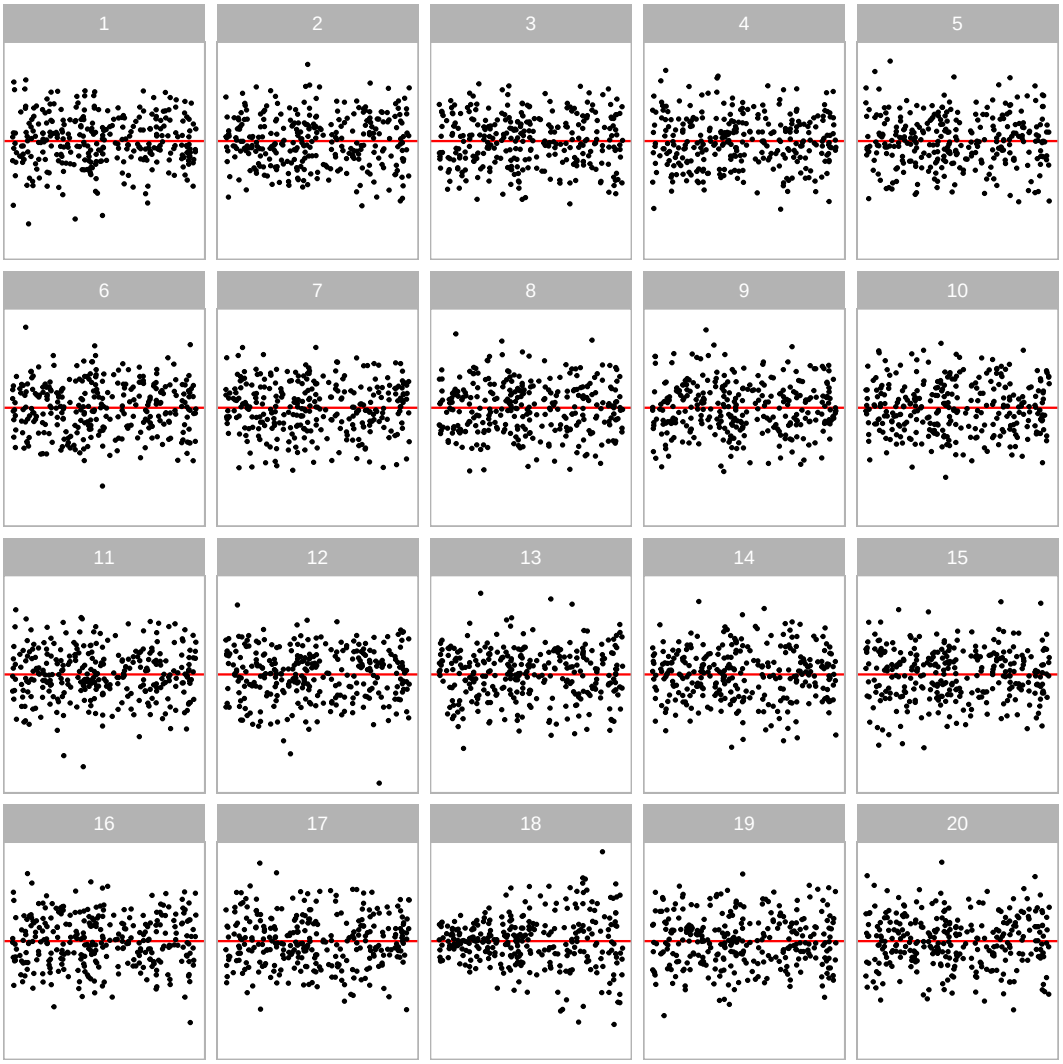

Figure 8. One of the lineups containing heteroskedasticity pattern used in data collection period II. Can you spot the most different plot? The data plot is positioned at $3^3 - 3^2$

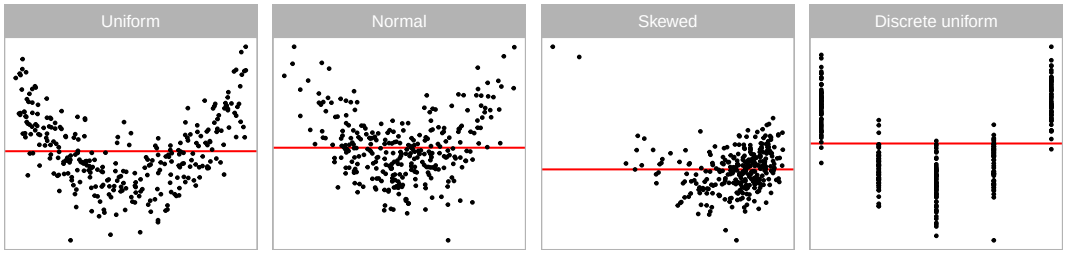

Figure 9. Variations in fitted values, that might affect perception of residual plots. Four different distributions are used.

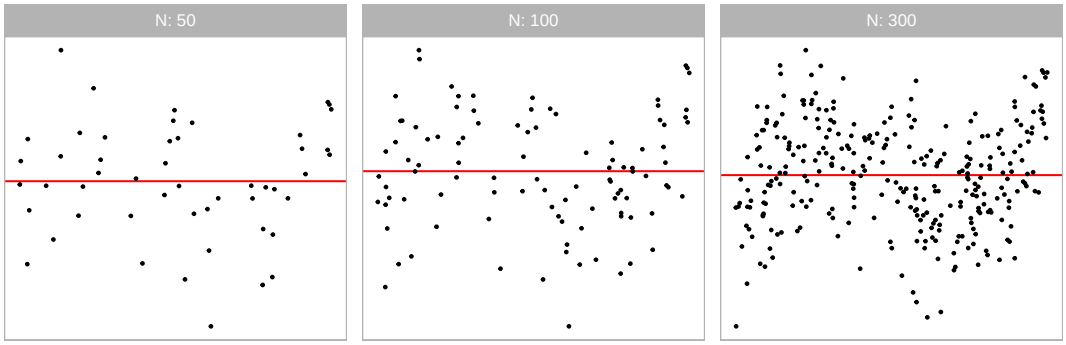

Figure 10. Examining the effect of signal strength for the three different values of n used in the experiment, for non-linear structure with fixed $\sigma = 1.5$, uniform fitted value distribution, and "S" shape. For these factor levels, only when $n = 300$ is the "S" shape clearly visible.

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 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 uniform. 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 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:

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

where $diag(.)$ is the diagonal matrix constructed from the diagonal elements of a matrix, $\boldsymbol{R}_a = \boldsymbol{I}_n - \boldsymbol{H}_a$ is the residual operator, $\boldsymbol{H}_a = \boldsymbol{X}_a(\boldsymbol{X}_a'\boldsymbol{X}_a)^{-1}\boldsymbol{X}_a'$ is the hat matrix, $\boldsymbol{\mu}_z = \boldsymbol{R}_a\boldsymbol{Z}\boldsymbol{\beta}_z$ 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:

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

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

5.1. *Overview*

There are 2880, 2700 and 1674 lineups evaluation made by 160, 150 and 133 subjects recruited for data collection periods 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 a more reliable analysis. Analysis of lineups with other fitted value distributions can be found in Section 5.6.

5.2. *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.

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 RE-

SET 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 predictors 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 predictors in the Breusch-Pagan test be the same as the White test. Thus, we include both x and 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 7% to 84% as $log_e(E)$ increases from 0 to 2.5, while power of other tests respond inactively to the change of effect, showing that RESET test is much more sensitive to weak non-linear structure. Meanwhile, no noticeable residual departures can be spotted from the example residual plots.

In terms of heteroskedasticity patterns, the power of Breusch-Pagan test is constantly greater than the power of visual test. At $log_e(E) \approx 2.5$, where the power curve of the visual test remains at a low level, the Breusch-Pagan test has around 50% chance of rejecting H_0. Similarly, the visual feature is nearly unobservable from the example residual plots at this level of effect size.

The power of visual test arises steadily as $log_e(E)$ increases from 3 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 $log_e(E) = 3.5$, a weak "S-shape" and a weak "right-triangle" shape are presented in Figure 11 and Figure 12 respectively. The visual pattern becomes clearer as $log_e(E)$ increases. At $log_e(E) = 5$, the power of visual tests for both patterns reaches almost 100%.

Power curves of inappropriate tests show 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 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 literally 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 to weak residual departures 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 to analysts and decision makers. In contrast, if the strict correctness of the model assumption is of particular interest, conducting a conventional test is still the recommended choice based on the findings.

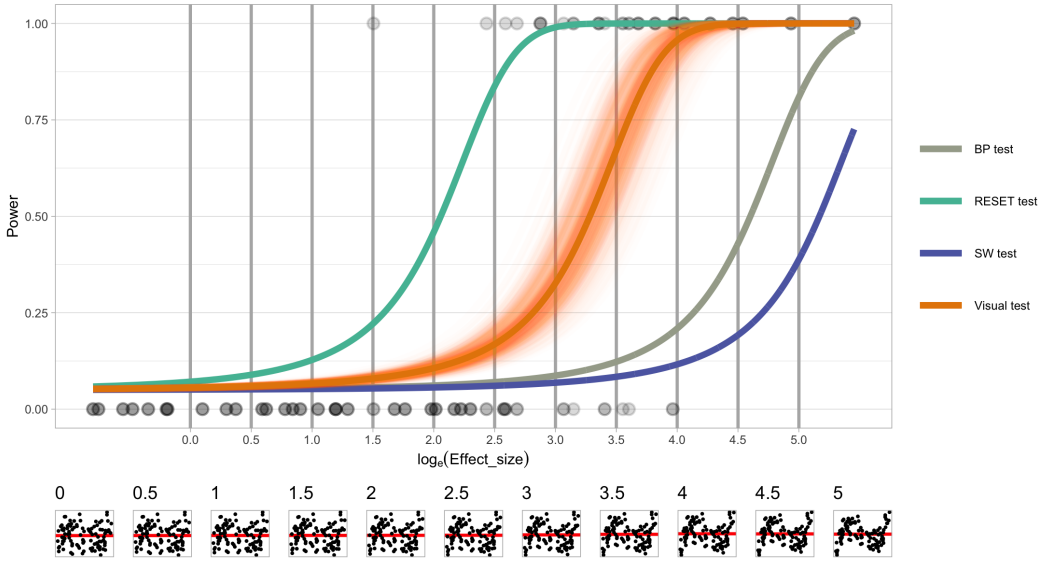

Figure 11. Comparison of power between different tests for non-linear patterns (uniform fitted values only). Main plot shows the power curves estimated using logistic regression, with dots indicating human evaluations of lineups. Surrounding lines of the visual test show the estimated power of 500 bootstrap samples. Small row of plots shows typical residual plots corresponding to specific effect sizes, marked by vertical 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 2, but for the visual test it is around 3.5.

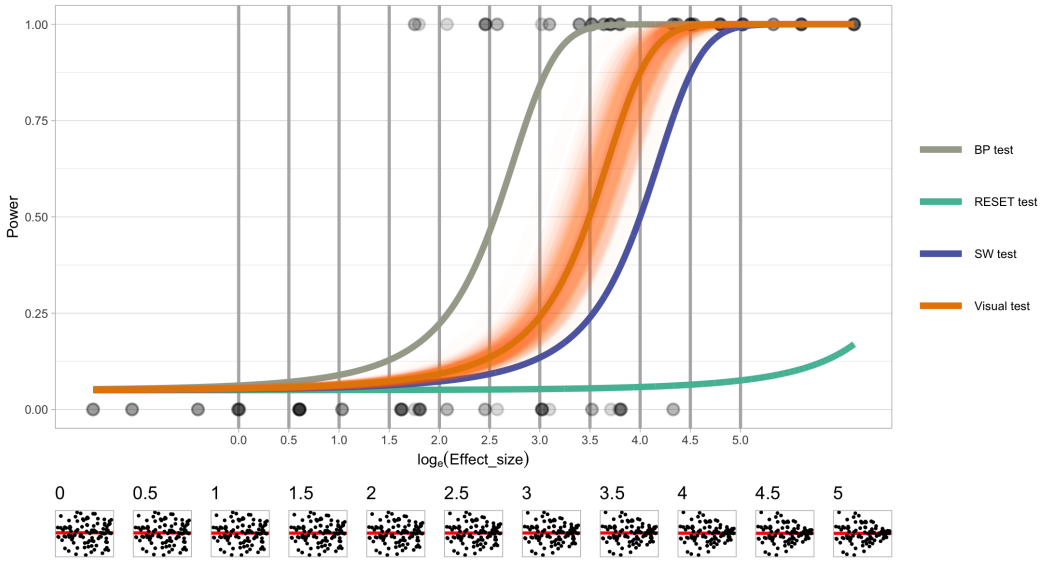

Figure 12. Comparison of power between different tests for heteroskedasticity patterns (uniform fitted values only). Main plot shows the power curves, with dots indicating human evaluations of lineups. Surrounding lines of the visual test show the estimated power of 500 bootstrap samples. Small row of plots shows typical residual plots corresponding to specific effect sizes, marked by vertical 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 2.5, but for the visual test it is around 3.5.

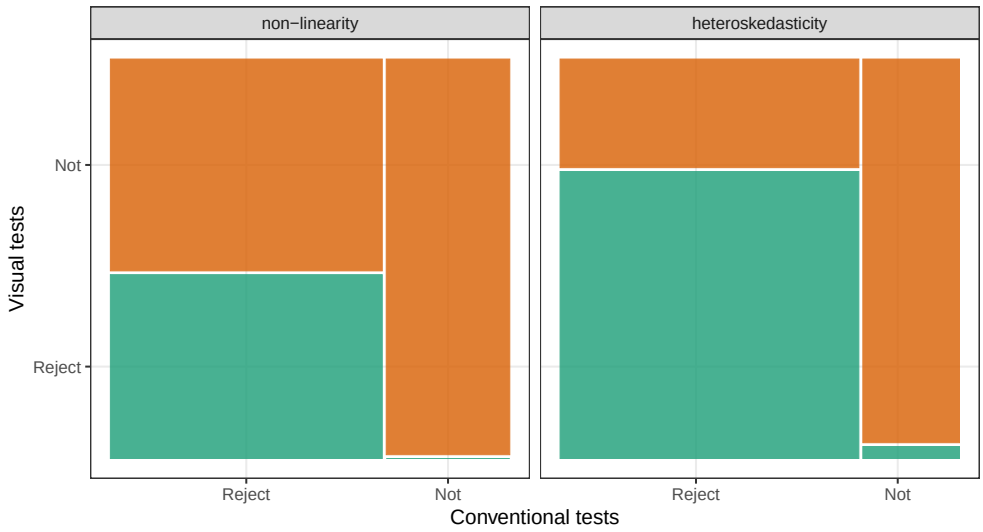

Figure 13. Rejection rate (p-value ≤ 0.05) of visual test conditional on the conventional test decision on non-linearity (left) and heteroskedasticity (right) lineups (uniform fitted values only) displayed using a mosaic plot. The visual test rejects less frequently than the conventional test. We would generally expect the visual test to only reject when the conventional test does. Surprisingly, one lineup containing a heteroskedasticity pattern does not follow this rule.

5.3. *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.

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 $= 0.056$, 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 must be the small sample size that may have resulted in the lack of detection.

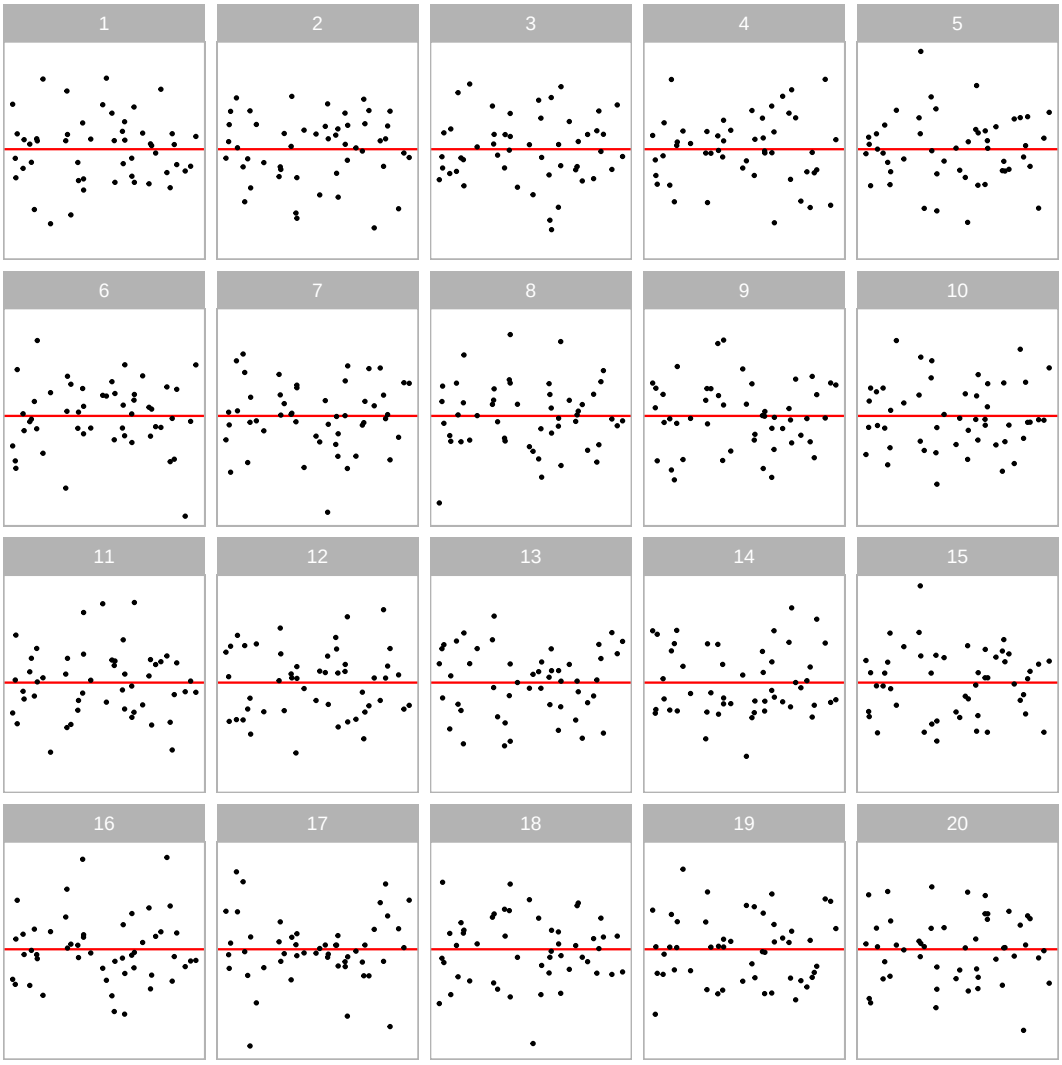

Figure 14. The single heteroskedasticity lineup that is rejected by the visual test but not by the BP test. The data plot at panel 17 contains a "butterfly" shape. It has effect size = 3.76, somewhat surprising that it is not detected by the BP test.

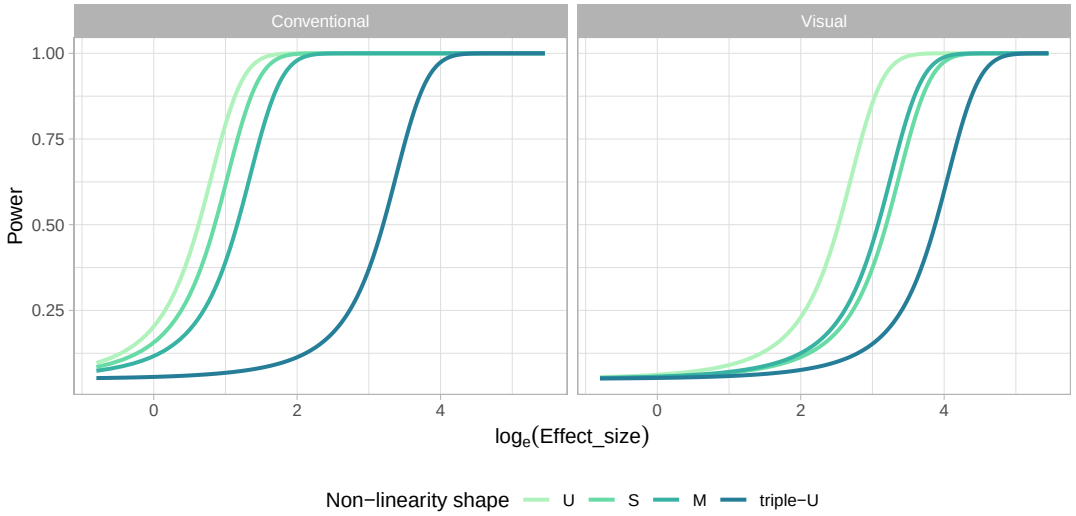

Figure 15. Power of conventional tests and visual tests on lineups containing four different non-linearity shapes. Power curves with higher order of non-linearity are drawn with deeper colours. The default RESET tests under-perform significantly in detecting "triple-U" shape. To achieve a similar power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

5.4. *Effect of shape of non-linearity*

A primary factor contributes to the non-linearity model is the shape of the non-linearity pattern. According to Figure 15, conventional tests have higher power in testing shapes constructed from lower orders of Hermite polynomials. But the chance of detecting the "triple-U" shape drops significantly compared to other shapes. 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 be only slight lower than the power curve of "M" shape. The recommendation of the polynomial power for the RESET test 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 generating process in order to set the parameters for the test.

For visual tests, based on the orders of the Hermite polynomials, 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 "S" shape. This implies, unlike the conventional test, the visibility of the shape do not strictly follow the degree of the polynomials.

5.5. *Effect of shape of heteroskedasticity*

We have also investigated 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

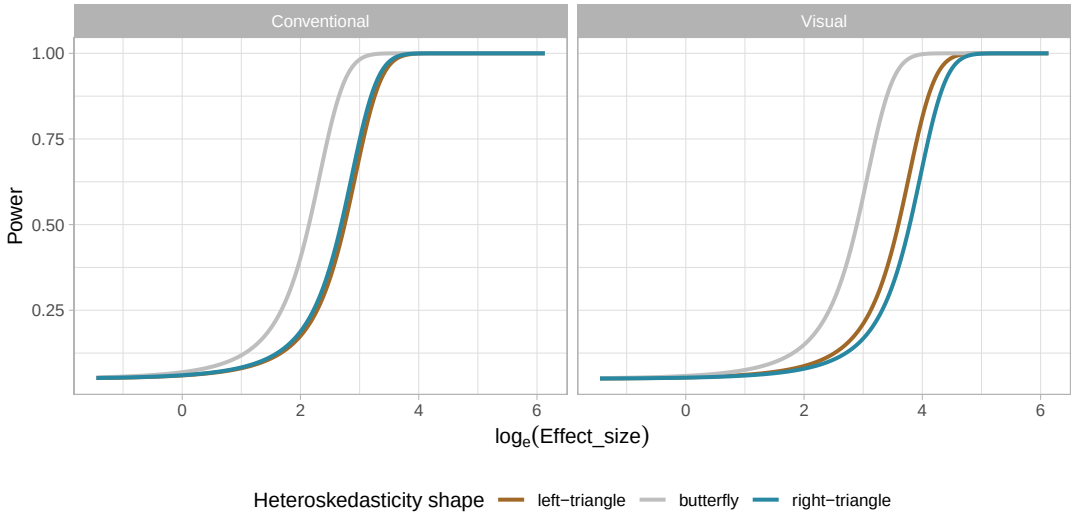

Figure 16. Power of conventional tests and visual tests on lineups produced by the heteroskedasticity model with three different shapes controlled by the parameter a. The visual test has an asymmetric power in testing the "left-triangle" shape and "right-triangle" shape, while these two shapes are equivalent in conventional testing.

test. As shown in Figure 16, this is indeed the case where little difference between the power curves can be perceived. Similarly, visual tests should have the same power in detecting these two shapes if they are equally likely to be identified. However, it can be observed from Figure 16 that the power curve of the "left-triangle" shape is higher than the one of the "right-triangle" shape, indicating a potential favour of orientation by human, which is worth to be explored in future studies. Besides, the chance of detecting the "butterfly" shape are higher than other two shapes in both conventional and visual testing.

5.6. *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 impact of fitted value distribution on the power of conventional tests and visual tests. For conventional tests, only the power curves of appropriate tests which are RESET tests and Breusch-Pagan tests for non-linearity pattern and heteroskedasticity pattern respectively are shown. For visual tests, note that we have collected more evaluations on lineup with uniform fitted value distribution, and the number of evaluations on a lineup will affect the power of the visual test. To have a fair comparison, for lineups with uniform fitted value distribution, we randomly sample five evaluations from the total eleven evaluations to estimate the power curves. The sampling process is repeated 500 times to produce an indication of the variation of the power.

In terms of visual tests, we do not observe significant difference between the power curve of the uniform distribution and the power curve of the discrete uniform distribution. Normal and lognormal distributions produce lower power than uniform and

discrete uniform distributions in regards of residual departure patterns, but the gap is larger when the residual plot contains heteroskedasticity pattern. This suggests that residual departures displayed with evenly distributed fitted values are visually more attractive and effective. Besides, the skewness of the fitted value distribution plays a role in the degree of presence of visual features as lognormal distribution produce the lowest power for both non-linearity and heteroskedasticity patterns. A symmetric fitted value distribution can better reveal the underlying shape.

Although it is unusual to consider fitted value distribution or predictor distribution in power analysis, it is not so surprised that the power of conventional tests vary because of fitted value distribution. For instance, the power of the RESET test which is effectively a F test depends on the realizations of the additional predictors included in the auxiliary regression equation (Jamshidian, Jennrich, and Liu 2007; Olvera Astivia, Gadermann, and Guhn 2019; Zhang and Yuan 2018). And those additional predictors are linear transformation of random vectors following certain predictor distributions. What is truly unexpected is the huge variability in power of conventional tests compared to visual tests. The discrete uniform distribution constantly produce the highest power while the normal distribution and the lognormal distribution produce the lowest power for non-linearity and heteroskedasticity pattern respectively. The difference in power corresponding to different distributions peak at around 90% when the effect size is at a moderate level. This huge difference is undesirable as is not uncommon for analysts to omit the factor of predictor distribution in power calculation and sample size calculation.

6. Conclusion

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

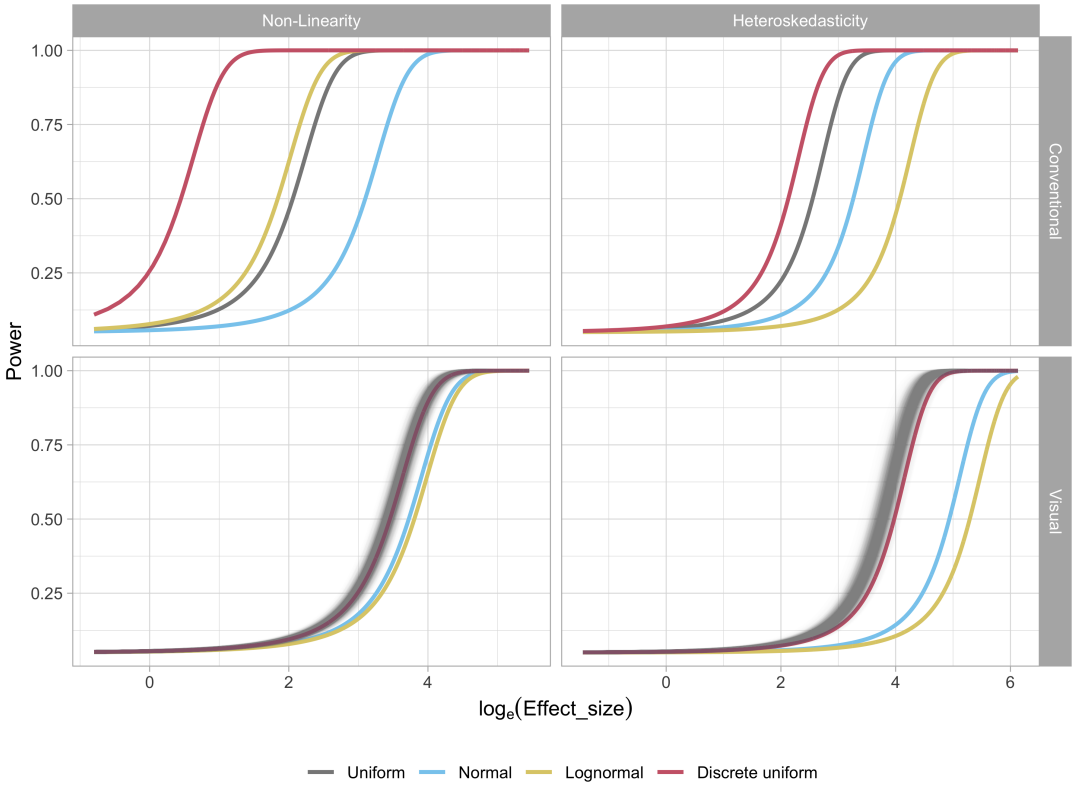

Figure 17. Comparison of power on lineups with different fitted value distributions for conventional tests and visual tests. Power curves of conventional tests for non-linearity and heteroskedasticity patterns are produced by RESET tests and Breusch-Pagan tests respectively. Power curves of visual tests are estimated using five evaluations on each lineup. For lineups with uniform fitted value distribution, the five evaluations are randomly sampled from the total eleven evaluations. The sampling process has been repeated for 500 times and the corresponding power curves are drawn with transparent grey lines. The fitted value distribution has greater impact on the power of conventional tests than visual tests. Meanwhile, uneven distributions including normal and lognormal distributions genearlly produce lower power.

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 (link here) contains all materials required to reproduce this article.

Supplementary material

The supplementary material is available at (link here). 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.

Appendix A. Experiment setup

A.1. *Mapping of subjects to experimental factors*

Mapping of subjects to experimental factors is an important part of experiment design. Essentially, we want to maximum the difference in factors exposed to a subject. For this purpose, we design an algorithm to conduct subject allocation. Let L be a set of available lineups and S be a set of available subjects. According to the experimental design, the availability of a lineup is associated with the number of subjects it can assign to. For lineups with uniform fitted value distribution, this value is 11. And other lineups can be allocated to at most five different subjects. The availability of a subject is associated with the number of lineups that being allocated to this subject. A subject can view at most 18 different lineups.

The algorithm starts from picking a random subject $s \in S$ with the minimum number of allocated lineups. It then tries to find a lineup $l \in L$ that can maximise the distance metric D and allocate it to subject s. Set L and S will be updated and the picking process will be repeated until there is no available lineups or subjects.

Let $F_1, ..., F_q$ be q experimental factors, and $f_1, ..., f_q$ be the corresponding factor values. We say f_i exists in L_s if any lineup in L_s has this factor value. Similarly, $f_i f_j$ exists in L_s if any lineup in L_s has this pair of factor values. And $f_i f_j f_k$ exists in L_s if any lineup in L_s has this trio of factor values. The distance metric D is defined between a lineup l and a set of lineups L_s allocated to a subject s if L_s is non-empty:

$$D = C - \sum_{1 \le i \le q} I(f_i \text{ exists in } L_s) - \sum_{\substack{1 \le i \le q-1 \\ i < j \le q}} I(f_i f_j \text{ exists in } L_s) - \sum_{\substack{1 \le i \le q-2 \\ i < j \le q-1 \\ j < k \le q}} I(f_i f_j f_k \text{ exists in } L_s)$$

where C is a sufficiently large constant such that $D > 0$. If L_s is empty, we define $D = 0$.

The distance measures how different a lineup is from the set of lineups allocated to the subject in terms of factor values. Thus, the algorithm will try to allocate the most different lineup to a subject at each step.

A.2. *Data collection process*

The survey data is collected via a self-hosted website designed by us. The complete architecture is provided in Figure A1. The website is built with the `Flask` (Grinberg 2018) web framework and hosted on `PythonAnywhere` (PythonAnywhere LLP 2023). It is configured to handle HTTP requests such that subjects can correctly receive webpages and submit responses. Embedded in the resources sent to subjects, the `jsPsych` front-end framework (De Leeuw 2015) instructs subjects' browsers to render an environment for running behavioral experiments. During the experiment, this framework will automatically collect common behavioral data such as response time and clicks on buttons. Subjects' responses are first validated by a scheduled `Python` script run on the server, then push to a Github repository. Lineup images shown to users are saved in multiple Github repositories and hosted in corresponding Github pages. The URLs to these images are resolved by `Flask` and bundled in HTML files.

Once the participant is recruited from Prolific (Palan and Schitter 2018), it will be redirected to the entry page of our study website. An image of the entry page is provided in Figure A2. Then, the participant needs to submit the online consent form and fill in the demographic information as shown in A3 and A4 respectively. Before

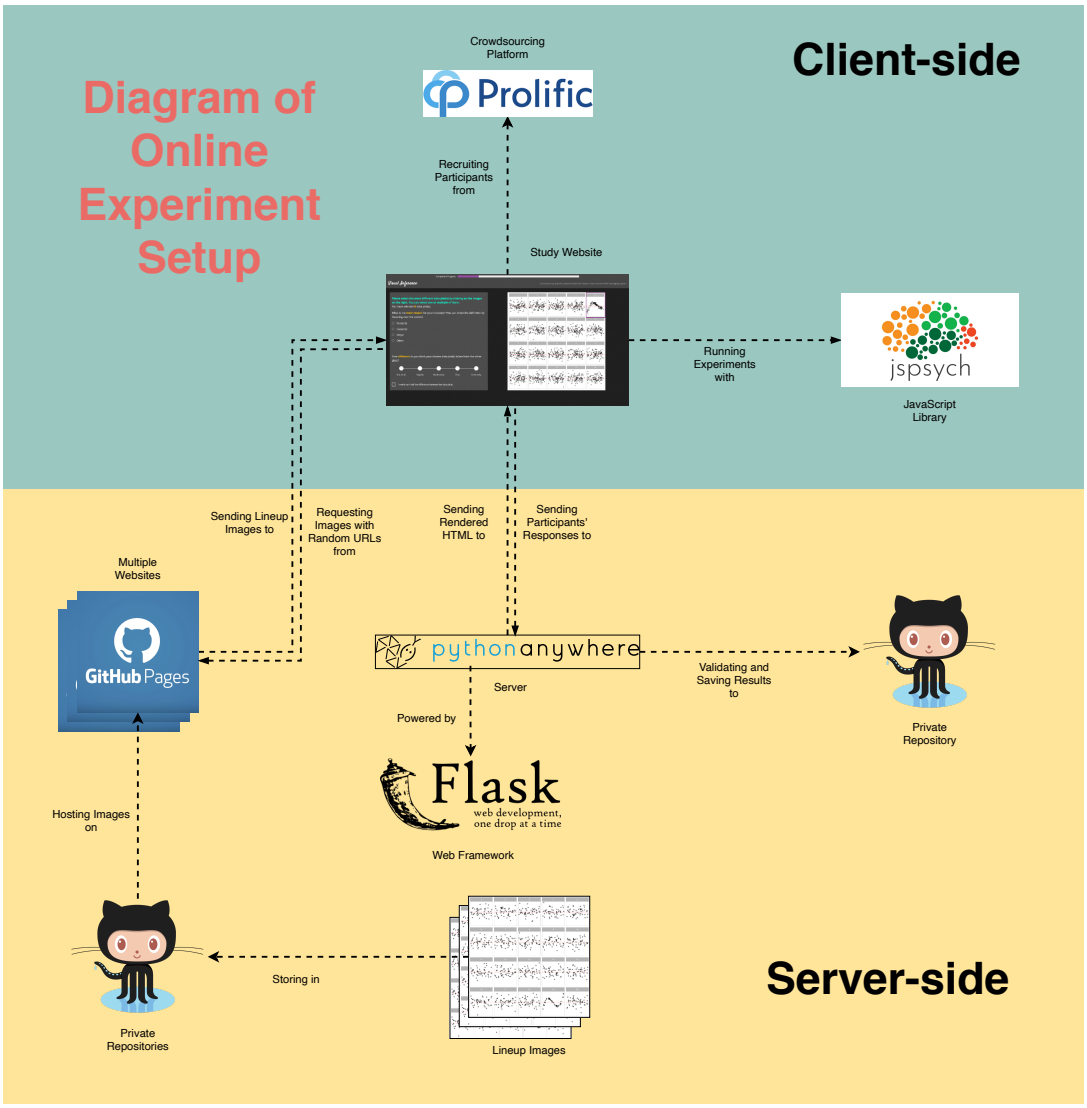

Figure A1. Diagram of online experiment setup. The server-side of the study website uses Flask as backend hosted on PythonAnywhere. And the client-side uses jsPsych to run experiment.

evaluating lineups, participant also need to read the training page as provide in Figure A5 to understand the process. An example of the lineup page is given in Figure A6. A half of the page is taken by the lineup image to attract participant's attention. The button to skip the selections for the current lineup is intentionally put in the corner of the bounding box with smaller font size, such that participants will not misuse this functionality.

Appendix B. Demographics

Along with the responses to lineups, we have collected a series of demographic information including age, pronoun, education background and previous experience in studies involved data visualization. Table B1, B2, B3 and B4 provide summary of the demographic data.

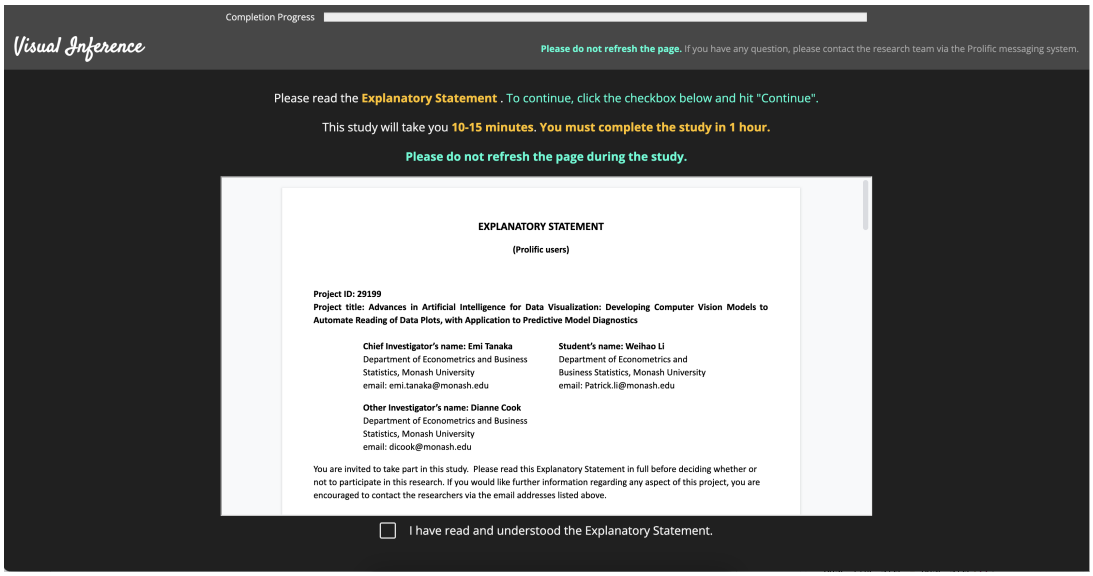

Figure A2. The entry page of the study website.

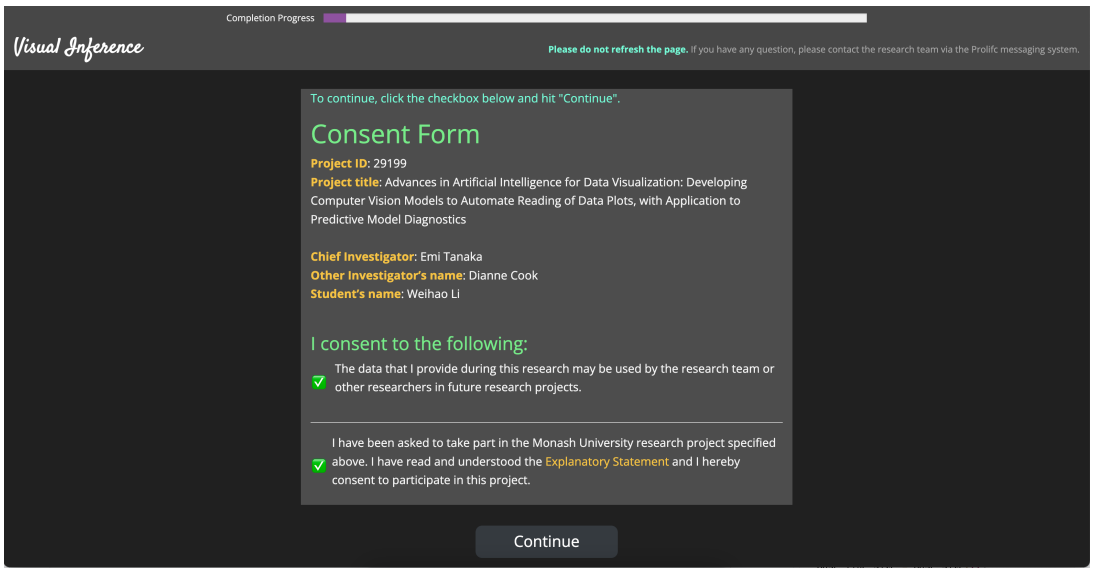

Figure A3. The consent form provided in the study website.

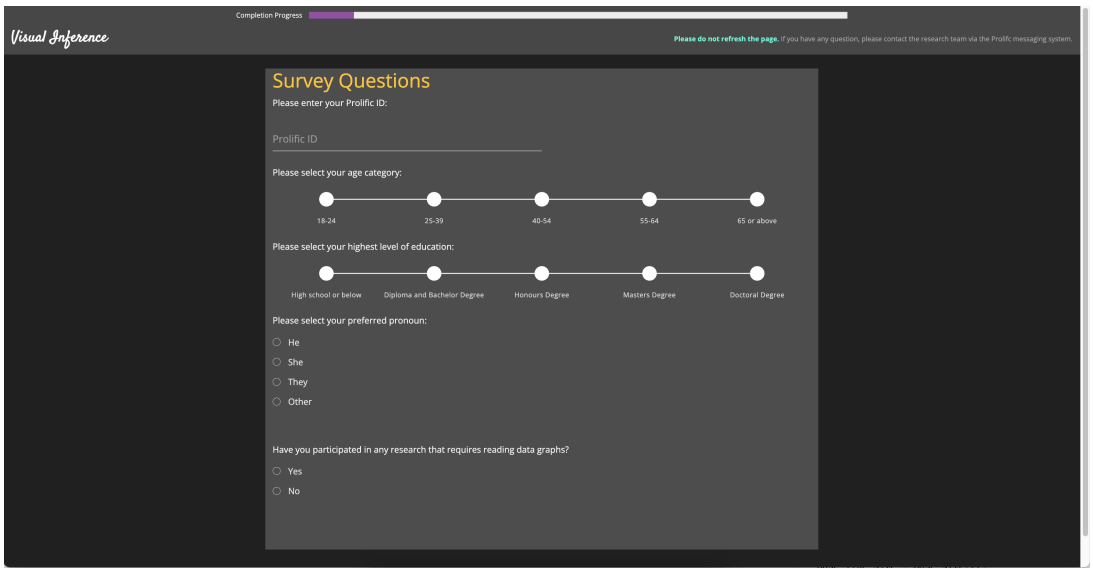

Figure A4. The form to provide demographic information.

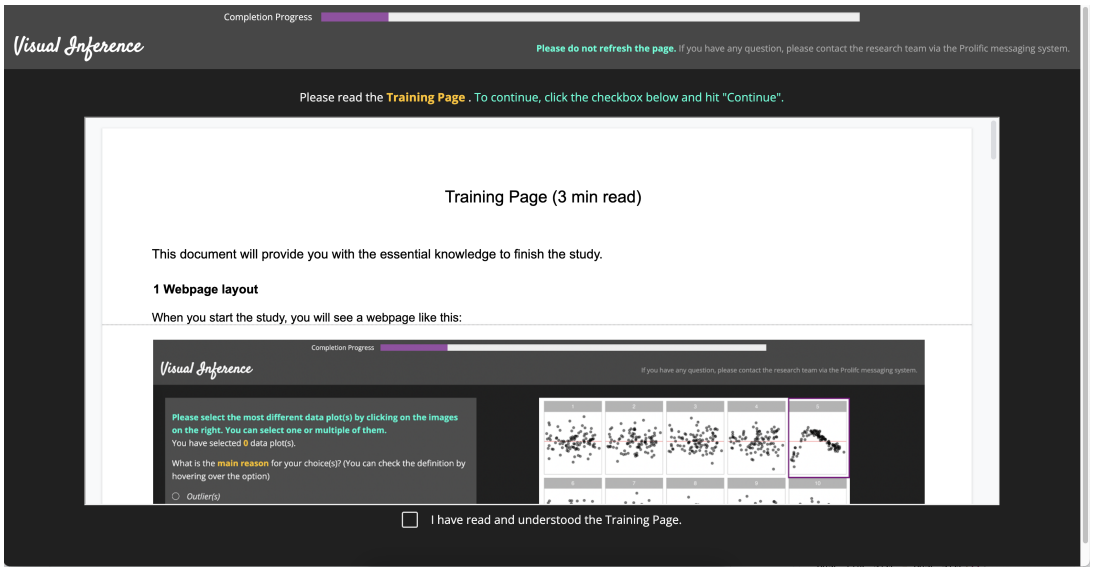

Figure A5. The training page of the study website.

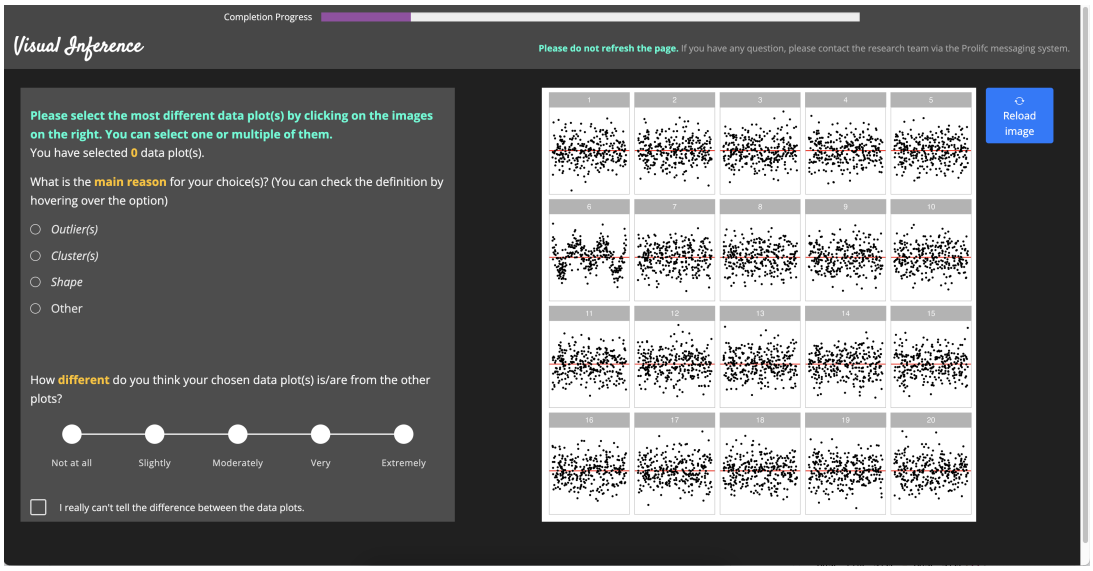

Figure A6. The lineup page of the study website.

Table B1. Summary of pronoun distribution of subjects recuritted in this study.

Pronoun	Period I	%	Period II	%	Period III	%	Total	%
He	77	17.4	79	17.8	61	13.8	217	49.0
She	78	17.6	77	17.4	61	13.8	216	48.8
Other	5	1.1	4	0.9	1	0.2	10	2.3
	160	36.1	160	36.1	123	27.8	443	100.0

It can be observed from the tables that most participants have Diploma or Bachelor degrees, followed by High school or below and the survey data is gender balanced. Majority of participants are between 18 to 39 years old and there are slightly more participants who do not have previous experience than those who have.

Appendix C. Effect size derivation

Effect size can be defined as the difference of a parameter for a particular model or distribution, or a statistic derived from a sample. Importantly, it needs to reflect the

Table B2. Summary of age distribution of subjects recuritted in this study.

Age group	Period I	Period II	Period III	Total
18-24	83	86	51	220
25-39	69	63	63	195
40-54	6	8	6	20
55-64	2	3	3	8

Table B3. Summary of education distribution of subjects recuritted in this study.

Education	Period I	Period II	Period III	Total
High School or below	41 (9.3%)	53 (12%)	33 (7.4%)	127 (28.7%)
Diploma and Bachelor Degree	92 (20.8%)	79 (17.8%)	66 (14.9%)	237 (53.5%)
Honours Degree	6 (1.4%)	15 (3.4%)	6 (1.4%)	27 (6.1%)
Masters Degree	21 (4.7%)	13 (2.9%)	16 (3.6%)	50 (11.3%)
Doctoral Degree	0 (0%)	0 (0%)	2 (0.5%)	2 (0.5%)
	160 (36.1%)	160 (36.1%)	123 (27.8%)	443 (100%)

Table B4. Summary of previous experience distribution of subjects recuritted in this study.

Previous experience	Period I	Period II	Period III	Total
No	96	88	67	251
Yes	64	72	56	192

treatment we try to measure. Centred on a conventional statistical test, we usually can deduce the effect size from the test statistic by substituting the null parameter value. When considering the diagnostics of residual departures, there exist many possibilities of test statistics for a variety of model assumptions. Meanwhile, diagnostic plots such as the residual plot have no general agreement on measuring how strong a model violation pattern is. To build a bridge between various residual-based tests, and the visual test, we focus on the shared information embedded in the testing procedures, which is the distribution of residuals. When comes to comparison of distribution, Kullback-Leibler divergence (Kullback and Leibler 1951) is a classical way to represent the information loss or entropy increase caused by the approximation to the true distribution, which in our case, the inefficiency due to the use of false model assumptions.

Following the terminology introduced by Kullback and Leibler (1951), P represents the measured probability distribution, and Q represents the assumed probability distribution. The Kullback-Leibler divergence is defined as $\int_{-\infty}^{\infty} log(p(x)/q(x))p(x)dx$, where $p(.)$ and $q(.)$ denote probability densities of P and Q.

Let $\boldsymbol{X}_a = (\mathbf{1}, \boldsymbol{X})$ denotes the p regressors with n observations, $\boldsymbol{R}_a = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ denotes the residual operator, and let $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$ denotes the error. Using the Frisch–Waugh–Lovell theorem, residuals $\boldsymbol{e} = \boldsymbol{R}_a\boldsymbol{\varepsilon}$. Because $rank(\boldsymbol{R}_a) = n - p < n$, e follows a degenerate multivariate normal distribution and does not have a density. Since the Kullback-Leibler divergence requires a proper density function, we need to simplify the covariance matrix of $\boldsymbol{e}$ by setting all the off-diagonal elements to 0. Then, the residuals will assumed to follow $N(\mathbf{0}, diag(\boldsymbol{R}_a\sigma^2))$ under the null hypothesis that the model is correctly specified. If the model is however misspecified due to omitted variables $\boldsymbol{Z}$, or a non-constant variance $\boldsymbol{V}$, the distribution of residuals can be derived as $N(\boldsymbol{R}_a\boldsymbol{Z}\boldsymbol{\beta}_z, diag(\boldsymbol{R}_a\sigma^2))$ and $N(\mathbf{0}, diag(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a'))$ respectively.

By assuming both P and Q are multivariate normal density functions, the Kullback-

Leibler divergence can be rewritten as

$$KL = \frac{1}{2}\left(log\frac{|\Sigma_p|}{|\Sigma_q|} - n + tr(\Sigma_p^{-1}\Sigma_q) + (\mu_p - \mu_q)'\Sigma_p^{-1}(\mu_p - \mu_q)\right).$$

Then, we can combine the two residual departures into one formula

$$KL = \frac{1}{2}\left(log\frac{|diag(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a')|}{|diag(\boldsymbol{R}_a\sigma^2)|} - n + tr(diag(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a')^{-1}diag(\boldsymbol{R}_a\sigma^2)) + \boldsymbol{\mu}_z^T(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a')^{-1}\boldsymbol{\mu}_z\right). \tag{C1}$$

When there are omitted variables but constant error variance, the formula can be reduced to

$$KL = \frac{1}{2}\left(\boldsymbol{\mu}_z^T(diag(\boldsymbol{R}_a\sigma^2))^{-1}\boldsymbol{\mu}_z\right).$$

And when the model equation is correctly specified but the error variance is non-constant, the formula can be reduced to

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

Since we assume $\sigma = 1$ for the heteroskedasticity model, the final form of the formula is

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

Appendix D. Effect of data collection period

We have the same type of model collected over different data collection periods, that may lead to unexpected batch effect. Figure D1 and D2 provide two lineups to examine whether there is an actual difference across data collection periods for non-linearity model and heteroskedasticity model respectively. To emphasize the tail behaviour and display fewer outliers, we use the "letter-value" boxplot (Hofmann, Wickham, and Kafadar 2017) which is an extension of the number of "letter value" statistics to check the weighed proportion of detect over different data collection period. The weighted proportion of detect is calculated by taking the average of c_i of a lineup over a data collection period. Within our research team, we can not identify the data plot from the null plots for these two lineups, result in p-values much greater than 5%. Thus, there is no clear evidence of batch effect.

Appendix E. Sensitivity analysis for α

The parameter α used for the p-value calculation needs to be estimated from responses to null lineups. With a greater value of $\hat{\alpha}$, the p-value will be smaller, resulting in

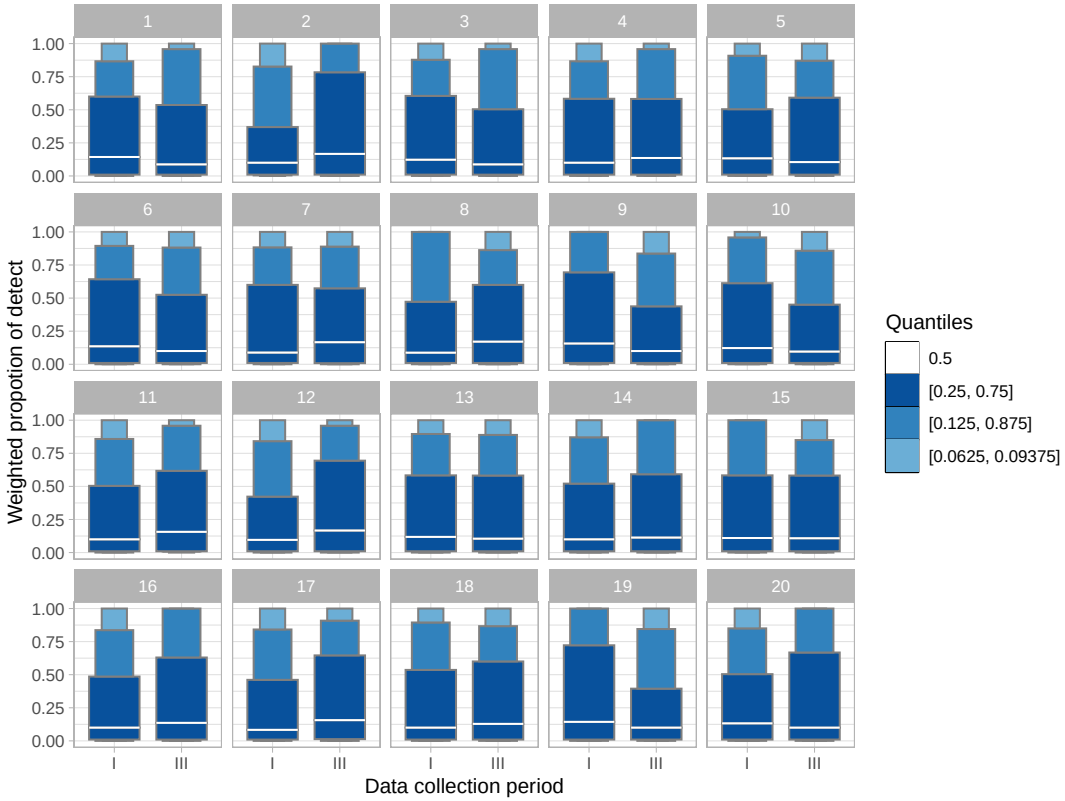

Figure D1. A lineup of "letter-value" boxplots of weighted propotion of detect for lineups over different data collection periods for non-linearity model. Can you find the most different boxplot? The data plot is positioned in panel $2^3 - 1$.

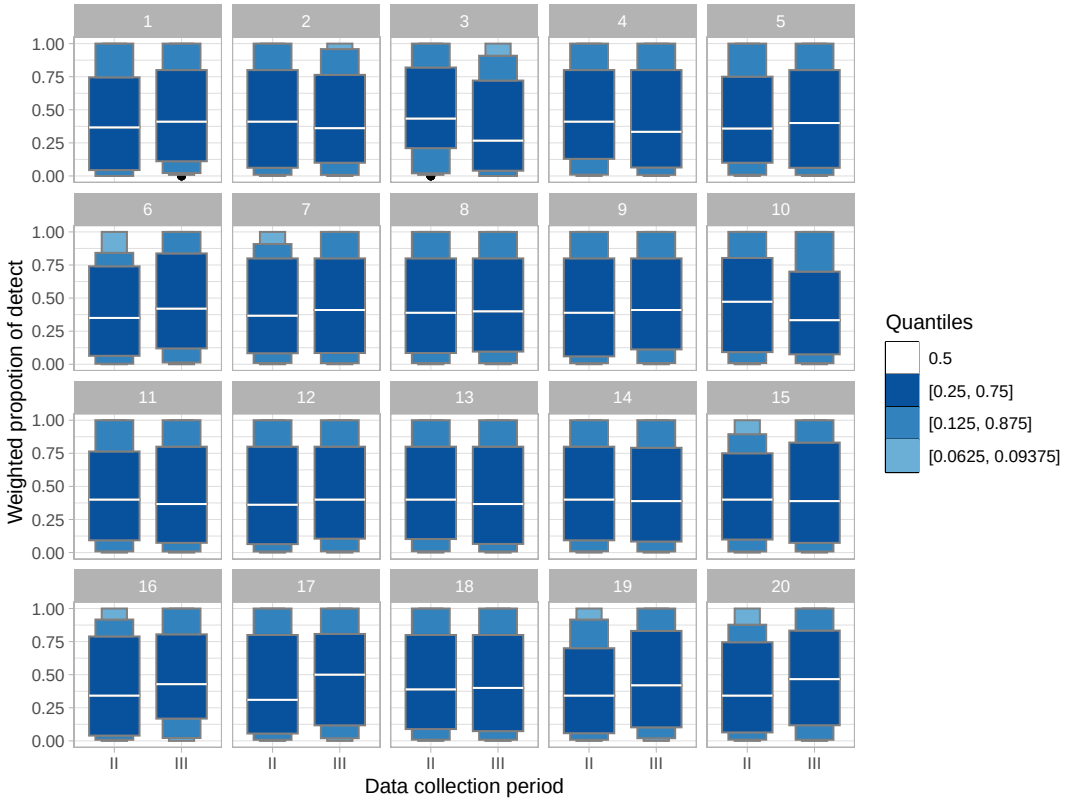

Figure D2. A lineup of "letter-value" boxplots of weighted propotion of detect for lineups over different data collection periods for heteroskedasticity model. Can you find the most different boxplot? The data plot is positioned in panel $2^4 - 2$.

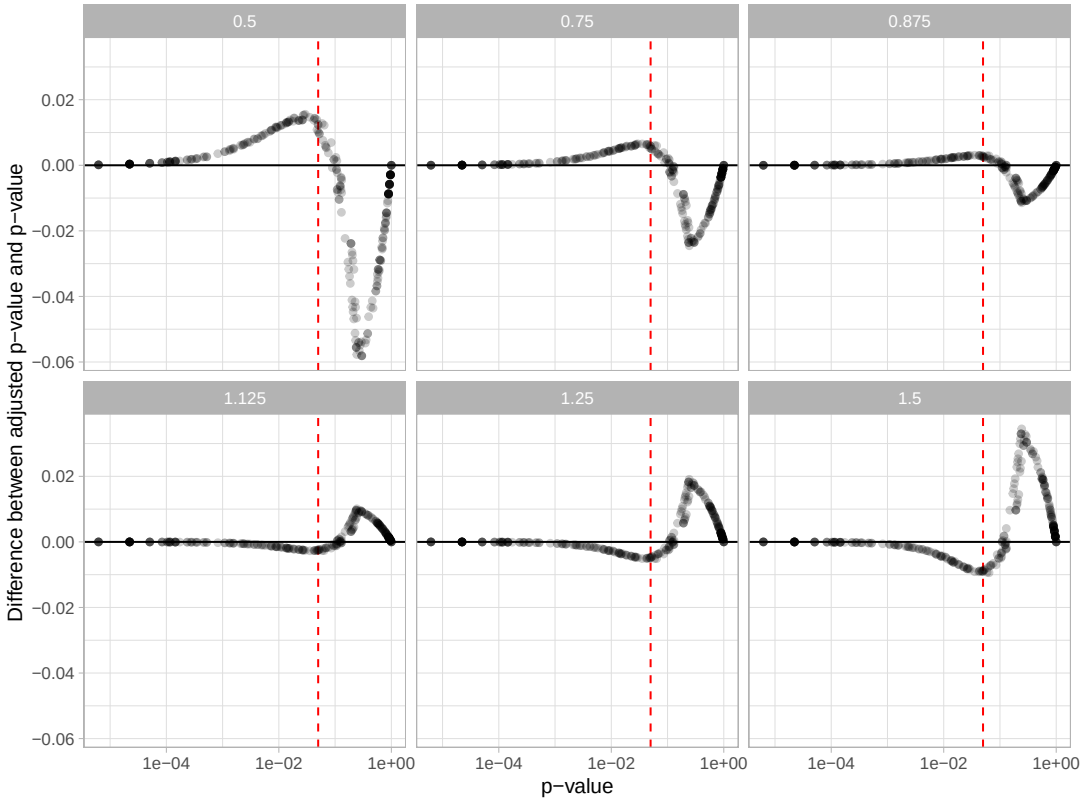

Figure E1. Change of p-values with $\hat{\alpha}$ multiplied by 0.5, 0.75, 0.875, 1.125, 1.25 and 1.5. The vertical dashed line is to indicate p-value = 0.05. The x-axis is drawn on logarithmic scale. For multiplier smaller than 1, the adjusted p-value will incrase then decrase as p-value increases. The trend is the opposite for multiplier greater than 1, but the difference will eventually reach 0.

more lineups being rejected. However, The way we generate Rorschach lineup is not strictly the same as what suggested in VanderPlas et al. (2021) and Buja et al. (2009). Therefore, we conduct a sensitivity analysis in this section to examine the impact of the variation of the estimator α on our primary findings.

The analysis is conducted by setting up several scenarios, where the α is under or overestimated by 12.5%, 25% and 50%. Using the adjusted $\hat{\alpha}$, we recalculate the p-value for every lineup and show the results in Figure E1. It can be observed that there are some changes to p-values, especially when the $\hat{\alpha}$ is multiplied by 50%. However, Table E1 shows that adjusting $\hat{\alpha}$ will not result in a huge difference in rejection decisions. There are only a small percentage of cases where the rejection decision change. It is very unlikely the downstream findings will be affected because of the estimate of α.

Appendix F. Effect of number of evaluations on the power of a visual test

When comparing power of visual tests across different fitted value distributions, we have discussed the number of evaluations on a lineup will affect the power of the visual test. Using the lineups with uniform fitted value distribution, we show in Figure F1 the change of power of visual tests due to different number of evaluations. It can be learned that as the number of evaluations increases, the power will increase but the margin will decrease. Considering we have eleven evaluations on lineups with uniform

Table E1. Change of rejection decision because of the modification of $\hat{\alpha}$.

multiplier	Proportion of lineups transforms to "not reject"	Proportion of lineups transforms to "reje
0.500	2.51%	0%
0.750	1.43%	0%
0.875	1.08%	0%
1.125	0%	1.08%
1.250	0%	1.43%
1.500	0%	1.79%

fitted value distribution, and five evaluations on other lineups, it is necessary to use the same number of evaluations for each lineup in comparison.

Appendix G. Power of a RESET test under different auxiliary regression formulas

It is found in the result that the power of a RESET test will be affected by the highest order of fitted values included in the auxiliary formula. And we suspect that the current recommendation of the highest order - four, is insufficient to test complex non-linear structures such as the "Triple-U" shape designed in this paper. Figure G1 illustrates the change of power of RESET test while testing the "U" shape and the "Triple-U" shape with different highest orders. Clearly, when testing a simple shape like the "U" shape, the highest order has very little impact on the power. But for testing the "Triple-U" shape, there will be a loss of power if the recommended order is used. To avoid the loss of power, the highest order needs to be set to at least six.

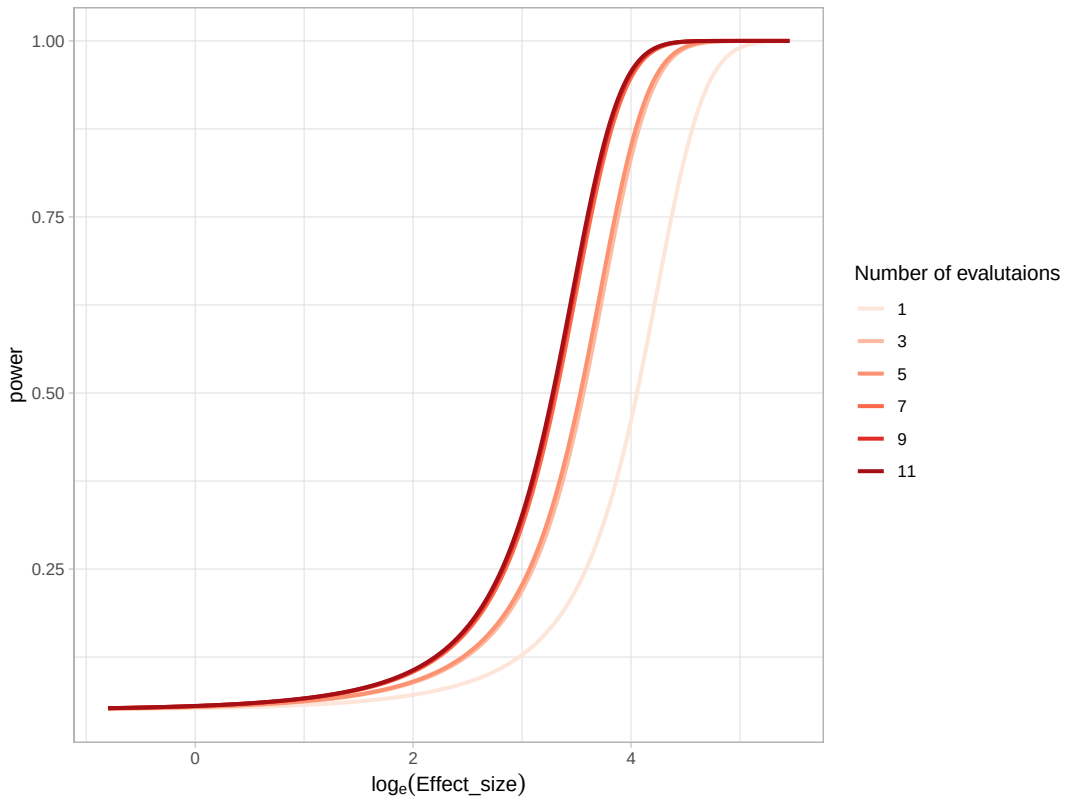

Figure F1. Change of power of visual tests for different number of evalutions on lineups with uniform fitted value distribution. The power will incrase as the number of evaluations increases, but the margin will decrase.

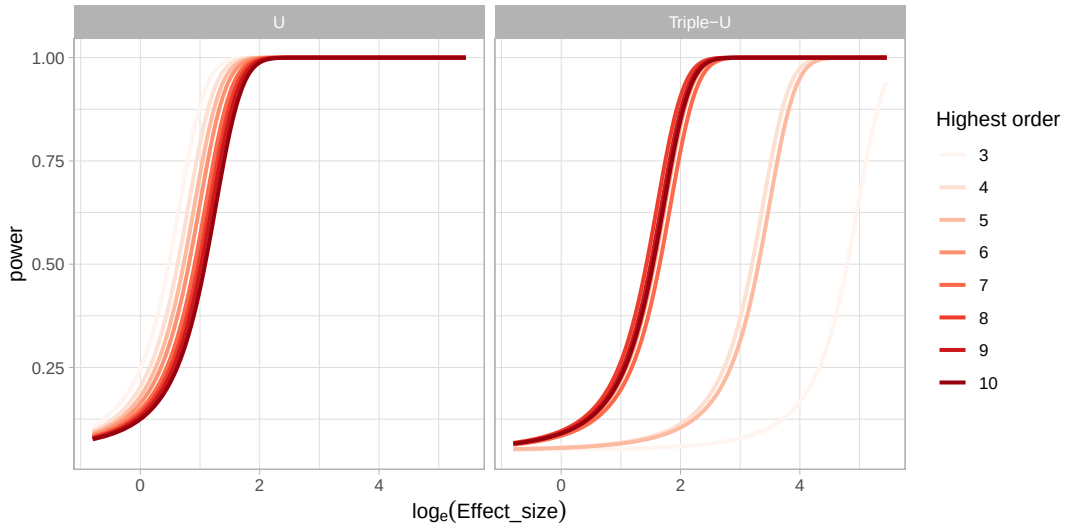

Figure G1. Change of power of RESET tests for different orders of fitted values included in the auxiliary formula. The left panel is the power of testing the "U" shape and the right panel is the power of testing the "Triple-U" shape. The power will not be greatly affected by the highest order in the case of testing the "U" shape. In the case of testing the "Triple-U" shape, the highest order needs to be set to at least six to avoid the loss of power.

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.

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.

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.

Hofmann, Heike, Hadley Wickham, and Karen Kafadar. 2017. "value plots: Boxplots for large data." *Journal of Computational and Graphical Statistics* 26 (3): 469–477.

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.

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.

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.