ARTICLE TEMPLATE

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

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

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

ARTICLE HISTORY
Compiled December 21, 2022

ABSTRACT
Abstract to fill.

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

1. Introduction

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

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

We can assess residuals in multiple ways. Residuals might be plotted, as a histogram or quantile-quantile plot to examine the distribution. Using the classical normal linear regression model as an example, if the distribution is symmetric and unimodal, it is well-behaved. But if the distribution is skewed, bimodal, multimodal, or contains outliers, there is cause for concern. The distribution could also be inspected 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 is any visually discoverable patterns, the model is potentially misspecified. However, correctly judging a residual plot where no pattern exists is a painstakingly difficult task for humans. It is especially common, particularly among new data analytics students to report patterns when an experienced data analyst might quickly conclude that there are none. Generally, one looks for noticeable departures from the model like non-linear dependency or heteroskedasticity. It is also possible to conduct hypothesis tests for non-linear dependence (Ramsey 1969), and use a Breusch-Pagan test (Breusch

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

and Pagan 1979) for heteroskedasticity.

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

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

2. Background

2.1. *Departures from good residual plots*

Graphical summaries in which residuals are plotted against fitted values or other functions of the regressors that are approximately orthogonal to residuals are referred to as standard residual plots in Cook and Weisberg (1982). As shown in Figure 1, the top-left panel is a good residual plot with residuals evenly distributed at both sides of the horizontal zero line showing no noticeable patterns.

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

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

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

Compared to non-linearity and heteroskedasticity, non-normality is usually harder to detect from a residual plot since scatter plot do not readily reveal the marginal distribution. A favourable graphical summary for this task is the quantile-quantile plot. For a consistent comparison, the residual plot of this departure is still presented in plot D of Figure 1. However, since the scatter plot is not suitable for this task, non-normality will not be the focus of this paper. Besides, it is important to note that

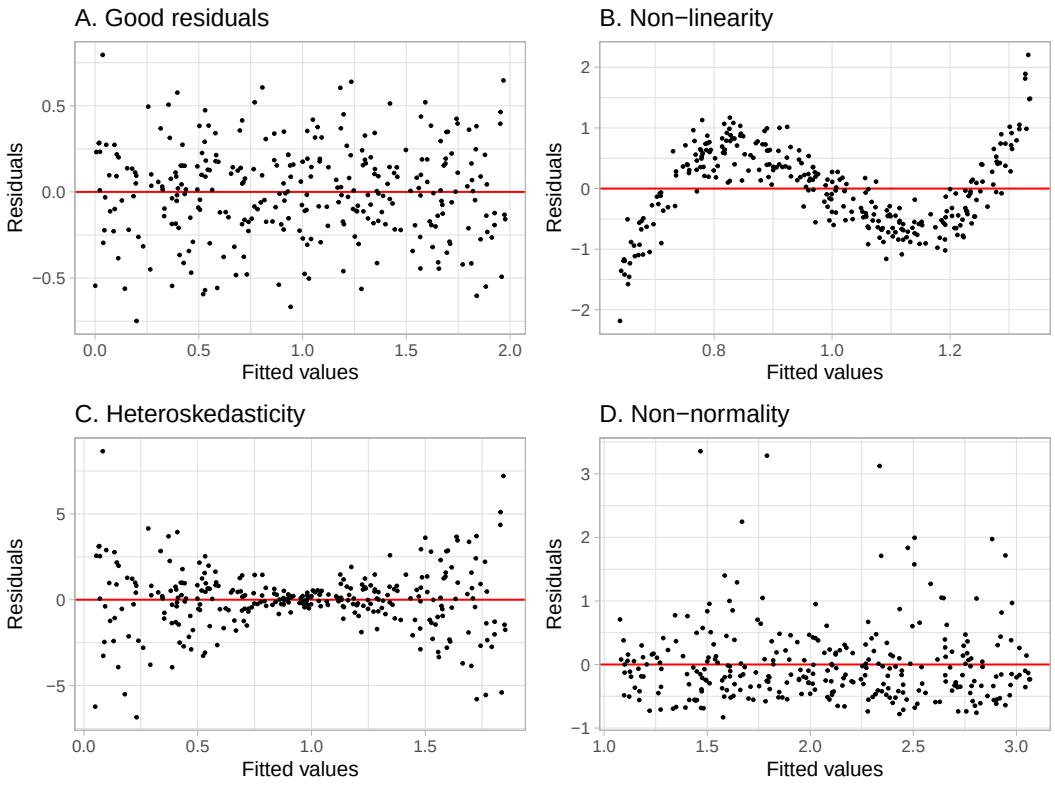


Figure 1. Example fitted vs residual plots: (A) classically good looking residuals, (B) non-linear pattern indicates that the model has not captured a non-linear association, (C) heteroskedasticity indicating that variance around the fitted model is not uniform, and (D) non-normality where the residual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

not all regression models assume normality for the error term, but a certain amount do including the classical normal linear regression model. In the case that the normality assumption is violated, it is expected to observe data points do not center around the horizontal axis and there is an uneven distribution of the number points at both below and above the horizontal axis. For example, given a skewed error distribution, fewer data points and more outliers are on one side of the horizontal axis as shown in plot D of Figure 1.

2.2. *Conventionally testing for departures*

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

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

2.3. *Visual test procedure based on lineups*

2.3.1. Lineup protocol

One may argue that reading diagnostic plots is to some extent subjective and indecisive compared to those rigorous statistical procedures as it relies on graphical perception - human's ability to interpret and decode the information embedded in the graph (Cleveland and McGill 1984). It is particularly true that visual discovery suffers from its unsecured and unconfirmed nature where 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. One such example is finding

Table 1. Statistical significance testing for departures from good residuals for plots in Figure 1. Shown are the p-values calculated for the conventional 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*

an over-interpretation of 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 (Gelman 2004). 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 actual data plot, and the remaining $m-1$ plots have the identical graphical production as the data plot except the data has been replaced with data consistent with H_0. Then, an observer who have not seen the actual data plot will be asked to point out the most different plot from the lineup. Under H_0, it is expected that the actual data plot would have no distinguishable difference with the null plots, and the probability of the observer correctly picks the actual data plot is $1/m$. If we reject H_0 as the observer correctly picks the actual data plot, then the Type I error of this test is $1/m$.

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

The effectiveness of lineup protocol is validated by Majumder, Hofmann, and Cook (2013) under relatively simple classical normal linear regression model settings with one or two regressors. Their results suggest that visual test is capable of testing the significance of a single regressor with a similar power as a t-test, though they express that in general it is unnecessary to use visual inference if there exists a conventional test and they do not expect the visual test to perform equally well as the conventional test. In their third experiment, where there does not exist a proper conventional test, visual test outperforms the conventional test for a large margin. This is encouraging as it promotes the use of visual inference in border field of data science where there are no existing statistical testing procedures. Visual inference have also already 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.

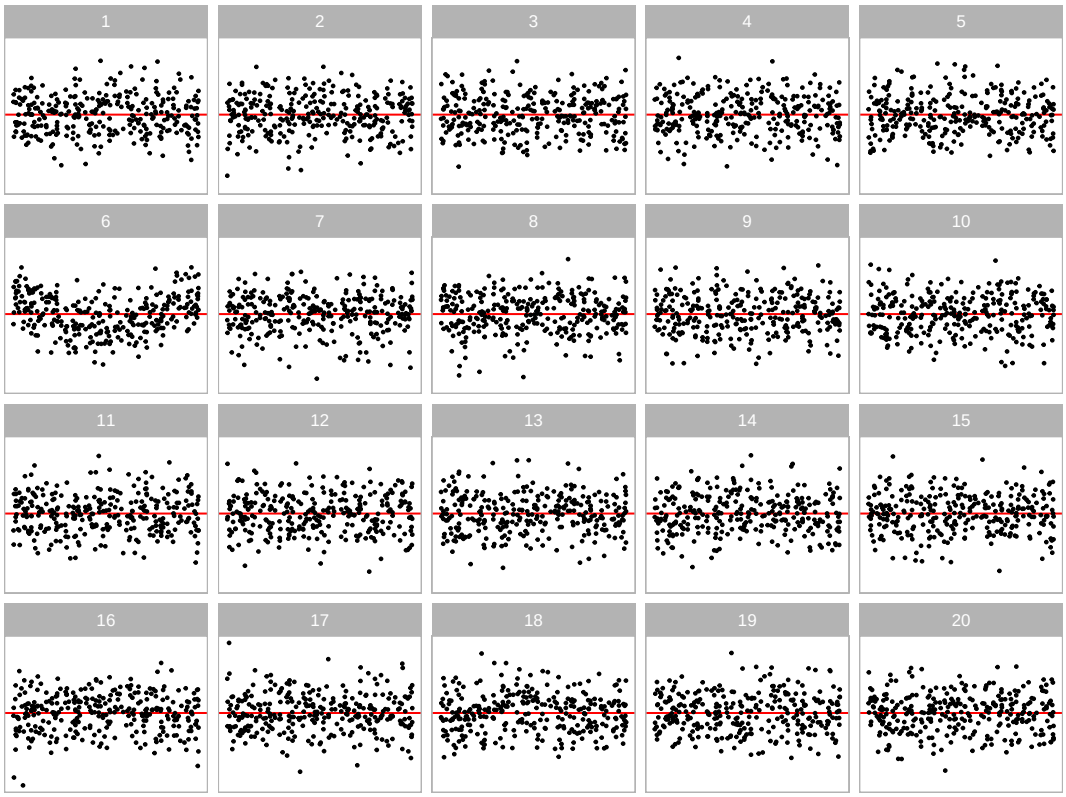


Figure 2. Visual testing is conducted using a lineup, as in the example here. The residual plot computed from the observed data (plot $2^2 + 2$, exhibiting non-linearity) is embedded among 19 null plots, where the residuals were simulated from a standard error model. Computing the p-value requires that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

2.3.2. Sampling from the null distribution

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

2.3.3. Calculating p-values for the visual test

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

As pointed out by VanderPlas et al. (2021), this basic binomial model doesn't take into account the possible dependencies in the visual test due to repeated evaluations of the same lineup. And it is inapplicable to visual test where subjects are asked to select one or more "most different" plots from the lineup. VanderPlas et al. (2021) summarises three common scenarios in visual inference: (1) K different lineups are shown to K subjects, (2) K lineups with different null plots but the same actual 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) model 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 $\alpha > 0$. And defined c_i to be the number of times plot i being selected in K evaluations. 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 ensured $\sum_i c_i = K$. Since we are only interested in the selections of the actual 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 only works with non-negative integer c_i. For non-negative real number c_i, linear approximation will be applied to calculate the p-value

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

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 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) define 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

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

VanderPlas et al. (2021) suggest that α can be estimated using MLE with Equation 5. But for precise estimate of α, additional responses to Rorschach lineups, which is a type of lineup that consists of plots constructed from the same null data generating mechanism, are required.

2.3.4. Power of a visual test

As discussed in Majumder, Hofmann, and Cook (2013), individual's skill will affect the number of observers who identify the actual data plot from the lineup. Thus, the power of a visual test depends on the subject-specific abilities. Previously, it is addressed by modelling the probability of a subject i correctly picking the actual data plot from a lineup l using a mixed-effect logistic regression with the subject being treated as a random effect (Majumder, Hofmann, and Cook 2013). However, having this probability is insufficient to determine the power of a visual test allowed for multiple selections as it doesn't provide information about the number of selections made by the subject for p-value calculation.

Instead, we omitted the individual effect and directly estimated the probability of a lineup being rejected using a logistic regression with the natural logarithm of the effect size as the only regressor formulated as

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

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

Effect E is derived from the Kullback-Leibler divergence (see [appendix ref here]) formulated as

$$E = \frac{1}{2\sigma^2} \boldsymbol{X}'_b \boldsymbol{R}'_a (diag(\boldsymbol{R}_a))^{-1} \boldsymbol{R}_a \boldsymbol{X}_b, \tag{7}$$

where $diag(.)$ is the diagonal matrix constructed from the diagonal elements of $\boldsymbol{R}_a$.

To study various factors contributing to the power of the visual test, the same logistic regression model is fit on different subsets of the collated data grouped by levels of factors. This includes [expansion].

3. Experimental design

Three experiments were conducted to investigate the difference between conventional hypothesis testing and visual inference in the application of linear regression diagnostics. The experiment I has ideal scenario for conventional testing, where the visual test is not expected to outperform the conventional test. Meanwhile, the experiment II is a scenario where the conventional test is an approximate test, in which the visual test may have a chance to match the performance of the conventional test. The experiment III is designed for collecting human responses to lineup with only good residual plots such that the parameter α in Equation 1 can be estimated. Overall, we planned to collect 7974 evaluations on 1152 uniqued lineups performed by 443 subjects throughout three experiment.

3.1. *Simulating departures from good residuals*

Two types of departures, namely non-linearity and heteroskedasticity, were considered with the corresponding data generating process being designed for experiment I and II.

3.1.1. Non-linearity

Experiment I is designed to study the ability of human subjects to detect the effect of a non-linear term $\boldsymbol{z}$ constructed using Hermite polynomials (Hermite 1864) on random vector $\boldsymbol{x}$ formulated as

$$\boldsymbol{y} = 1 + \boldsymbol{x} + \boldsymbol{z} + \boldsymbol{\varepsilon}, \tag{8}$$
$$\boldsymbol{x} = g(\boldsymbol{x}_{raw}, 1), \tag{9}$$
$$\boldsymbol{z} = g(\boldsymbol{z}_{raw}, 1), \tag{10}$$
$$\boldsymbol{z}_{raw} = He_j(g(\boldsymbol{z}, 2)), \tag{11}$$

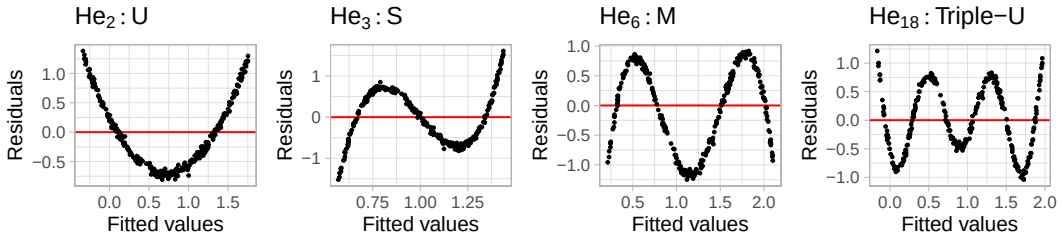


Figure 3. Polynomial forms generated for the residual plots used in experiment I. The four shapes are generated by varying the order of polynomial given by j in $He_j(.)$.

Table 2. Description of all factors involved in the non-linear and heteroskedasticity studies.

Poly Order (j)	Distribution of X_{raw}	SD (σ)	Heteroskedasticity Shape (a)	Heteroskedasticity (b)	Size (n)
2	$U(-1,1)$	0.25	-1	0.25	50
3	$N(0,0.3^2)$	1.00	0	1.00	100
6	$lognormal(0,0.6^2)/3$	2.00	1	4.00	300
18	$U\{1,5\}$	4.00		16.00	
				64.00	

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 $\{-k, k\}$ defined as

$$g(\boldsymbol{x}, k) = (\boldsymbol{x} - min(\boldsymbol{x}))/max(\boldsymbol{x} - min(\boldsymbol{x}))2k - k, \quad \text{for} \quad k > 0. \tag{12}$$

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 Z_{raw}, the function `hermite` from the R package `mpoly` (Kahle 2013) was used to generate Hermite polynomials.

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

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{u}, \tag{13}$$

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 result in model misspecification. Visual patterns of non-linearity were simulated using four different order of probabilist's Hermite polynomials ($j = 2, 3, 6, 18$) and four different distribution of X_{raw}: (1) $U(-1,1)$, (2) $N(0,0.3^2)$, (3) $lognormal(0,0.6^2)/3$ and (4) $u\{1,5\}$. A summary of the factors is given in Table 2.

The values of j was chosen so that distinct shapes of non-linearity were included in the residual plot. A greater value of j will result in a curve with more turning points. As shown in Figure 3, it includes "U" shape, "S" shape, "M" shape and "Triple-U" shape. 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.

Four different distribution were used to generate X_{raw} as shown in Figure 4. The uniform and the normal distribution are symmetric and commonly assumed in statis-

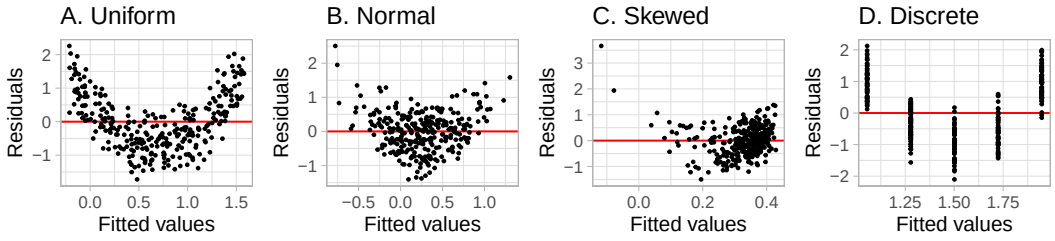


Figure 4. Variations in fitted values, that might affect perception of residual plots. Four different distribution of X_{raw} are used in the experiment to provide various visual patterns.

tical models. The adjusted log-normal distribution provides skewed density, while the discrete uniform distribution provides discreteness in residual plot, which could enrich the pool of visual patterns.

Figure 5 shows one of the lineups used in experiment I. This lineup is produced under $j = 6$ and $X_{raw} \sim N(0.0.3^2)$. The actual data plot location is $2^3 - 4$. All five subjects correctly identified the actual data plot for this lineup.

3.1.2. Heteroskedasticity

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

$$\boldsymbol{y} = 1 + \boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{14}$$

$$\boldsymbol{x} = g(\boldsymbol{x}_{raw}, 1) \tag{15}$$

$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, 1 + 2 - |a|b(\boldsymbol{x} - a)^2\boldsymbol{I}), \tag{16}$$

$$\tag{17}$$

where $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{\varepsilon}$ are vectors of size n and $g(.)$ is the scaling function defined in 12.

The null regression model used to fit the realizations generated by the above model is formulated exactly the same as Equation 13.

For $b \neq 0$, the variance-covariance matrix of the error term $\boldsymbol{\varepsilon}$ is correlated with the regressor $\boldsymbol{x}$, which will lead to the presence of heteroskedasticity. Visual patterns were simulated using three different shapes (a = -1, 0, 1) and the same four different distribution of X_{raw} used in experiment I. A summary of the factors is given in Table 2.

The values of a was chosen so that different shapes of heteroskedasticity were included in the residual plot. These include left-triangle shape, butterfly shape and right-triangle shape as displayed in Figure 6.

An example lineup of this model is shown in Figure 7 with $a = -1$ and $X_{raw} \sim U(-1, 1)$. The actual data plot location is $2^4 + 2$. Nine out of 11 subjects correctly identified the actual data plot for this lineup.

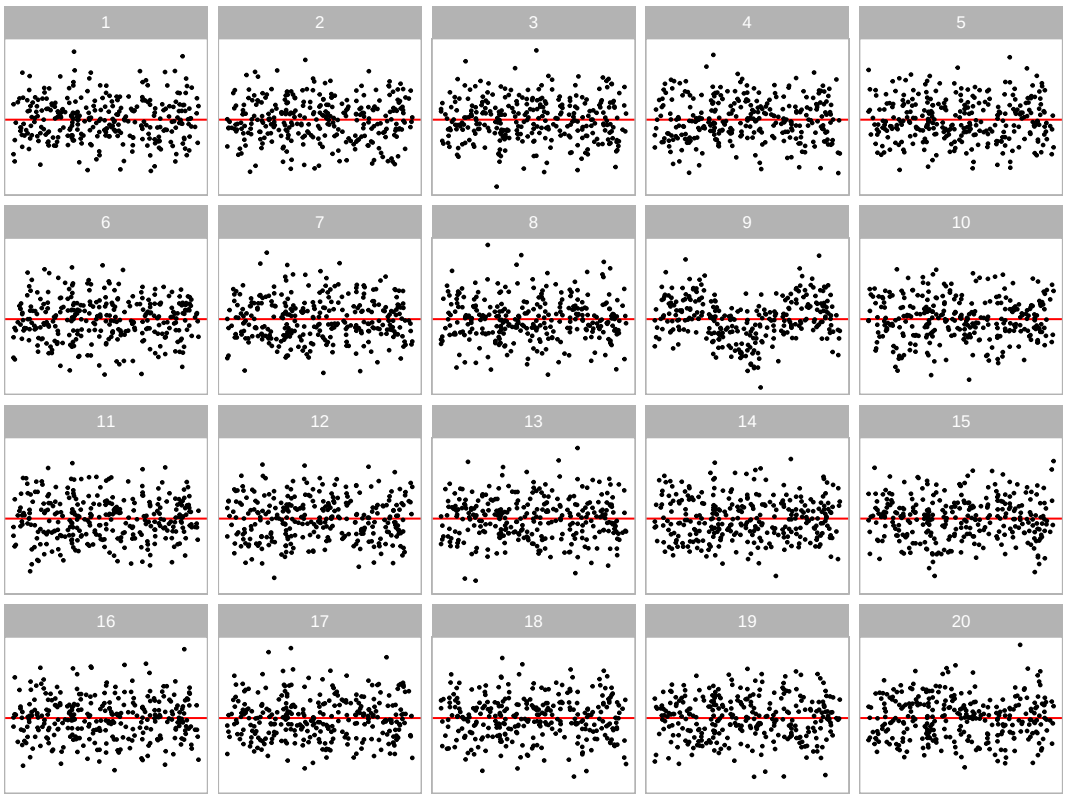


Figure 5. Lineup poly-24 in experiment I. Can you spot the most different plot?

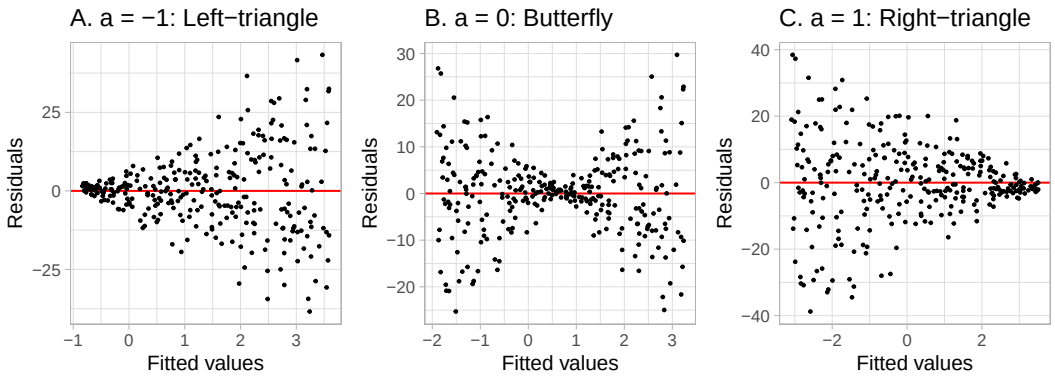


Figure 6. Heteroskedasticity forms used in experiment II. Three different shapes (a = -1, 0, 1) are used in the experiment to create left-triangle, butterfly and right-triangle shapes.

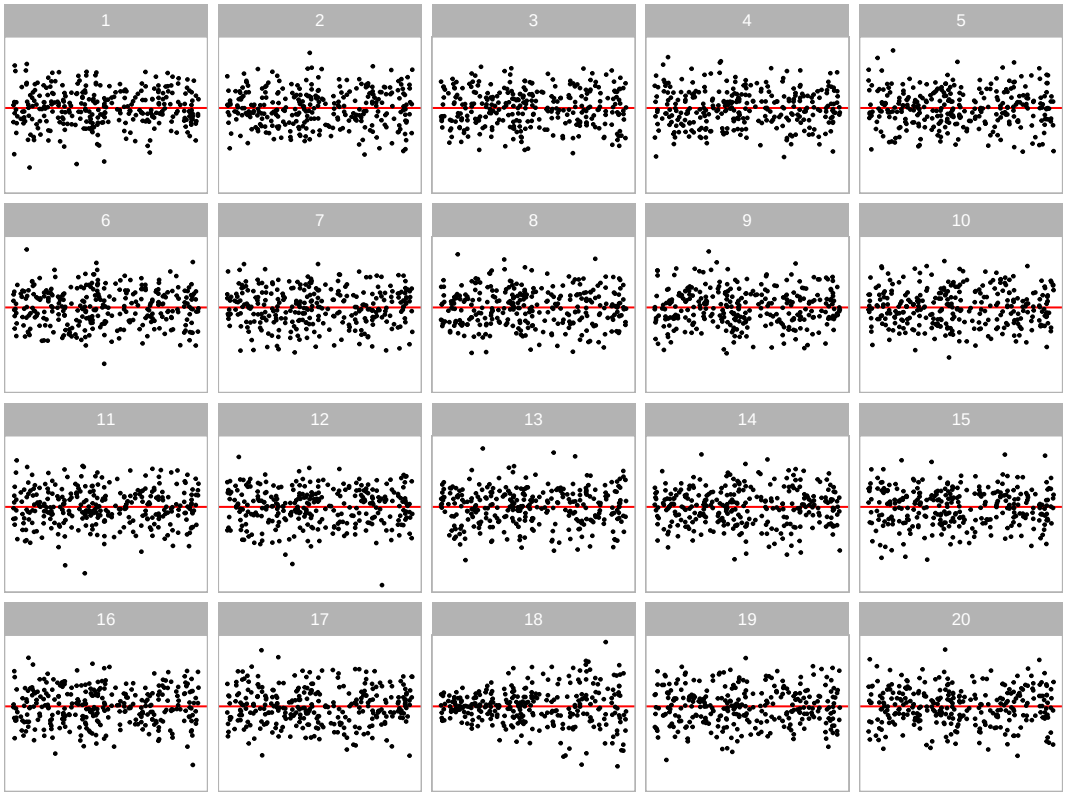


Figure 7. Lineup heter-169 in experiment II. Can you spot the most different plot?

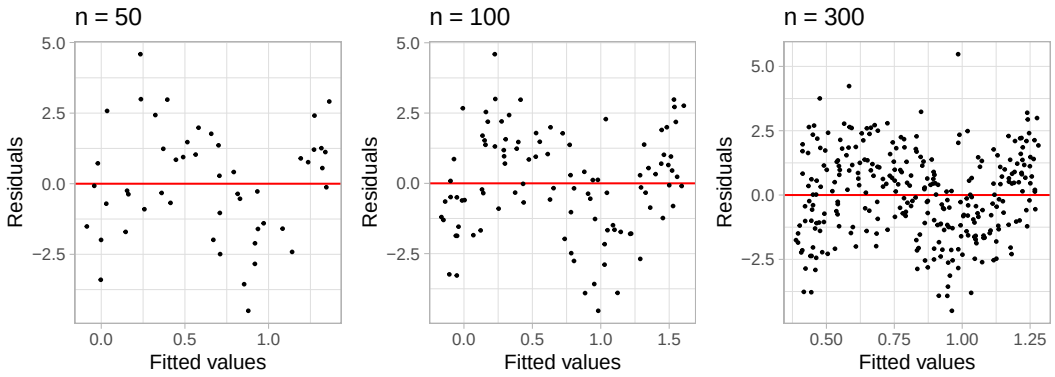


Figure 8. Three different values of n are used in experiment I, II and III to control the strength of the signal.

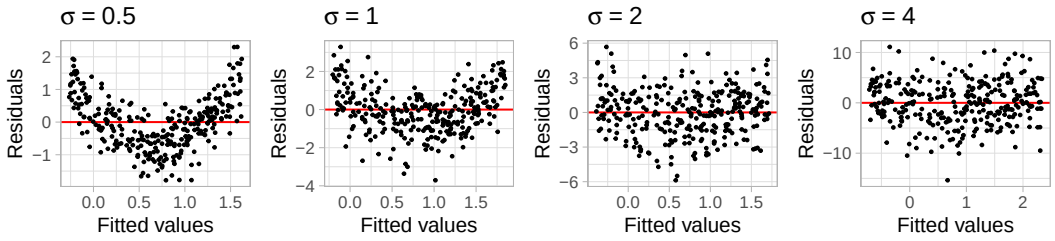


Figure 9. Four different values of σ are used in the experiment I to control the strength of the signal.

3.2. *Experimental setup*

3.2.1. Controlling the strength of the signal

As summarised in Table 2, three additional parameters n, σ and b were used to control the strength of the signal so that different difficulty levels of lineups were generated, and therefore, the estimated power curve would be smooth and continuous. Parameter $\sigma \in \{0.5, 1, 2, 4\}$ and $b \in \{0.25, 1, 4, 16, 64\}$ were used in experiment I and II respectively. Figure 9 and 10 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 ratio of the standard deviation of the heteroskedasticity across the domain of the regressor. Given $x \neq a$, a larger value of b will lead to a larger ratio of the variance at x to the variance at $x - a = 0$, making the visual pattern more obvious.

Three different sample sizes were used (n = 50, 100, 300) in all three experiments. It can be observed from Figure 8 that with fewer data points drawn in a residual plot, the visual pattern is more difficult to be detected.

3.2.2. Subject allocation

Three replications were made for each of the parameter values shown in Table 2 resulting in $(4 \times 4 \times 4 \times 3 + 4 \times 3 \times 5 \times 3) \times 3 = 1116$ different lineups. In addition, each lineup was designed to be evaluated by five different subjects. After attempting some pilot studies internally in our research group, we decided to present a block of 20 lineups to every subject. And to ensure the quality of the survey data, two lineups with obvious

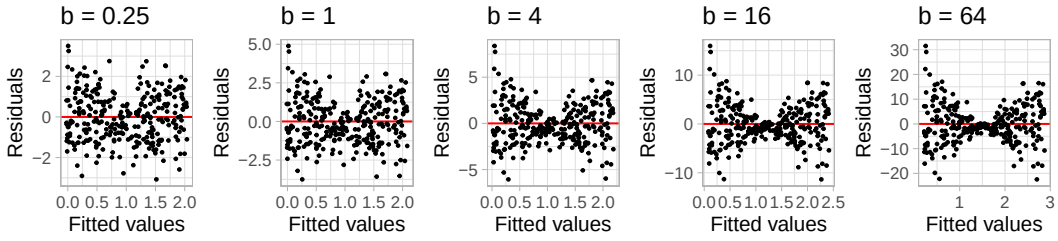


Figure 10. Five different values of b are used in experiment II to control the strength of the signal.

visual patterns were included as attention checks. Thus, $576 \times 5 \div (20 - 2) = 160$ and $540 \times 5 \div (20 - 2) = 150$ subjects were recruited to satisfy the design of the experiment I and experiment II respectively.

As mentioned in Section 2.3.3, α used in Equation 1 needs to be estimated using Rorschach lineups. Hence, 36 lineups with all combinations of n and X_{raw} and three replications were included in experiment III. In these lineups, the data of the data plot was generated from a model consistent with the null hypothesis, while the data of the 19 null plots were generated using the same simulation method discussed in Section 2.3.2. Technically, this was an alternative to Rorschach lineup but with the null distribution further conditioning on the estimated coefficients and the residual sum of squares. Thus, we used another valid method suggested in VanderPlas et al. (2021) which was designed for typical lineups to estimate α. Considering the differnece between typical lineups and Rorschach lineups, we have also included a sensitivity analysis in the Appendix to examine the impact of the variance of the 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. We planned to also collect 6 more evaluations on the 279 lineups with $X_{raw} \sim U(-1, 1)$, resulted in $(36 \times 20 + (4 \times 4 \times 3 + 3 \times 5 \times 3) \times 3) \div (20 - 2) = 133$ subjects recruited for experiment III.

3.2.3. Collecting results

Subjects for all three experiments were recruited from an crowdsourcing platform called Prolific [palan2018prolific]. Prescreening procedure was applied during the recruitment, subjects were required to be fluent in English, with 98% minimum approval rate in other studies and 10 minimum submissions. During the experiment, every subject was presented with a block of 20 lineups. For each lineup, the actual data plot was drawn as a standard residual plot of the null model with raw residuals on the y-axis and fitted values on the x-axis. An additional horizontal red line was added at $y = 0$ as a helping line. The 19 null datasets were generated by the residual rotation technique, and plotted in the same way. The lineup consisted of 20 residual plots with one randomly placed actual data plot. And for every lineup, the subject was 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 were from others. If there was no noticeable difference between plots in a lineup, subjects were permitted to select zero plots without providing the reason. No subject was shown the same lineup twice. Information about preferred pronoun, age group, education, and previous experience in visual experiment were also collected. A subject's submission

was only accepted if the actual data plot was identified for at least one attention check. Data of rejected submissions were discarded automatically to maintain the overall data quality.

4. Results

4.1. *Overview*

There were 2880, 2880 and 2214 lineups evaluation made by 160, 160 and 123 subjects recruited for experiment I, II and III respectively. In the total of 7974 lineup evaluations, 3744 used lineups produced by the non-linearity model, 3510 used lineups produced by the heteroskedasticity model, and 720 used lineups consists of only null plots.

Besides, there were 886 attention checks not included in the following analysis. The collated dataset is provided in `vi_survey` of the `visage` `R` package.

4.2. *Comparison of the power of different tests*

(discuss choice of tests, mention pkg `lmtest`, `skedastic`)

Figure 11 shows the estimated power of visual test on lineups produced by the non-linearity model with $X_{raw} \sim U(-1, 1)$, against the natural logarithm of the effect $log_e(E)$.

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

To compare the power of visual test and conventional tests, we picked 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`. These three tests can be used to detect non-linearity, heteroskedasticity and non-normality respectively. Their estimated power is shown in Figure 11. For the RESET test, we included different power of fitted values as proxies. According to (Ramsey 1969), there is no general rule for the power of the fitted values needed by the RESET test, but it found power up to four is usually sufficient. Thus, we followed this guideline to conduct the RESET test. For the Breusch-Pagan test, both $\boldsymbol{x}$ and $\boldsymbol{x}^2$ were included in the auxiliary regression equation to [justify this choice?].

At the bottom of the figure 11, there are a sequence of example residuals 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.

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 the non-linearity model, the power of all four tests increases as the effect gets larger. The power curve of RESET test climbs aggressively from 13% to around 70% as $log_e(E)$ increases from 0 to 2, while power of other tests respond inactively to the

change of effect and remain lower than 25% throughout the period, showing that F-test is way more sensitive to the type of model defects that being considered. Meanwhile, no noticeable visual features can be spotted from the example residuals plots.

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

The power of visual test arises steadily as $log_e(E)$ increases from 2 to 5 in both non-linearity model and heteroskedasticity model, 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 residuals plots that when $log_e(E) = 2.5$, a weak "S-shape" and a weak "triangle" shape are presented in Figure 11 and Figure 12 respectively. And as $log_e(E)$ increases, the visual pattern becomes much clearer, and the power reaches almost 100% at $log_e(E) \approx 6$.

The power of all inappropriate tests except for F-test shows improvement as the effect increases but at a lower rate than the visual test in both scenarios. This coincides the point made by Cook and Weisberg (1982) that residual-based tests for a specific type of model defect may be sensitive to other types of model defects. The power curve of F-test remains at around 5% in Figure 12 since there is no non-linear term leave out in the heteroskedasticity model and H_0 of the F-test will be always satisfied.

Overall, the power analysis suggests that conventional tests do not match how humans perceive residual plots and thus determine model violations in these two scenarios. Non-severe model violations with human acceptable residual departures could often alarm by the conventional tests. This could limit the popularity of conventional tests in residual diagnostics among analysts.

4.3. *Comparison of rejection rates based on p-values*

(style similar to mosaic plot)

(lineups detected by visual test but not by conventional tests)

Visual inference and conventional hypothesis testing are two distinct statistical procedures. As displayed in Figure 13, the visual test gives higher p-values than the conventional test for both polynomial and heteroskedasticity models in most of the time, suggesting that humans have lower confidence and much higher tolerance of the residual departures than the conventional test.

In particular, given 95% confidence level, both tests will reject H_0 in only around 58% of the cases. And in 41% of the cases, the visual test does not reject H_0 but the conventional test does. Since fail to reject H_0 in a visual test means that there is no obvious visual discoveries found in the residual plot, analysts and the general public as the consumers of the output may not fully believe in the rejection of H_0 even though the alternative hypothesis is true. Even if the rejection is accepted, the model violation may not be considered as impactful due to the fact that the departures are not clearly visible.

This makes it sounds like people are arrogantly ignoring the accuracy of conventional tests and trust too much to their visual perception, but it is not. In fact, the sensitivity of the conventional test could sometimes distract and discourage analysts from finding a simple but good linear approximation to the data. The rejection of H_0 because of acceptable and negligible residual departure is not practically meaningful and useful.

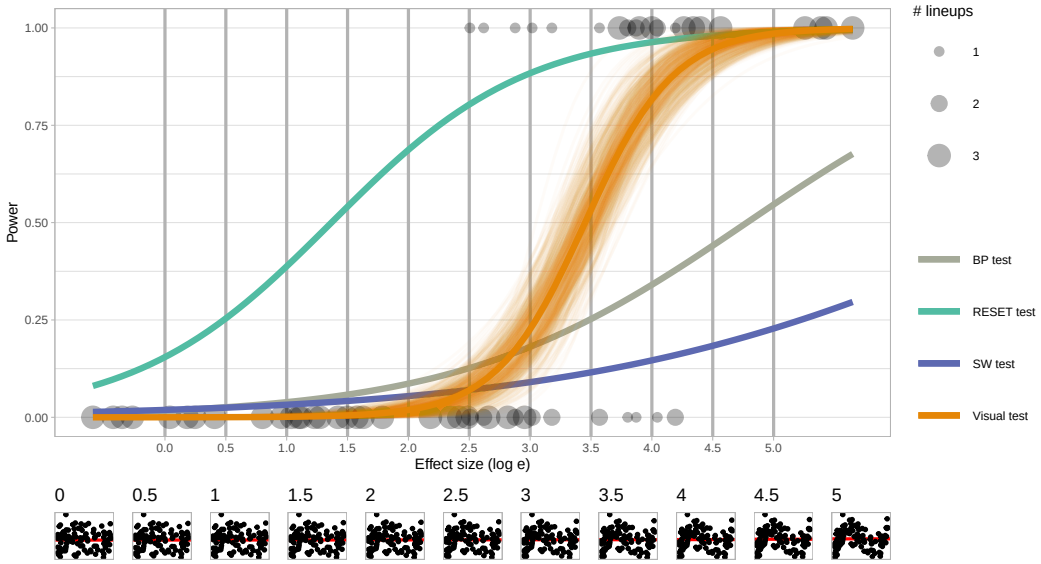


Figure 11. Comparison of power between different tests for non-linear patterns. Main plot shows the power curves, with dots indicating human evaluations of lineups. Small row of plots shows typical residual plots corresponding to specific effect sizes, marked by dashed lines in main plot. Where would you draw the line of too much non-linearity in the residuals? For the RESET test this is around log effect size 1, but for the visual test it is around 3.

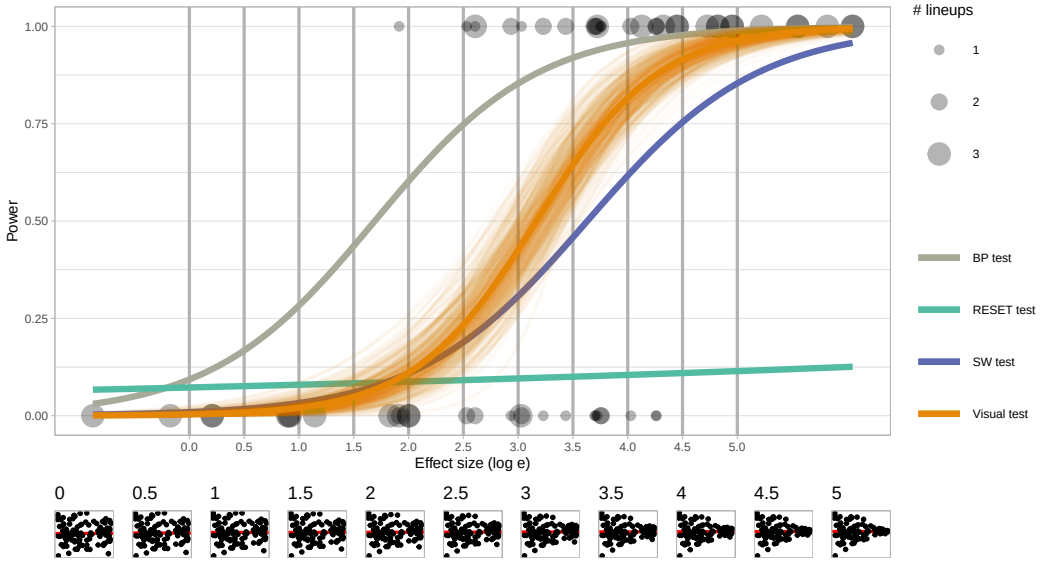


Figure 12. Comparison of power between different tests for heteroskedasticity patterns. Main plot shows the power curves, with dots indicating human evaluations of lineups. Small row of plots shows typical residual plots corresponding to specific effect sizes, marked by dashed lines in main plot. Where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

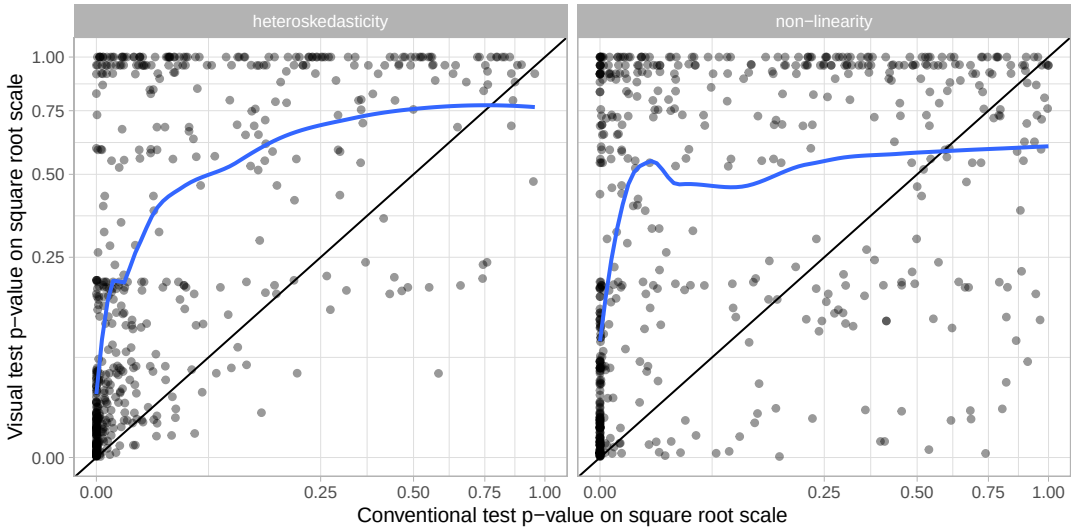
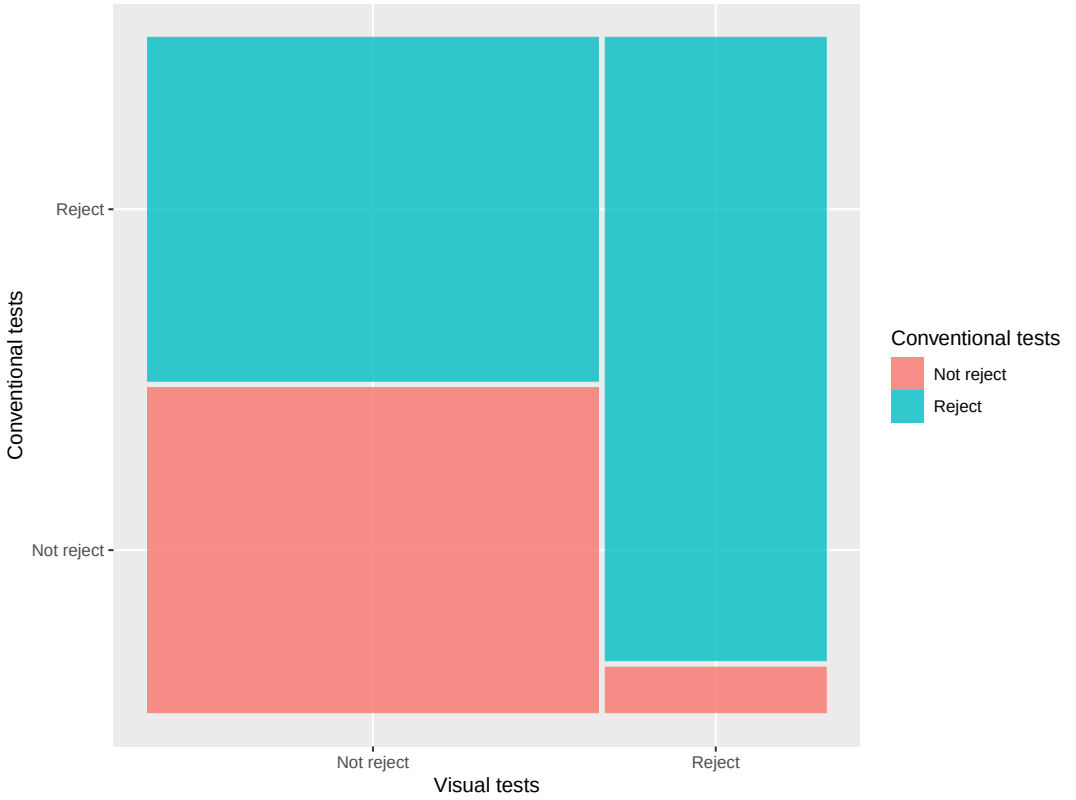


Figure 13. Visual test p-value compared to conventional test p-value for lineups produced by heteroskedasticity and non-linearity model. The scatter plot is drawn on a square root scale. Every dot on the plot corresponds to a lineup. Most of the dots are fallen on the left of the y = x line indicating visual test p-value is almost always higher than conventional test p-value. The blue curve is a smoothing of the dots showing a positive trend.

This also partly explains why abundance of literature suggesting the use of data plot in regression diagnostics.



4.4. *Effect of the distribution of the independent variable*

5. change of x_dist -> power

6. appendix (contains all the details)

6.0.1. A collection interesting lineups (unusual results)

why unusual? what is the possible explanations?

7. targe journal

JRSSB: Journal of the Royal Statistical Society Series B (Statistical Methodology)
Deadline: Jan 1, Apr 1

Reading: style of writing, author guideline (`https://rss.onlinelibrary.wiley.com/hub/journal/14679868/author-guidelines`)

JCGS JDSS

References

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

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

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

Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne, and Hadley Wickham. 2009. "Statistical inference for exploratory data analysis and model diagnostics." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367 (1906): 4361–4383.

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

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

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

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

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

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

Gelman, Andrew. 2004. "Exploratory Data Analysis for Complex Models." *Journal of Computational and Graphical Statistics* 13 (4): 755–779.

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

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

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

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

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

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

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

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

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

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

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

Shapiro, Samuel Sanford, and Martin B Wilk. 1965. "An analysis of variance test for normality (complete samples)." *Biometrika* 52 (3/4): 591–611.

Silvey, Samuel D. 1959. "The Lagrangian multiplier test." *The Annals of Mathematical Statistics* 30 (2): 389–407.

VanderPlas, Susan, Christian Röttger, Dianne Cook, and Heike Hofmann. 2021. "Statistical significance calculations for scenarios in visual inference." *Stat* 10 (1): e337.

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

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