ARTICLE TEMPLATE

Why aren't significance tests commonly used for linear regression diagnostics?

Weihao Li[a], Dianne Cook[a], Emi Tanaka[a]

[a]Department of Econometrics and Business Statistics, Monash University, Clayton, VIC, Australia

ARTICLE HISTORY
Compiled September 7, 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, it is typical to interrogate the residuals. Residuals summarise what is not captured by the model, and thus provide the capacity to identify what might be wrong. There are many ways that residuals could be assessed.

Residuals might be plotted, as a histogram or quantile-quantile plot to examine the distribution. Using the classical 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).

Plotting the residuals against predicted values and each of the explanatory variables on a scatter plot is a recommend way to scrutinize their relationships. If there is any visually discoverable patterns, the model is potentially misspecified. However, it is a very difficult task for a human judge, though to make a decision that there's nothing there. It is especially common, particularly among new data analysts to report patterns when an experienced data analyst might quickly conclude that there are none. Generally, one looks for departures from nothingness 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 and Pagan 1979) for heteroskedasticity.

CONTACT Weihao Li. Email: weihao.li@monash.edu, Dianne Cook. Email: dicook@monash.edu, Emi Tanaka. Email: emi.tanaka@monash.edu

There is an abundance of literature describing 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.* There is a common guidance by experts that plots are the best for diagnosing model fits.

This 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 the types of departures that one expects to detect, and describes a formal process for reading residual plots, called visual inference, that can avoid the concerns about subjectiveness of human readers. Section 3 describes the experimental setup to enable a comparison between decision made by formal hypothesis testing, and how humans would read diagnostic plots. The results are reported in Section **??**. 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 predictor variables that are approximately orthogonal to residuals are referred to as standard residual plots in Cook and Weisberg (1982). As shown in Figure [ref here], 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. We will only discuss three most commonly checked departures in this paper. Namely, non-linearity, heterskedasticity and non-normality.

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. At the top-right of Figure [ref here], there is a residual plot giving an example of the visual pattern of non-linearity when a cubic term is not captured by the regression model.

Heterskedasticity refers to the presence of nonconst 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 heterskedasticity on a residual plot is the inconsistent spread of the residuals at different x values. Visually, it sometimes results in the so-called "butterfly" shape as shown in the bottom-left panel of Figure [ref here], or the "left-triangle" and "right-triganle" shape where the smallest variance occurs at the edges of the x-axis.

Non-normality is usually harder to be detected from a residual plot compared to non-linearity and heterskedasticity. Besides, a quantile-quantile plot can often do a better job for this task. Note that not all regression models assume normality of the error term, but a certain amount of them does. If this assumption happens to be false, then it is expected to see data points do not distribute noramlly on the y-axis. For example, given a skewed error distribution, one will see fewer data points and more

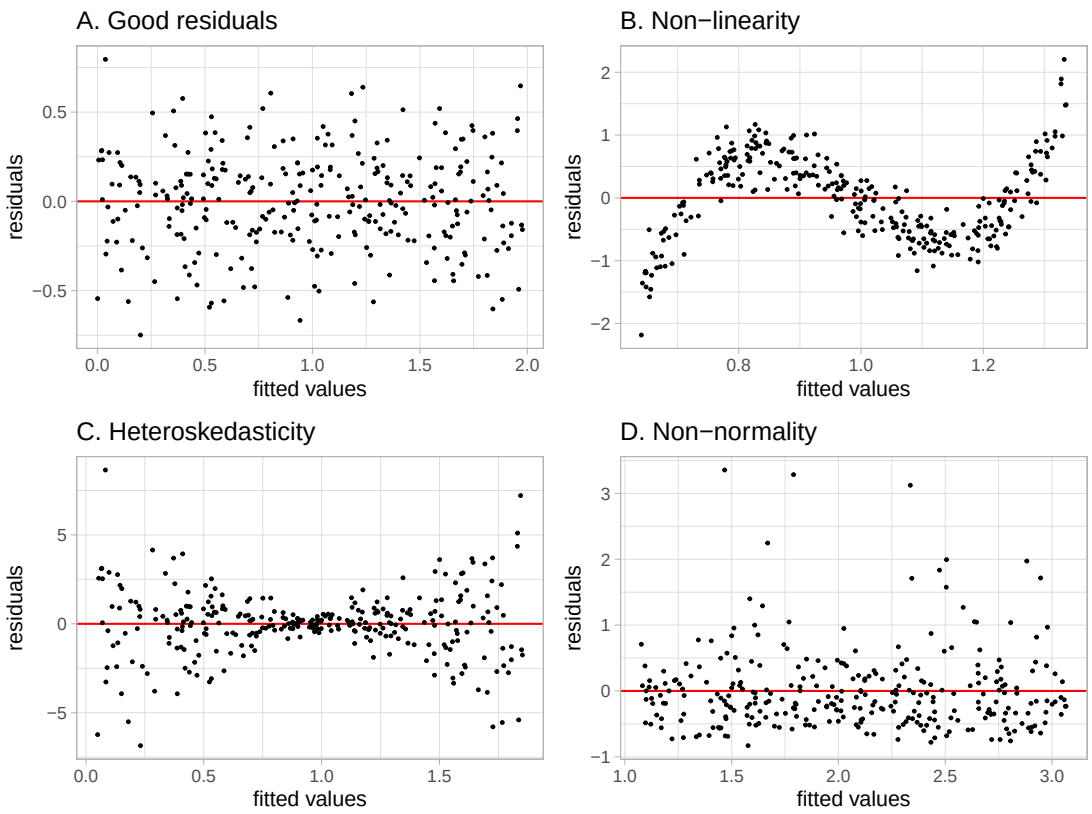
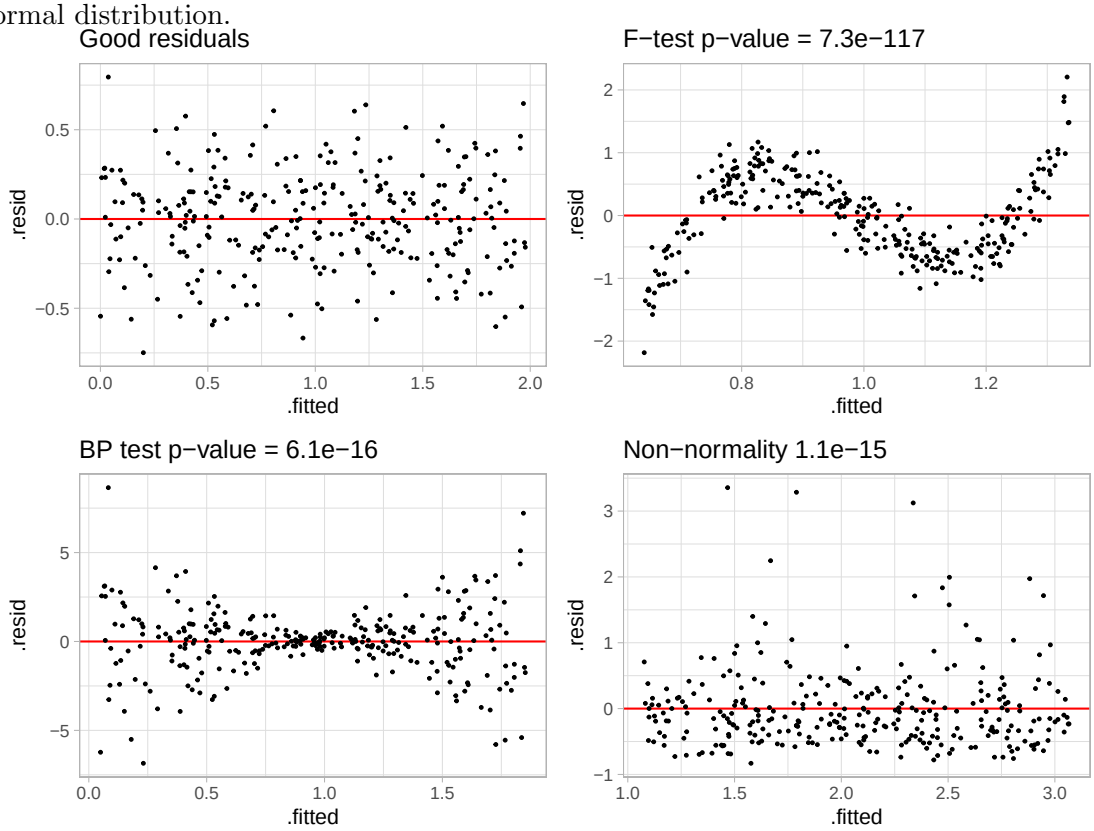


Figure 1. Example fitted vs residual plots: (A) classically good looking residuals, (B) non-linear pattern indicates that the model has not captured a non-linear association, (C) heteroskedasticity indicating that variance around the fitted model is not uniform, and (D) non-normality where the residual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

outliers on one side of the zeroline as shown in the bottom-right of Figure [ref here].

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_heteroskedasticity-consistent_1980] or the Breusch-Pagan test [@breusch_simple_1979], which are both derived from the Lagrange multiplier test [@silvey1959lagrangian] principle that relies on the asymptotic properties of the null distribution. For testing non-linearity, one may apply the F-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_tests_1969]. And for testing normality, the shapiro–Wilk test [ref here] is perhaps the most widely used test reported by many of the statistical softwares. Another choice will be the Jarque–Bera test [ref here] which directly checks if the sample skewness and kurtosis match a normal distribution.



Explain the tests you are showing later

- RESET
- BP
- SW

and show the results for the residual plots displayed in previous section
Briefly mention any others

2.3. *Visual testing for departures*

Unlike hypothesis testing built upon rigorous statistical procedures, reading diagnostic plots relies on graphical perception - human's ability to interpret and decode the information embedded in the graph [@cleveland_graphical_1984], which is to some extent subjective and indecisive. Further, 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 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_using_2015].

- Very briefly explain what a lineup test is
- Lineup (maybe using your previous residual plot for nonlinearity)
- p-value calulcation, briefly, including how multiple selections are handled
- power calculation, briefly, including how multiple selections are handled

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, in the Bayesian context, data plots was systematically considered as model diagnostics by taking advantage of the data simulated from the assumed statistical models (Gelman 2003, 2004).

It was surprising that the essential components of visual inference had actually been established in Buja, Cook, and Swayne (1999), but it was not until 10 years later that Buja et al. (2009) formalized it as an inferential framework to extend confirmatory statistics to visual discoveries. This framework redefines the test statistics, tests, null distribution, significance levels and p-value for visual discovery modelled on the confirmatory statistical testing. Figure **??** outlines the parallelism between conventional tests and visual discovery.

In visual inference, a collection of test statistics $T^{(i)}(\mathbf{y})$ $(i \in I)$ is defined, where $\mathbf{y}$ is the data and I is a set of all possible visual features. Buja et al. (2009) described each of the test statistics $T^{(i)}(\mathbf{y})$ as a measurement of the degree of presence of a visual feature. Alternatively, Majumder, Hofmann, and Cook (2013) avoids the use of visual features and defined the visual statistics $T(.)$ as a mapping from a dataset to a data plot. Both definitions of visual test statistics are valid, but in the rest of the paper the first definition will be used as it covers some details needed by the following discussion. A visual discovery is defined as a rejection of a null hypothesis, and the same null hypothesis can be rejected by many different visual discoveries (Buja et al. 2009). For regression diagnostics, the null hypothesis would be the assumed model, while the visual discoveries would be any findings that are inconsistent with the null hypothesis. The same regression model can be rejected by many reasons with residual plot, including non-linearity and heteroskedasticity as shown in Figure 2.

2.3.1. Sampling from the null distribution

The null distribution of plots refers to the infinite collection of plots of null datasets sampled from H_0. It is defined as the analogue of the null distribution of test statistics in conventional test (Buja et al. 2009). In practice, a finite number of plots of null

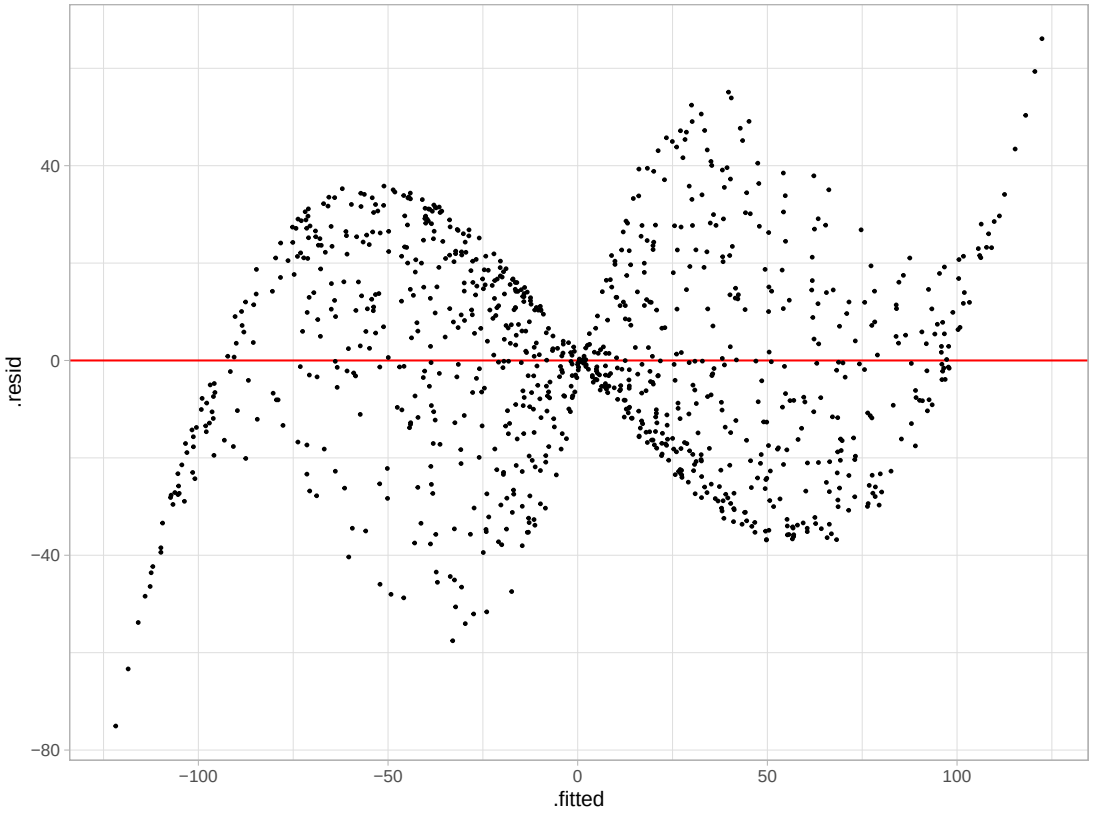


Figure 2. Residuals vs. fitted values plot for a classical linear regression model. The residuals are produced by fitting a two-predictor multiple linear regression model with data generated from a cubic linear model. From the residual plot, "butterfly shape" can be observed which generally would be interpretd as evidence of heteroskedasticity. Further, from the outline of the shape, nonlinear patterns exist. Both visual discoveries are evidence against the null hypothesis, though heteroskedasticity actually does not exist in the data generating process.

datasets could be generated, called null plots. 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). Suppose there exists a minimal sufficient statistic $\boldsymbol{S}(\boldsymbol{y})$ under the null hypothesis, any null datasets $\boldsymbol{y^*}$ should fulfil the condition $\boldsymbol{S}(\boldsymbol{y}) = \boldsymbol{s}$. Using the classical normal linear regression model as example, the minimal sufficient statistic is $\boldsymbol{S}(\boldsymbol{y}) = (\hat{\boldsymbol{\beta}}, \boldsymbol{e'e})$, where $\hat{\boldsymbol{\beta}}$ are the coefficient estimators and $\boldsymbol{e'e}$ is the residual sum of square. Alternatively, the minimal sufficient statistic can be constructed as $\boldsymbol{S}(\boldsymbol{y}) = (\hat{\boldsymbol{y}}, ||\boldsymbol{e}||)$, where $\hat{\boldsymbol{y}}$ are the fitted values and $||\boldsymbol{e}||$ is the length of residuals, which is more intuitive as suggested by Buja et al. (2009). Since the fitted values are held fixed, the variation can only occur in the residual space. And because the length of residual is also held fixed, residuals obtained from a null dataset has to be a random rotation of $\boldsymbol{e}$ in the residual space. With this property, 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.

2.3.2. Lineup protocol

With the simulation mechanism of null plots being provided, another aspect of hypothesis testing that needs to be addressed is the control of false positive rate or Type I error. Any visual statistic $T^{(i)}(\mathbf{y})$ needs to pair with a critical value $c^{(i)}$ to form a hypothesis test. When a visual feature i is discovered by the observer from a plot, the corresponding visual statistic $T^{(i)}(\mathbf{y})$ may not be known as there is no general agreement on the measurement of the degree of presence of a visual feature. It is only the event that $T^{(i)}(\mathbf{y}) > c^{(i)}$ is confirmed. Similarly, if any visual discovery is found by the observer, we say, there exists $i \in I : \ T^{(i)}(\mathbf{y}) > c^{(i)}$ (Buja et al. 2009).

Using the above definition, the family-wise Type I error can be controlled if one can provide the collection of critical values $c^{(i)}$ $(i \in I)$ such that $P($there exists $i \in I : T^{(i)}(\mathbf{y}) > c^{(i)}|\mathbf{y}) \leq \alpha$, where α is the significance level. However, since the quantity of $T^{(i)}(\mathbf{y})$ may not be known, such collection of critical values can not be provided.

Buja et al. (2009) proposed the lineup protocol as a visual test to calibrate the Type I error issue without the specification of $c^{(i)}$ $(i \in I)$. It is 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 the null hypothesis. 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 the null hypothesis, 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 the null hypothesis as the observer correctly picks the actual data plot, then the Type I error of this test is $1/m$.

This provides us with an mechanism to control the Type I error, because m - the

number of plots in a lineup can be chosen. A larger value of m will result in a smaller Type I error, but the limit to the value of m depends on the number of plots a human is willing to view (Buja et al. 2009). Typically, m will be set to 20 which is equivalent to set $\alpha = 0.05$, a general choice of significance level for conventional testing among statisticians.

Further, a visual test can involve K independent observers. Let $D_i = \{0, 1\}$ 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 the null hypothesis, $X \sim \text{Binom}_{K,1/m}$. Therefore, the p-value of a lineup of size m evaluated by K observer is given as

$$P(X \geq x) = \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right)^{i} \left(\frac{m-1}{m}\right)^{k-i}, \tag{1}$$

where x is the realization of number of observers correctly picking the actual data plot (Majumder, Hofmann, and Cook 2013).

The multiple individuals approach avoids the limit of m, while provides visual tests with p-value much smaller than 0.05. In fact, the lower bound of p-value decreases exponentially as K increases. With just 4 individuals and 20 data plots in a lineup, the p-value could be as small as 0.0001. Additionally, by involving multiple observers, variation of individual ability to read plots can be addressed to some degree as different opinions about visual discoveries can be collected.

As pointed out by VanderPlas et al. (2021), though Equation (1) is trivial, but it 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. They summarized three common different 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 experimental design. For Scenario 3, VanderPlas et al. (2021) modelled 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, they decided to add $1/s_j$ to c_i instead of one, where s_j is the number of plots subject j selected for $j = 1, ...K$. This ensured $\sum_i c_i = K$.

The full model was a Dirichlet-multinomial mixture distribution

$$\begin{aligned} \boldsymbol{\theta}|\alpha &\sim Dirichlet(\alpha) \\ (c_i, ..., c_m)|\boldsymbol{\theta} &\sim Multinomial(K, \boldsymbol{\theta}). \end{aligned} \tag{2}$$

Since the p-value calculation only needs to concern the number of times the actual data plot being selected denoted by C_i, they showed the model can be simplified to a beta-binomial mixture distribution

$$\begin{aligned} \theta_i|\alpha &\sim Beta(\alpha, (m-1)\alpha) \\ C_i|\theta_i &\sim Binomial(K, \theta_i). \end{aligned} \tag{3}$$

Thus, the visual p-value followed by the beta-binomial model is given as

$$P(C \geq c_i) = \sum_{x=c_i}^{K} \binom{K}{x} \frac{1}{B(\alpha, (m-1)\alpha)} B(x+\alpha, K-x+(m-1)\alpha), \tag{4}$$

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

$$B(a,b) = \int_0^1 t^{\alpha-1}(1-t)^{b-1}dt, \quad \text{where} \quad a, b > 0. \tag{5}$$

Note that the use of Equation (4) requires the estimation of $\hat{\alpha}$, which largely depends on the null model, the type of the plot and other aesthetic features. They suggested to estimate $\hat{\alpha}$ visually based on the selections of null plots of the experimental data, or to estimate $\hat{\alpha}$ numerically based on several additional Rorschach lineups, which is a type of lineup containing only null plots. However, when the number of null models are large, it could be expensive to manually estimate each α or include additional Rorschach lineups in the experiment.

Instead, in the experiments that will be described in section 3, we adopt a simpler model implicitly used by the `pmulti()` function of the `vinference R` package. We assume the attractiveness of the plot i modelled as $w_i \sim Uniform(0,1)$ for $i = 1, .., m$. Let $\theta_i = w_i / \sum_{i=1}^{m} w_i$ be the probability of plot i being selected by a subject. Then, given the number of selections s_j, for $j = 1, ..., K$, the distribution of C_i can be approximated by simulating the random selection process with computer. The simulated visual test p-value is formulated as

$$\text{p-value} = \frac{\#draws\ that\ the\ actual\ data\ plota\ being\ selected\ more\ than\ c_i\ times}{\#simulation}. \tag{6}$$

2.4. *Effectiveness of visual test in regression diagnostics*

The effectiveness of visual inference has already been validated by Majumder, Hofmann, and Cook (2013) under relatively simple classical normal linear regression model settings with only one or two regressors. Their results suggest visual test is capable of testing the significance of a single regressor with a similar power as a t-test, though they expressed that in general it is unnecessary to use visual inference if there exists a conventional test and they didn't 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. In fact, lineup protocol has been integrated into some model diagnostic tools such as Loy and Hofmann (2013).

With our knowledge, what haven't been examined so far is the effectiveness of visual test relative to the equivalent conventional test in regression diagnostics. Particularly, its ability to detect non-linearity and heteroskedasticity compared to F-test and BP-test.

3. Experimental design

3.1. *Motivation and overview*

3.2. *Simulating departures*

3.2.1. Nonlinearity

3.2.2. Heteroskedasticity

3.3. *Experimental setup*

3.3.1. Factors

- Treatments for nonlinearity experiment (plots summarising levels)
- Treatments for heteroskedasticity experiment (plots summarising levels)

3.3.2. Assigning subjects

3.3.3. Collecting results

software/technical

3.4. *Computing power curves*

For the purpose of examining the effectiveness of visual test in regression diagnostics, two experiments were conducted. The experiment I has ideal scenario for conventional testing, where the visual test is not expected to outperform the conventional test. 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.

Subjects for both experiments were recruited from an crowdsourcing platform called Prolific [ref here]. 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. 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 pool of 12 different lineups with obvious visual patterns were generated for experiment I and experiment II respectively. In every block of 20 lineups that presented to a subject, two out of 12 lineups were included as attention checks. A subject's submission was only accepted if the actual data plot was identified for at least one at-

tention check. Data of rejected submissions were discarded automatically to maintain the overall data quality.

3.5. *Non-linearity*

Experiment I is designed to study the ability of human subjects to detect the effect of a random vector $\boldsymbol{z}$ which is a probabilist's Hermite polynomial [Herimite ref here] of another random vector $\boldsymbol{x}$ in a two variable statistical model formulated as:

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

where $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{x}_{raw}$, $\boldsymbol{z}_{raw}$ are vector 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

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

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

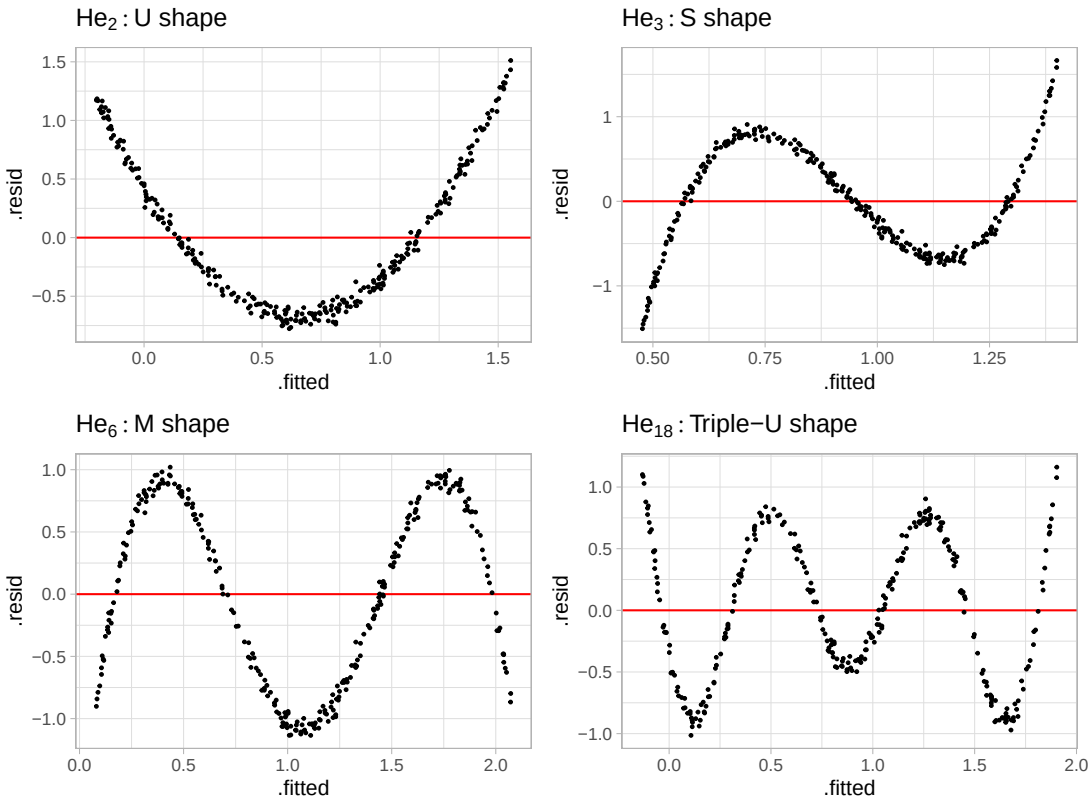
$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{u}, \tag{12}$$

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

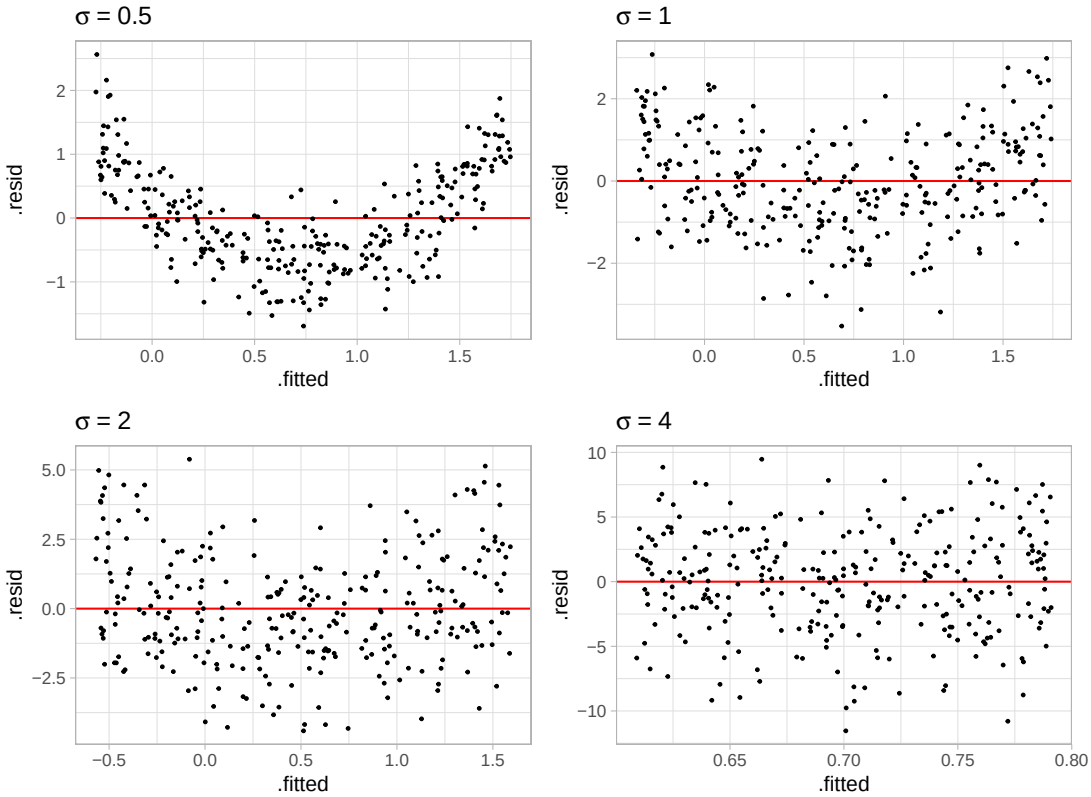
Model misspecification presents since the null model leaves out the higher order term.

Experiment data were simulated using four different order of probabilist's Hermite polynomials ($j = 2, 3, 6, 18$), three different sample sizes (n = 50, 100, 300) , four different standard deviations of the error (σ = 0.5, 1, 2, 4) 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 parameters used in this experiment is given in Table 1.

The values of j was chosen so that different shapes of non-linearity were included in the residual plot. These include "U" shape, "S" shape, "M" shape and "Triple-U" shape.



The range of σ, which is a factor controlling the strength of the signal, was chosen so that different difficulty levels of lineups were generated, and therefore, the estimated power curve would be smooth and continuous.

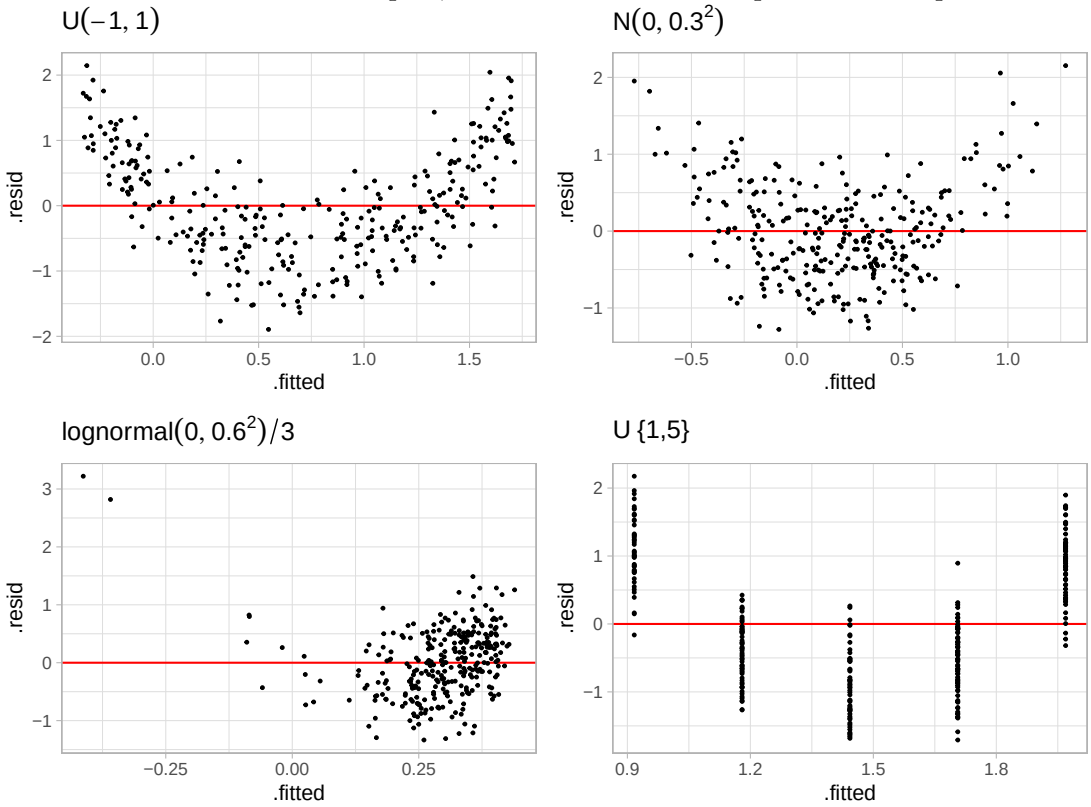


Four different distribution were used to generate X_{raw}. The uniform and the normal

Table 1. Parameter values for n, j σ, X_{raw}

Sample size (n)	Order of Hermite polynomial (j)	Error SD (σ)	Distribution of X_{raw}
50	2	0.5	$U(-1, 1)$
100	3	1.0	$N(0, 0.3^2)$
300	6	2.0	$lognormal(0, 0.6^2)/3$
	18	4.0	$U\{1, 5\}$

distribution are symmetric and commonly assumed in statistical models. The adjusted log-normal distribution provides skewed density. And the discrete uniform distribution provides discreteness in residual plot, which could enrich the pool of visual patterns.



Three replications are made for each of the parameter values shown in Table 1 resulting in 192 different 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. 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. Figure 3 is an example of one of these lineups. It was produced by using $n = 300$, $j = 6$, $\sigma = 0.5$ and $X_{raw} \sim N(0.0.3^2)$. The actual data plot location was four. All five subjects correctly identified the actual data plot for this lineup.

In addition, each lineup is designed to be evaluated by five different subjects to provide reasonable estimates of the visual p-value. Thus, $192 \times 3 \times 5/(20 - 2) = 160$ subjects were recruited to satisfy the design of the experiment I.

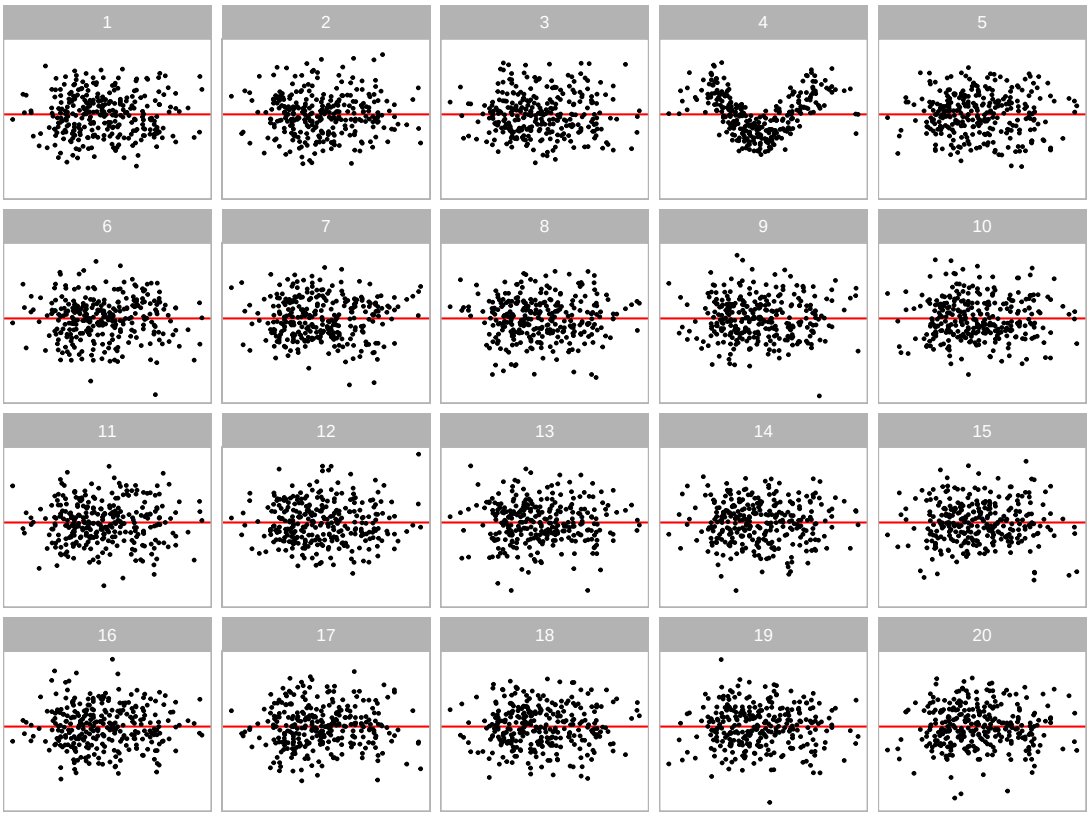


Figure 3. Example lineup

3.6. *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:

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

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, 1 + 2 - |a|b(\boldsymbol{x} - a)^2\boldsymbol{I}), \tag{14}$$

$$\tag{15}$$

where $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{\varepsilon}$ are vector of size n.

The null regression model used to fit the realizations generated by the above model is formulated exactly the same as the model used in the first experiment:

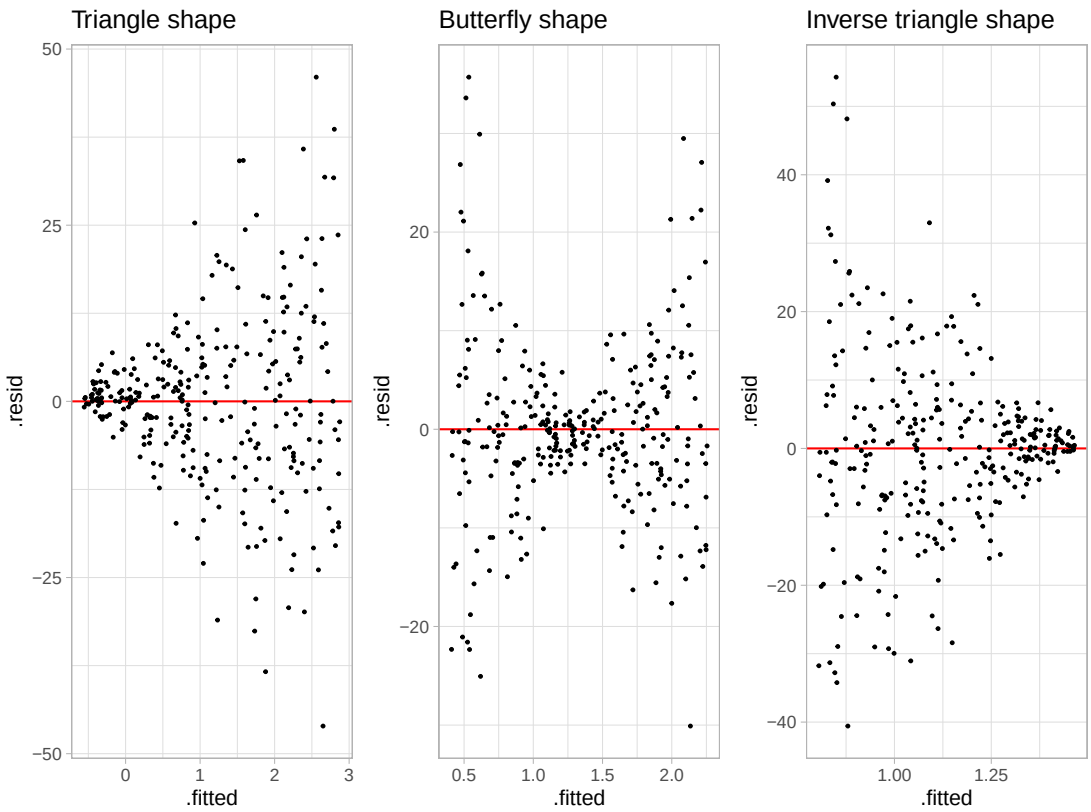
$$\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{x} + \boldsymbol{u}, \tag{16}$$

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

Model misspecification presents since the assumption about the constant error variance is violated.

Experiment data were simulated using three different shapes (a = -1, 0, 1), five different values of b (b = 0.25, 1, 4, 16, 64), three different sample sizes (n = 50, 100, 300) 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 parameters used in this experiment is given in Table 1.

The values of a was chosen so that different shapes of heteroskedasticity were included in the residual plot. These include triangle shape, butterfly shape and inverse triangle shape.



The range of b, which is a factor controlling the strength of the signal, was chosen so that different difficulty levels of lineups were generated, and therefore, the estimated power curve would be smooth and continuous.

4. Analysis and results

4.1. *Data overview*

How many people? How many lineups?

Subjects recruited from Prolific received a fixed payment for participating in the experiment. However, some subjects will try to maximize their earnings for minimum effort. During the review of submissions, if we found a subject objectively demonstrated clear low-effort throughout the experiment, i.e., failed all attention checks, we rejected the submission. The rejected submissions will be removed immediately, and Prolific will automatically recruit another subject as substitution. Therefore, we only paid for approved submissions and no further data screening procedure needed to be applied on the collected data.

A subject was allowed to select zero plots for a lineup if there was no visible difference between plots, but the simulated visual p-value given in Equation (6) will effectively drop the subject from the simulation for this case, leading to inaccurate estimation of p-value. Therefore, we treated this case as making one selection but failing to identify the actual data plot so that Equation (6) can be applied correctly.

In overall, there were a total of 3200 lineup evaluations made by 160 subjects in both experiment I and experiment II respectively, where 320 lineup evaluations were attention checks and were not used in the following analysis.

The collated dataset is provided in `polynomials` and `heter` of the `visage` `R` package.

4.2.

4.3. *Nonlinearity*

- Power curve overall
- Four lineups also shown, selected from ones close to the visual power curve from the uniform treatment
- Compare visual power against different factors

4.4. *Heteroskedasticity*

- Power curve overall
- Four lineups also shown, selected from ones close to the visual power curve from the uniform treatment
- Compare visual power against different factors

4.5. *Demographic summary*

Table 2 tabulates the number of subjects, preferred pronouns, education backgrounds, age groups, and previous experience in visual experiment. Figure 4 visualizes the marginal distribution of each of the category. The collated data was a balanced sample among male and female. Most of the ages of subject were between 18 to 39, while many of them were between 18 to 24. Very few subjects were awarded degree higher than Bachelor Degree. Around 40% of subjects had previous experience in visual experiment.

4.6. *Model fitting*

For each parameter combination, effect E is derived from the Kullback-Leibler divergence (see [appendix ref here]) formulated as:

$$E = \frac{1}{2\sigma^2} \boldsymbol{X}'_b \boldsymbol{R}'_a (diag(\boldsymbol{R}_a))^{-1} \boldsymbol{R}_a \boldsymbol{X}_b, \tag{17}$$

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

The logistic regression is fit using $log_e(E)$ as the only fixed effect covariate for the power of visual test formulated as:

$$Pr(\text{reject } H_0 | H_1, E) = \Lambda(\beta_0 + \beta_1 log_e(\boldsymbol{E})), \tag{18}$$

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

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].

Table [table ref here] shows the parameter estimates of the logistic regressions. [Discussion about the numeric estimates here]

Table 2. Summary of demographic information

Preferred pronoun	Education	Age group	Previous experience	Subject
She	Diploma and Bachelor Degree	18-24	Yes	26
She	Diploma and Bachelor Degree	18-24	No	24
He	Diploma and Bachelor Degree	18-24	No	22
He	High School or below	18-24	No	22
She	Diploma and Bachelor Degree	25-39	No	21
He	Diploma and Bachelor Degree	18-24	Yes	19
He	Diploma and Bachelor Degree	25-39	No	19
She	High School or below	18-24	No	19
He	High School or below	18-24	Yes	15
She	Diploma and Bachelor Degree	25-39	Yes	15
He	Diploma and Bachelor Degree	25-39	Yes	12
She	High School or below	18-24	Yes	10
He	Masters Degree	25-39	No	8
He	Masters Degree	25-39	Yes	8
He	High School or below	25-39	No	7
She	Masters Degree	25-39	Yes	7
She	High School or below	25-39	No	6
He	High School or below	25-39	Yes	5
She	Honours Degree	25-39	No	5
She	Honours Degree	25-39	Yes	4
She	Masters Degree	25-39	No	4
He	Honours Degree	25-39	Yes	3
She	Honours Degree	18-24	No	3
He	Diploma and Bachelor Degree	40-54	Yes	2
He	Diploma and Bachelor Degree	55-64	No	2
He	Honours Degree	25-39	No	2
He	Masters Degree	40-54	No	2
She	Diploma and Bachelor Degree	40-54	No	2
She	High School or below	25-39	Yes	2
She	Masters Degree	18-24	No	2
They	Diploma and Bachelor Degree	18-24	No	2
They	Diploma and Bachelor Degree	25-39	No	2
He	Diploma and Bachelor Degree	40-54	No	1
He	High School or below	40-54	Yes	1
He	High School or below	55-64	No	1
He	Honours Degree	18-24	No	1
He	Honours Degree	18-24	Yes	1
He	Honours Degree	40-54	Yes	1
He	Masters Degree	40-54	Yes	1
He	Masters Degree	55-64	No	1
Other	Diploma and Bachelor Degree	18-24	No	1
Other	High School or below	25-39	No	1
She	Diploma and Bachelor Degree	40-54	Yes	1
She	High School or below	40-54	No	1
She	High School or below	55-64	Yes	1
She	Honours Degree	40-54	Yes	1
She	Masters Degree	40-54	No	1
They	High School or below	18-24	No	1
They	High School or below	18-24	Yes	1
They	High School or below	25-39	No	1

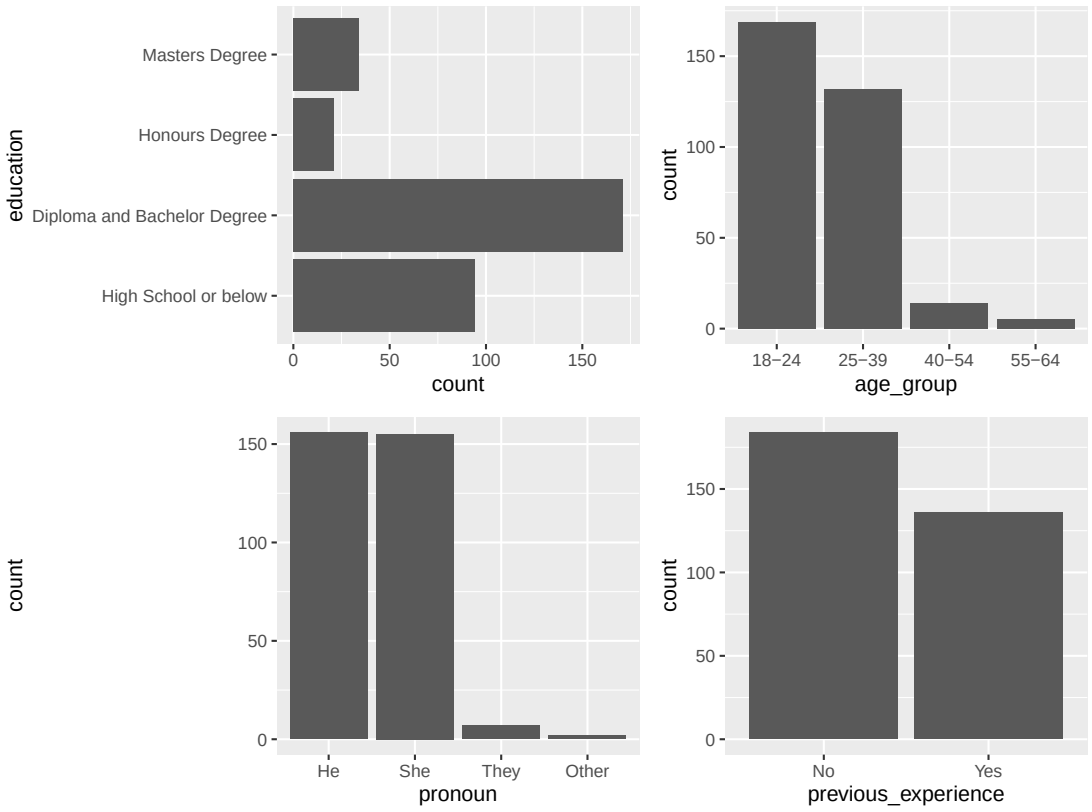


Figure 4. Summary of demographic information.

4.7. *Power comparison*

4.7.1. Experiment I

Figure 5 shows an overview of estimated power of visual test against natural logarithm of the effect with comparison to the power of an exact test - F-test, and the power of two other residual-based conventional tests commonly used in regression diagnostics but for testing other departures from the model assumptions. In overall, the power of all four tests increases as the effect becomes larger. The power curve of F-test climbs aggressively from 25% to around 90% as $log_e(E)$ increases from 0 to 2, while others respond inactively to the change of effect and remain lower than 25% throughout the period, showing that as an exact test, the F-test is relatively more sensitive to the type of model defects that being considered. The power of visual test arises steadily and nearly linearly to around 90% as $log_e(E)$ increases from 2 to 5, suggesting that the effect starts to make noticeable impact on the degree of the presence of the designed visual features. Other two inappropriate conventional tests shows improvement at the same time but at a lower rate. This coincides the point made by Cook and Weisberg (1982) mentioned in **??** that residual-based tests for a specific type of model defect are sensitive to other types of model defects. At $log_e(E) = 6$, the power curve of F-test reaches almost 100% followed by the visual test by a small margin. The power of Breusch–Pagan test and Shapiro–Wilk test reach around 75% and 63% respectively.

What truly impress us is the huge difference between the estimated power of visual test and the estimated power of F-test. The margin is largest at around $log_e(E) = 2$. An example lineup is included in Figure 5 where none of subjects detect the actual data plot positioned at panel 14. It demonstrates that at this level of difficulty, the designed visual feature is rarely visible, making the actual data plot indistinguishable from residual plots simulated from the assumed model. From a communication perspective, given the fact that the visual difference is unperceivable, the argument that non-linearity present in the fitted model is less convincing to the public even though it is true. At around $log_e(E) = 3$, the margin gets smaller as the chance of identifying the actual data plot becomes larger. At this level of difficulty, the designed visual features are usually detectable but it may not stand out from the lineup as other null plots may happen to include outliers or visual patterns that are are considered to be more attractive by human, and thus recognized as the most different plot. Without knowing the designed visual features beforehand, it is actually hard to identify the actual data plot by pure image comparison. The corresponding example lineup for $log_e(E) = 3$ shown in Figure 5 has the actual data plot positioned at panel 20, where two out of five subjects detect it. It can be observed that a M-shape is presented in plot 20, but the signal is not strong enough to attract all five subjects, resulting in a visual p-value sightly above the desired significance level $\alpha = 0.05$. At $log_e(E) = 4$ and $log_e(E) = 6$, the designed visual features become much clear and attractive, leading to a high percentage of rejection of the null hypothesis. Figure 5 gives example lineups of such cases.

4.7.2. Distributions of regressor

The impact of the distribution of X_raw on the power is shown in Figure 6. The power curve of F-test is stable across different distributions, while the visual test has a steeper power curve for normal and uniform distribution. BP-test performs worse for discrete uniform distribution and uniform distribution but has relatively high power for normal distribution. SW-test outperforms BP-test for discrete uniform distribu-

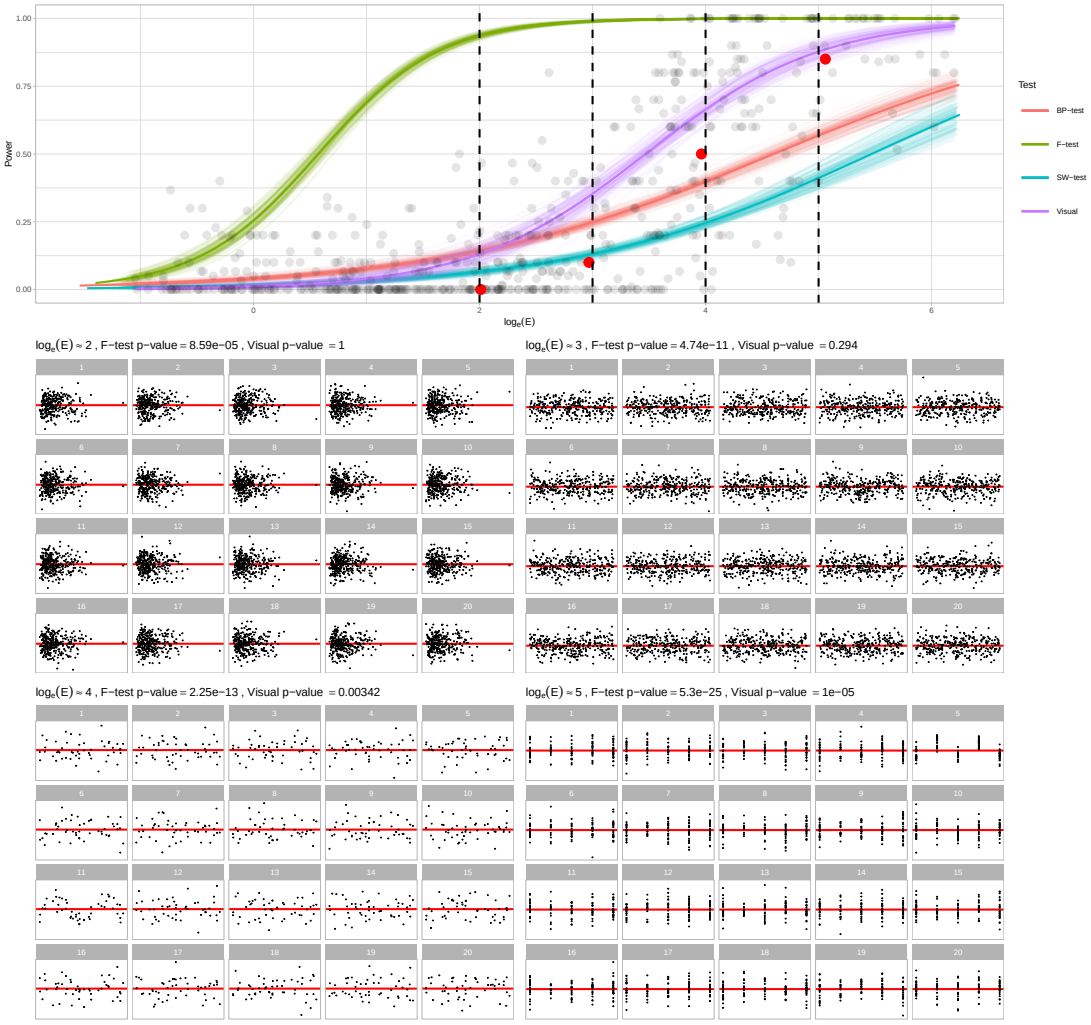


Figure 5. Power overview.

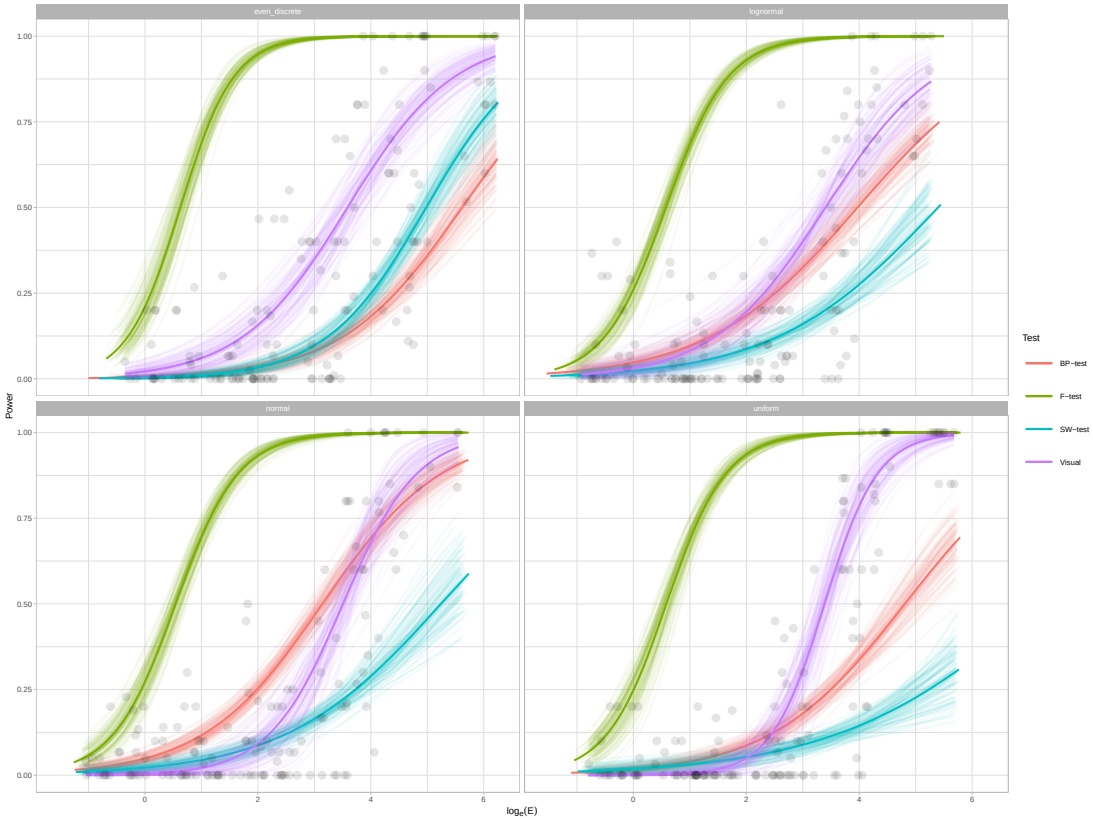


Figure 6.

tion but remains as the worst test for other distributions. The results indicate those inappropriate residual-based tests are sensitive to the distribution of the regressor.

4.7.2.1. Shpaes of non-linearity. Figure 7 illustrates the change of power under different shape of non-linearity. Similar to the power curves shown in 6, F-test is stable under different shapes. The power curve of visual test also behaves similarly across different shapes. What vary are the power curve of BP-test and SW-test. For Triple-U shape, both BP-test and SW-test are insensitive to the change of the effect. And for W shape, both tests have almost identical power curves. It can be observed that both tests performs the best for U-shape.

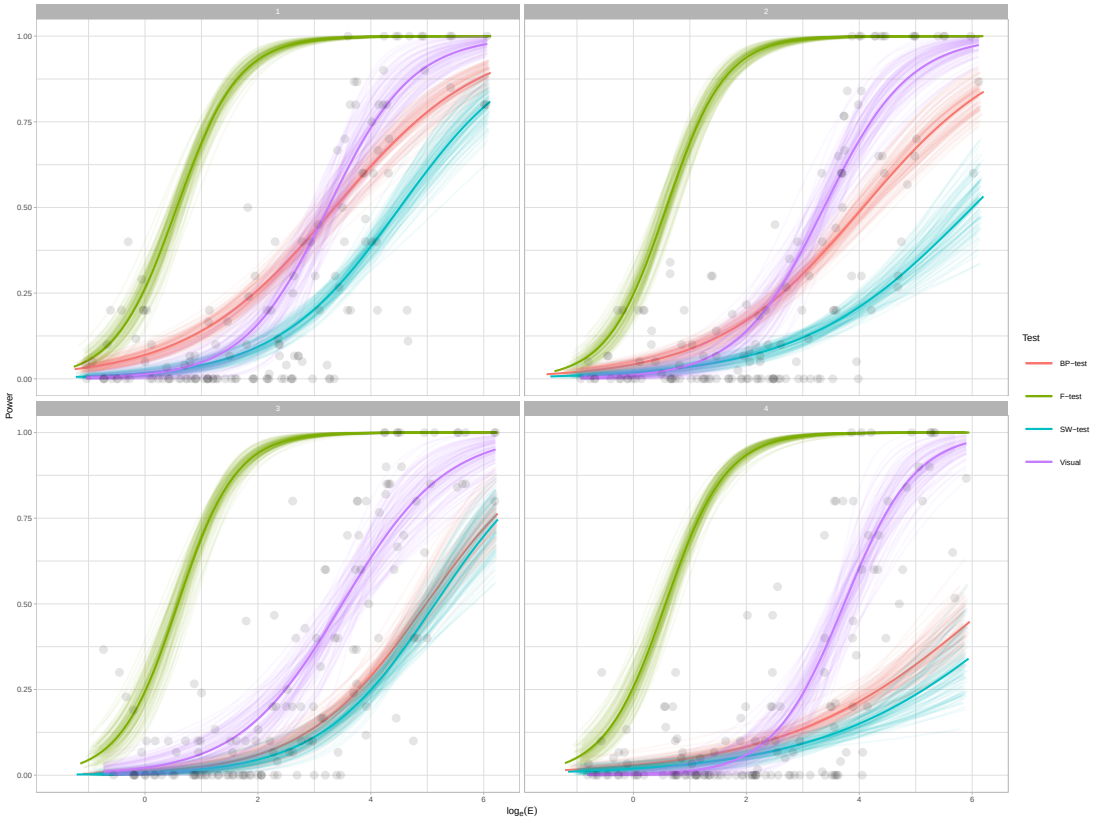
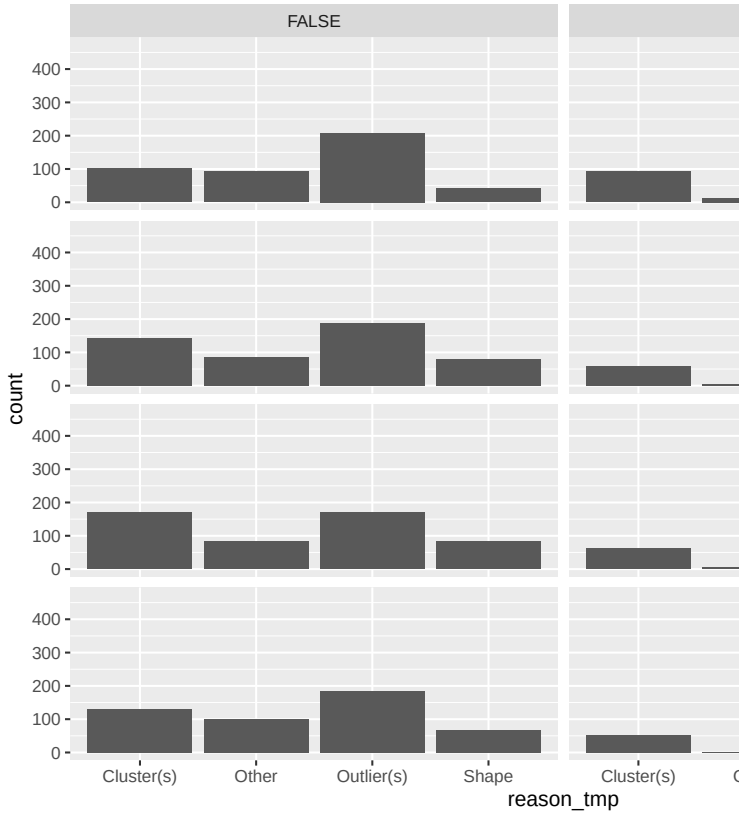
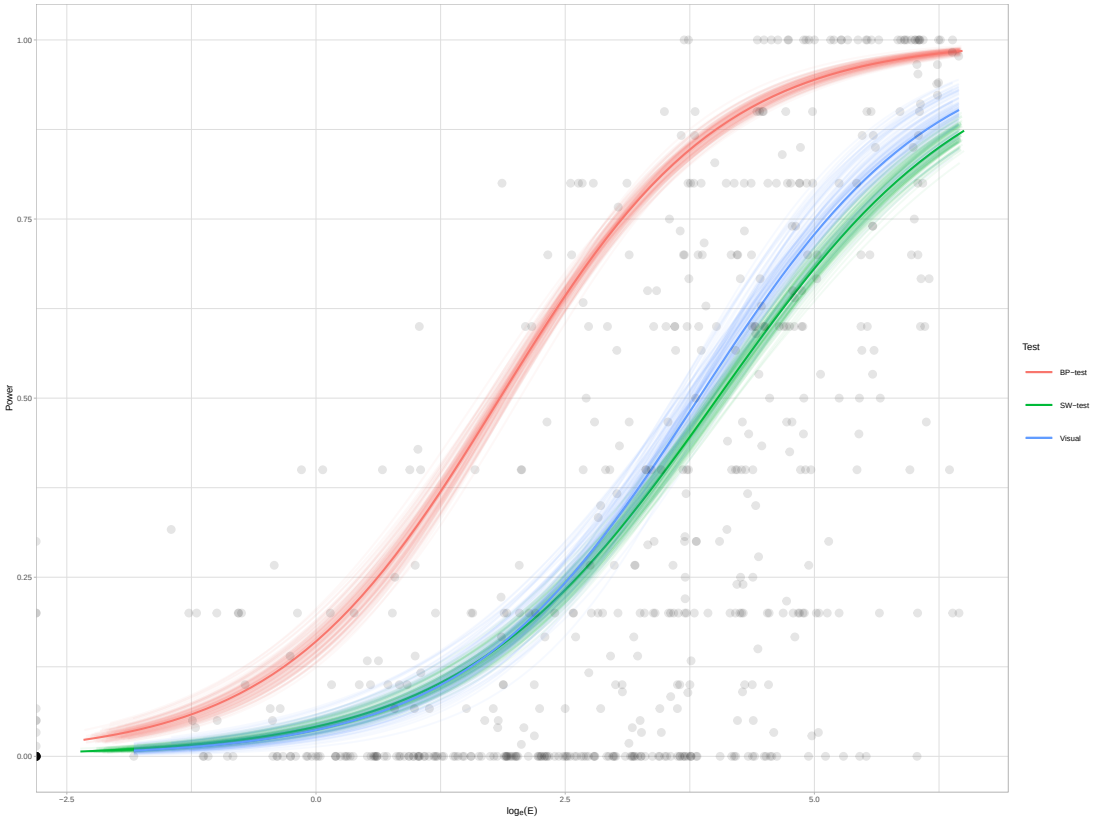

1
2
3
4
Power
1.00
0.75
0.50
0.25
0.00
0
2
4
6
log$_e$(E)
Test
BP−test
F−test
SW−test
Visual

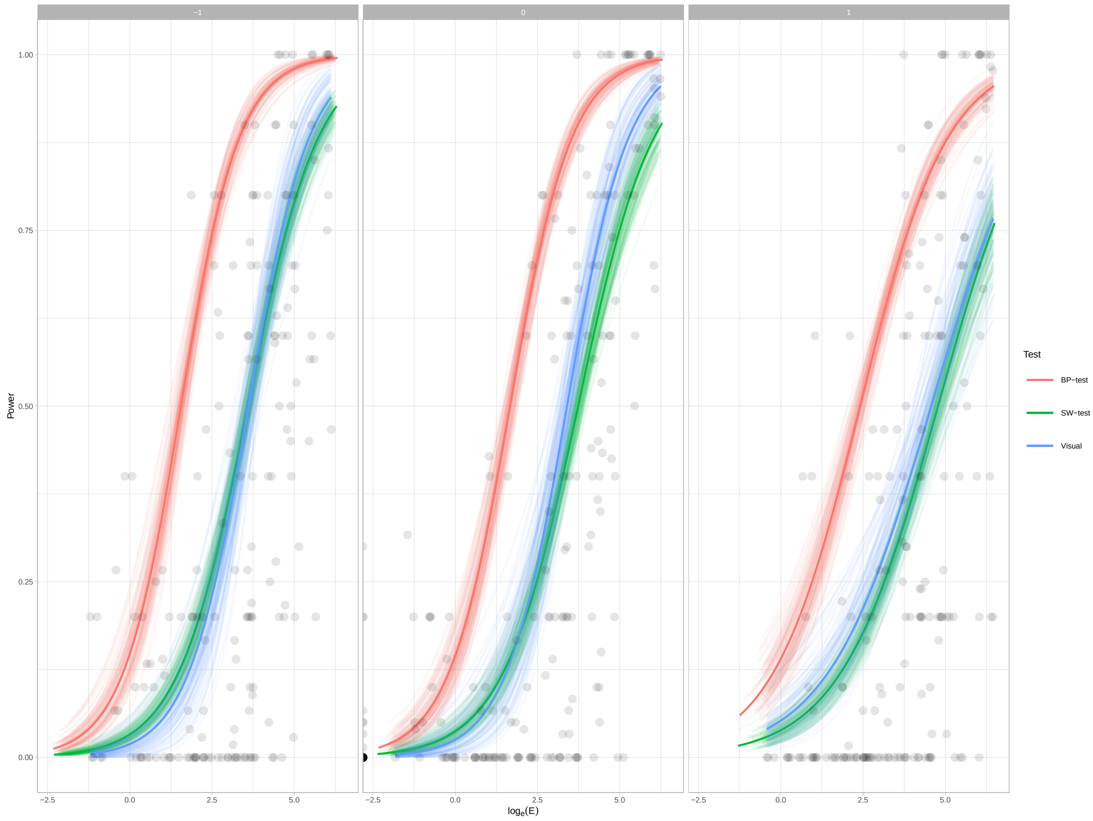
Figure 7.

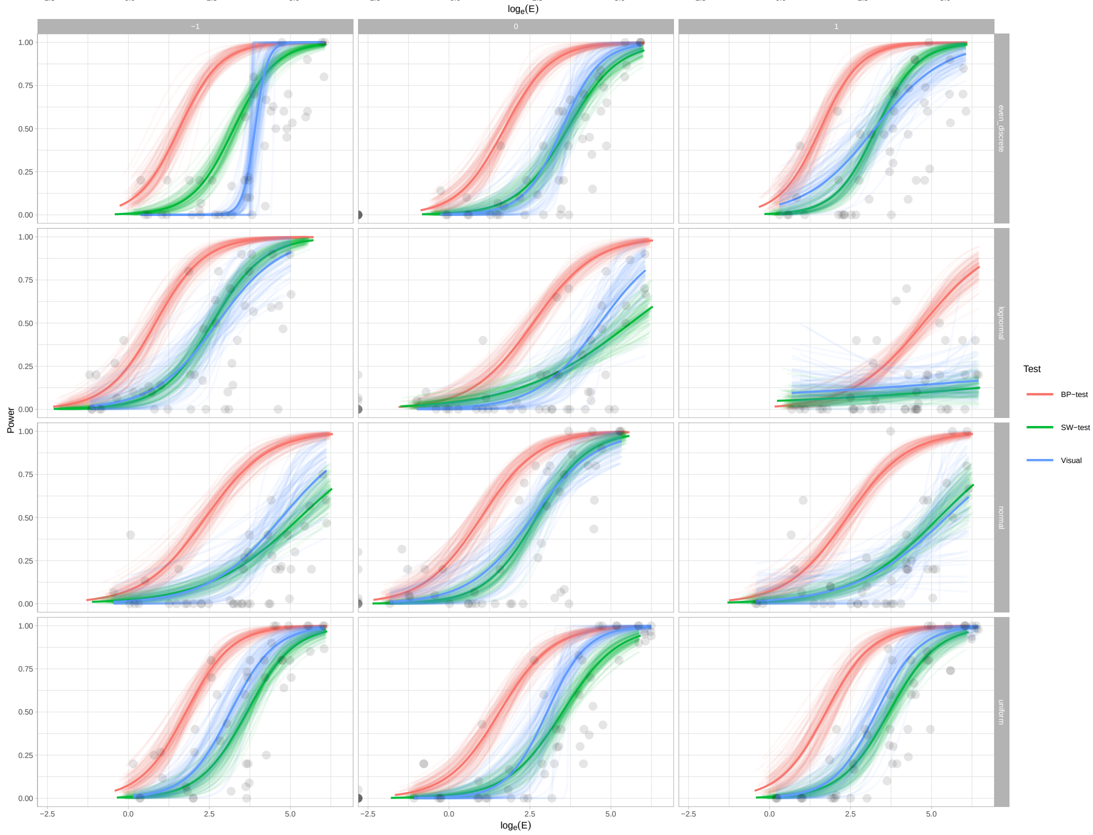


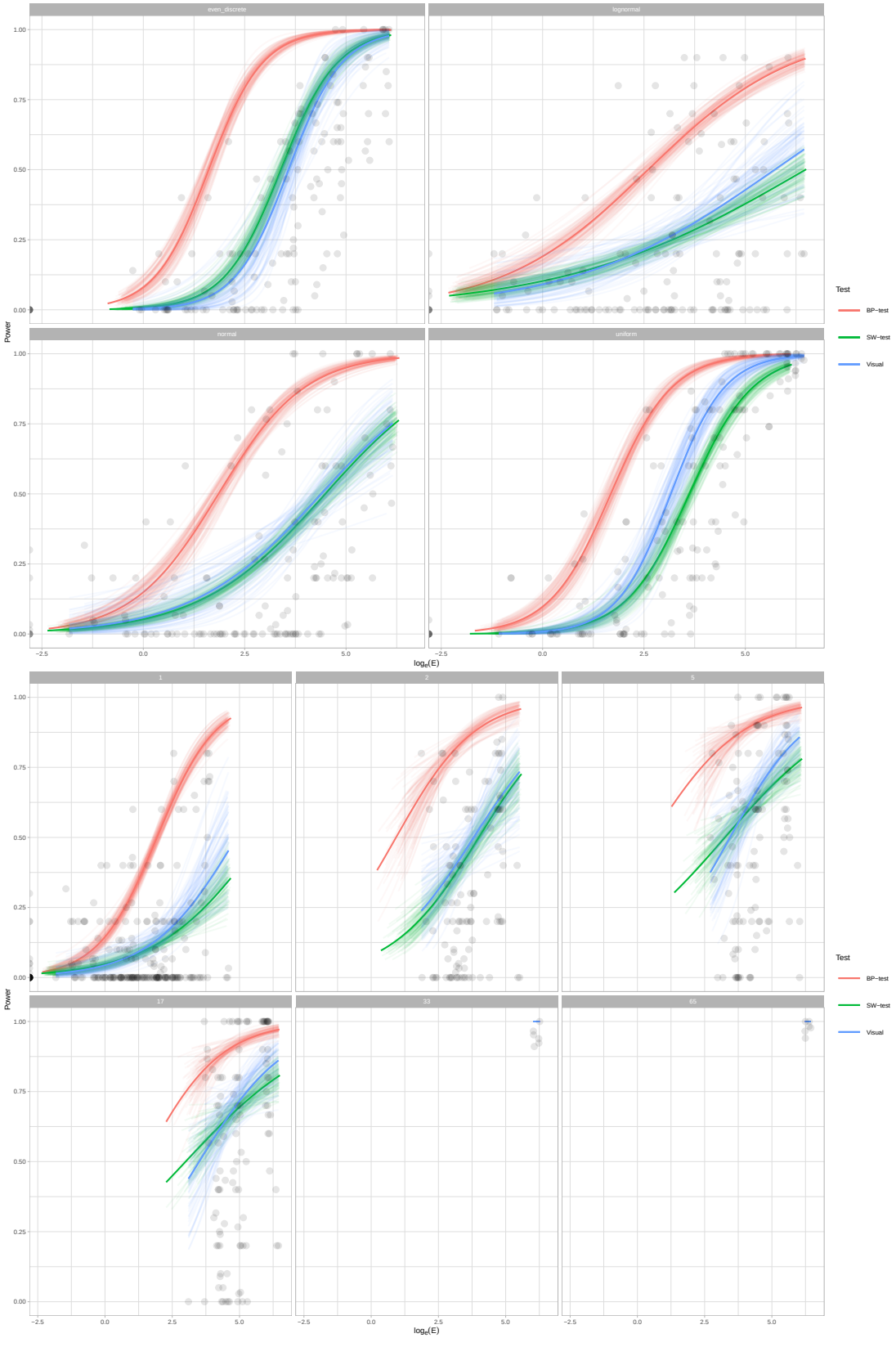
4.7.2.2. Reasons for making selections.

4.7.3. Experiment II




−1
0
1
Power
1.00
0.75
0.50
0.25
0.00
−2.5
0.0
2.5
5.0
log$_e$(E)
Test
BP−test
SW−test
Visual


−1
0
1
even_discrete
lognormal
normal
uniform
Power
1.00
0.75
0.50
0.25
0.00
−2.5
0.0
2.5
5.0
log$_e$(E)
Test
BP−test
SW−test
Visual


even_discrete
lognormal
normal
uniform
1
2
5
17
33
65
Power
$\log_{10}(E)$
Test
BP−test
SW−test
Visual

$->$

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*, .

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.

Gelman, Andrew. 2003. "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-fit Testing." *International Statistical Review* 71 (2): 369–382.

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

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

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.

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

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