ARTICLE TEMPLATE

Effectiveness of visual test in linear regression diagositics

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

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

ABSTRACT
Abstract to fill.

1. Introduction

Regression diagnostics is an essential step in regression analysis which is a field of study with at least a hundred years of history. The diagnostic procedure conventionally involves evaluating the fitness of the proposed model, detecting the presence of influential observations and outliers, checking the validity of model assumptions and many more. In terms of diagnostic techniques, data plots, hypothesis testing, and summary statistics are vital tools for a systematic and detailed examination of the regression model (Mansfield and Conerly 1987).

Many of those regression diagnostic methods and procedures are mature and well-established in books first published in the twentieth century, such as Draper and Smith (1998), Montgomery and Peck (1982), Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1999) and Cook and Weisberg (1982). Regardless of the level of difficulty, one will find the importance and usefulness of diagnostic plots being emphasized in those books repeatedly. Checking diagnostic plots is also the recommended starting point for validating model assumptions such as normality, homoscedasticity and linearity (Anscombe and Tukey 1963).

1.1. *Diagnostic 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 suggested by Cook and Weisberg (1982), these kinds of diagnostic plots are commonly used to identify patterns that are indicative of nonconstant error variance or non-linearity. Raw residuals and studentized residuals are the two most frequently used residuals

---

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

in standard residual plots. The debt on which type of residuals should be used is always present. While raw residuals are the most common computer regression software package output, by applying a scaling factor, the ability to reveal nonconstant error variance in standard residual plots will often be enhanced by studentized residuals in small sample size (Gunst and Mason 1980). As a two-dimensional representation of a model in a p-dimensional space, standard residual plots project data points onto the variable of the horizontal axis, which is a vector in p-dimensional space. Observations with the same projection will be treated as equivalent as they have the same abscissa. Consequently, standard residual plots are often useful in revealing model inadequacies in the direction of the variable of the horizontal axis but could be inadequate for detecting patterns in other directions, especially in those perpendicular to the variable of the horizontal axis. Hence, in practice, multiple standard residual plots with different horizontal axes will be examined (Cook and Weisberg 1982). Overlapping data points is another general issue in scatter plots not limited to standard residual plots, which often makes plots difficult to interpret because visual patterns are concealed. Thus, for a relatively large sample size, Cleveland and Kleiner (1975) suggests the use of robust moving statistics as reference lines to give aid to the eye in seeing patterns, which nowadays, are usually replaced with a spline or local polynomial regression line.

Other types of data plots that are often used in regression diagnostics include partial residual plots and probability plots. Partial residual plots are useful supplements to standard residual plots as they provide additional information on the extent of the non-linearity. Probability plots can be used to compare the sampling distribution of the residuals to the normal distribution for assessing the normality assumptions.

1.2. *Hypothesis testing*

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 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).

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 the null hypothesis 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 the null hypothesis despite the residuals are well-behaved (Cook and Weisberg 1999). This can be largely avoided in diagnostic plots as experienced analysts can evaluate the acceptability of assumptions flexibly, even in the presence of outliers. Montgomery and Peck (1982) suggested that 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. Not to mention, it is almost impossible to have an exactly correctly specified model in reality. There is a well-known aphorism in statistics stated by George Box - "All models are wrong, but some are useful". This indicates proper

hypothesis tests will always reject the null hypothesis as long as the sample size is large enough. The outcome "Not reject" can be interpreted as either "effect size is small" or "sample size is small". and the outcome "reject" doesn't inform us whether and how much the model defects are of actual consequence to the inference and prediction. But still, the effectiveness of statistical tests shall not be disrespected. Statistical tests have a chance to provide analysts with unique information. In situations where no suitable diagnostic plots can be found for a particular violation of the assumptions, or excessive diagnostic plots are needed to be checked, one will have no choice but to fall back on statistical tests if there exists any. A good regression diagnostic practice should be a balanced combination of both methods.

1.3. *Visual inference*

However, 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 and McGill 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 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, 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 1 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
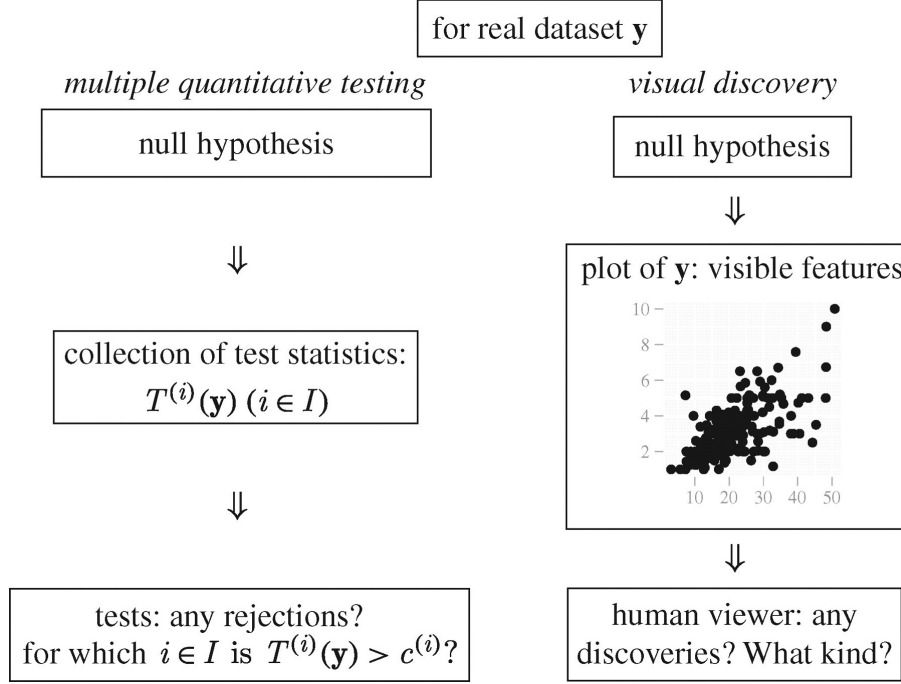
Figure 1. Parallelism between multiple quantitative testing and visual discovery (Buja et al. 2009). Visible features in a plot are viewed as a collection of test statistics $T^{(i)}(\mathbf{y})$ ($i \in I$), and any visual discoveries that are inconsistent with the null hypothesis are treated as evidence against the null. For regression diagnostics, the null hypothesis would be the assumed model, and visual discoveries could be any visual features in favour of any alternatives.

plot, including non-linearity and heteroskedasticity as shown in Figure 2.

1.3.1. 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 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 $\boldsymbol{S}(\boldsymbol{y}) = (\hat{\boldsymbol{\beta}}, \boldsymbol{e}'\boldsymbol{e})$, where $\hat{\boldsymbol{\beta}}$ are the coefficient estimators and $\boldsymbol{e}'\boldsymbol{e}$ 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).
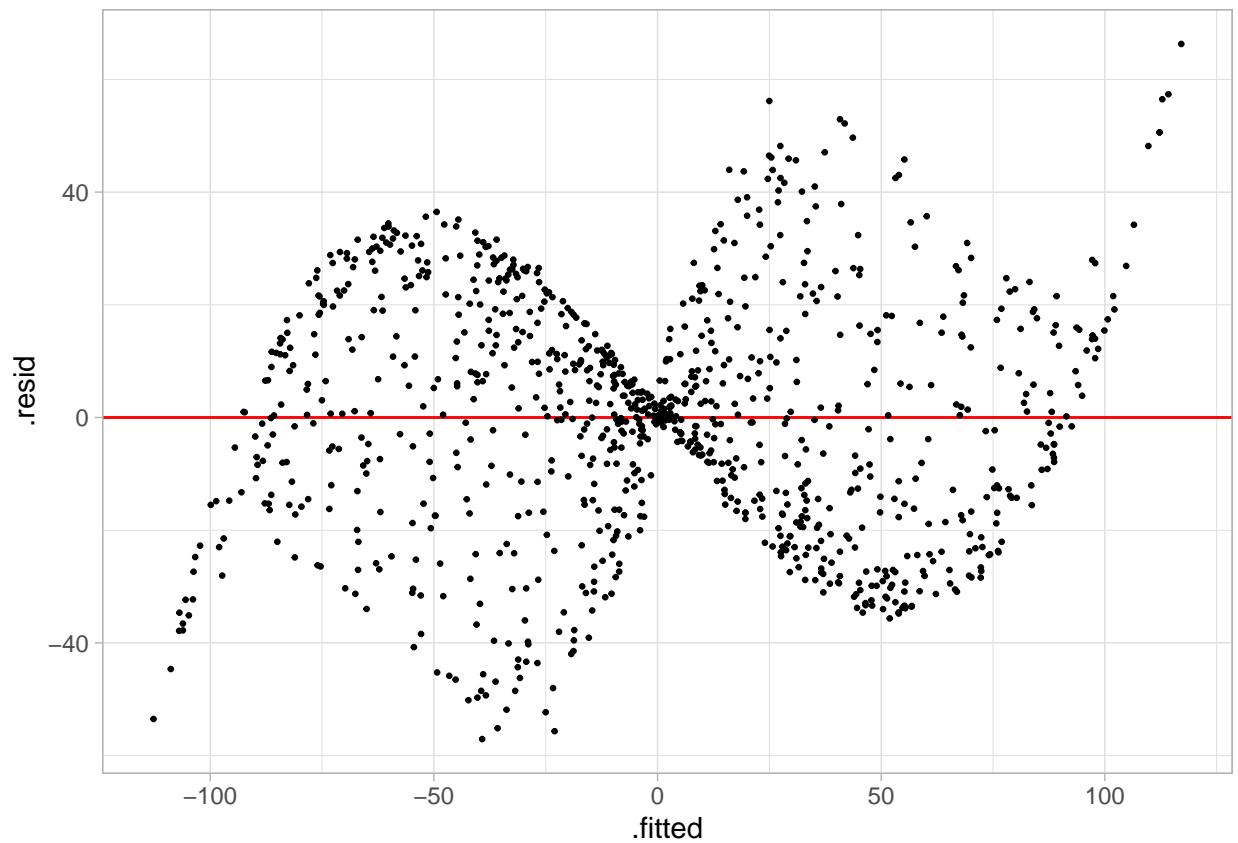
Figure 2. Residuals vs. fitted values plot for a classical linear regression model. The residuals are produced by fitting a two-predictor multiple linear regression model with data generated from a cubic linear model. From the residual plot, "butterfly shape" can be observed which generally would be 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.

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.

1.3.2. 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 2, 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

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

1.4. *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 suggests 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.

2. Experimental design

For the purpose of examining the effectiveness of visual test in regression diagnostics, two experiments were conducted. The first experiment has ideal scenario for conventional testing, where we would not expect the visual test to do better than the conventional test. The second experiment is a scenario where the conventional test is an approximate test, which visual test may have a chance to match the performance of the conventional test.

As suggested by many previous studies [many studies refs here], it is reasonable for an subject to evaluate a block of 20 lineups. Subjects for both two experiments were recruited from Prolific web service [ref here]. Data screening procedure was applied during the recruitment, subjects were required to be fluent in English, with 98% minimum approval rate and 10 minimum submissions. During the experiment, they were asked to select one or more plots that are mostly different from others, provide a reason

for their selections, and express how different they think the selected plots are from others. If there was no noticeable difference between plots in a lineup, subject could choose to select zero plots without providing the reason. Information about preferred pronoun, age group, education, and previous experience in visual experiment were also collected. No subject was shown the same lineup twice.

For each of the experiments, a pool of 12 lineups with obvious visual patterns were generated. In every block of 20 lineups that presented to subjects, two different lineups sampled from the corresponding pool of lineups were included as attention checks. Every subject should get at least one attention check correct such that the quality of the subject effort could be ensured.

2.1. Non-linear covariates

The first experiment is designed to study the ability of human subjects to detect the effect of a continuous variable $\boldsymbol{Z}$ which is a probabilist's Hermite polynomial [Herimite ref here] of another variable $\boldsymbol{X}$ in a two variable statistical model formulated as:

$$\boldsymbol{Y} = 1 + \boldsymbol{X} + \boldsymbol{Z} + \boldsymbol{\varepsilon}, \tag{7}$$

$$\boldsymbol{X} = g(\boldsymbol{X}_{raw}, 1), \tag{8}$$

$$\boldsymbol{Z} = g(\boldsymbol{Z}_{raw}, 1), \tag{9}$$

$$\boldsymbol{Z}_{raw} = He_j(g(\boldsymbol{X}, 2)), \tag{10}$$

where $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\varepsilon}, \boldsymbol{X}_{raw}, \boldsymbol{Z}_{raw}$ are $n \times 1$ matrices, $He_j(.)$ 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 variable 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 data generated by the above model is formulated as:

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X} + \boldsymbol{u}, \tag{12}$$

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

Clearly, omitted-variable bias will present since the null model leaves out the higher order term. 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 ($\sigma = 0.5, 1, 2, 4$) and four different distribution of $\boldsymbol{X}_{raw}$: (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 are included in the residual plot. These include "U" shape, "S" shape, "M" shape and "Triple-U" shape.
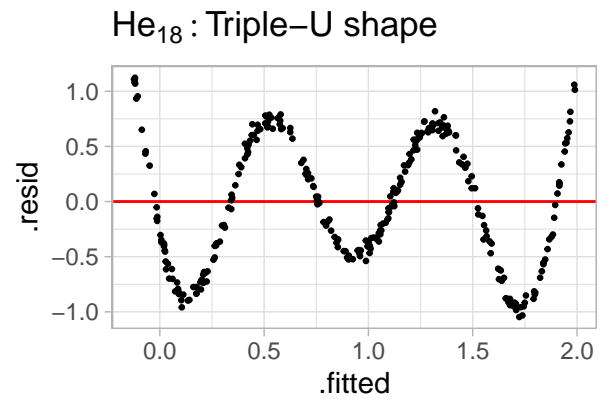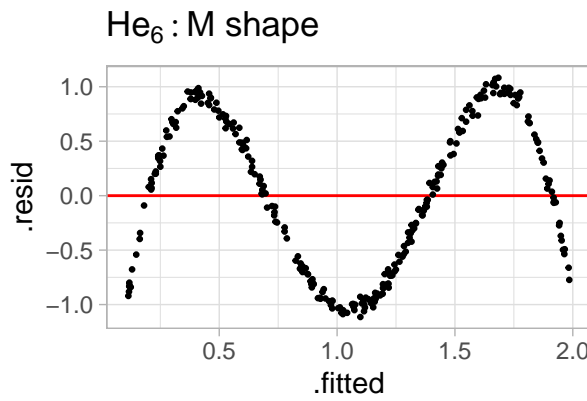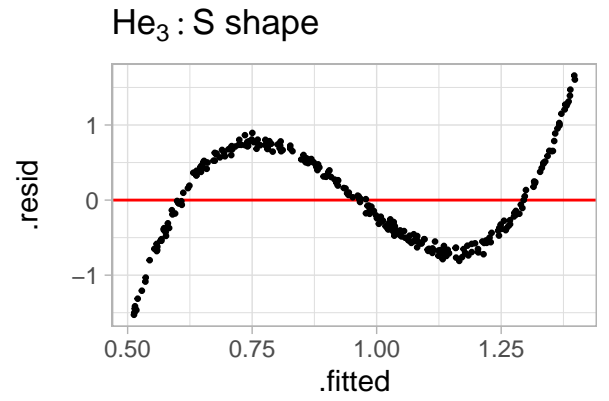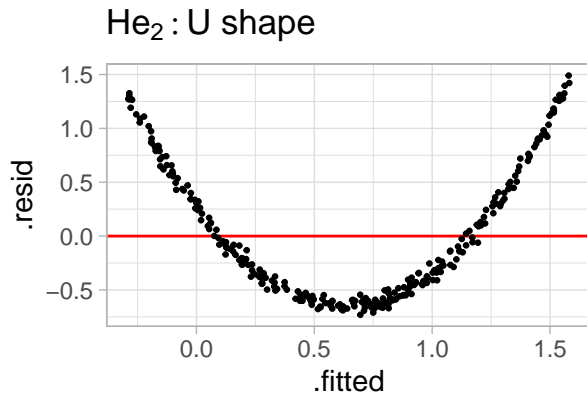
The range of σ, which is a factor controlling the strength of the signal, was chosen so that different difficulty levels of lineups were generated and the estimated power curve would be smooth and continuous.
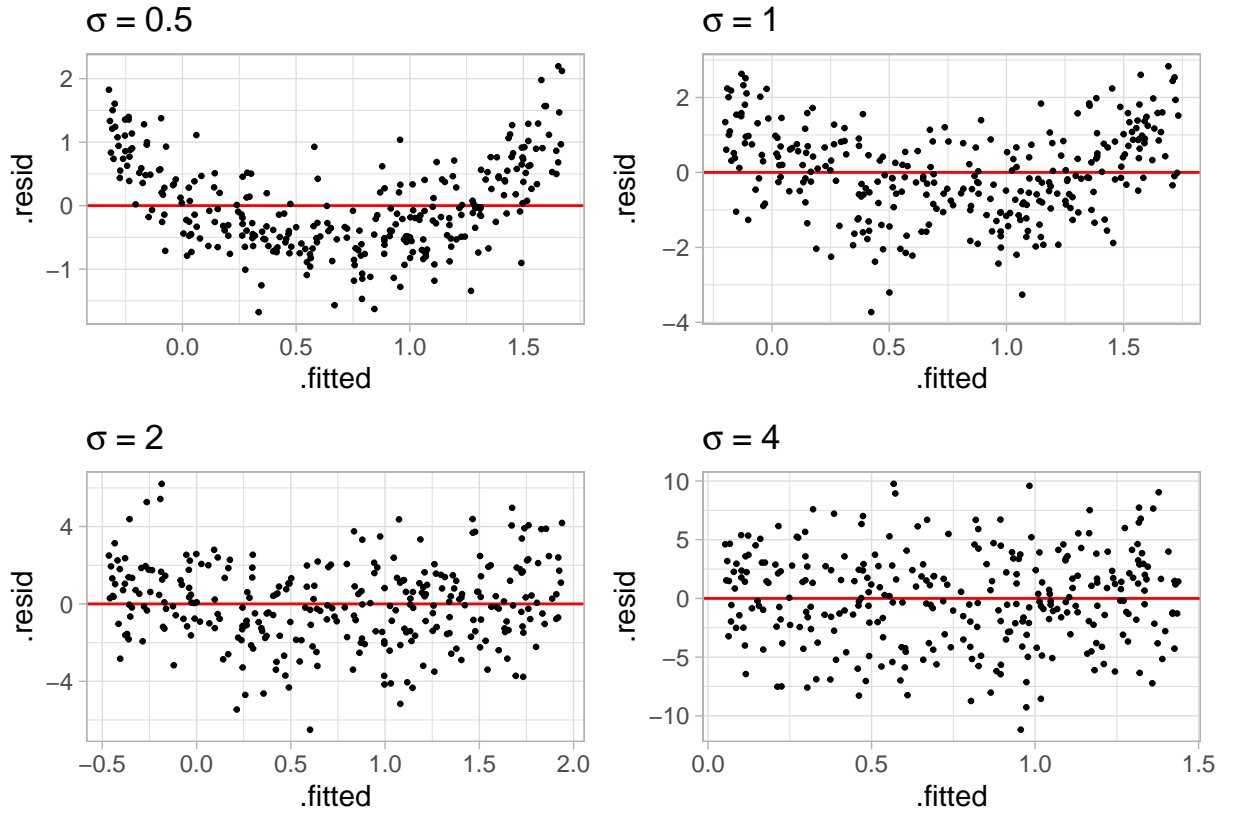
Four different distribution were used to generate $\boldsymbol{X}_{raw}$. The uniform and the normal distribution are symmetric and commonly assumed in statistical models. The adjusted log-normal distribution provides skewed density. And the discrete uniform distribution provides discreteness in residual plot, which could enrich the pool of visual patterns.

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	$U\{1, 5\}$
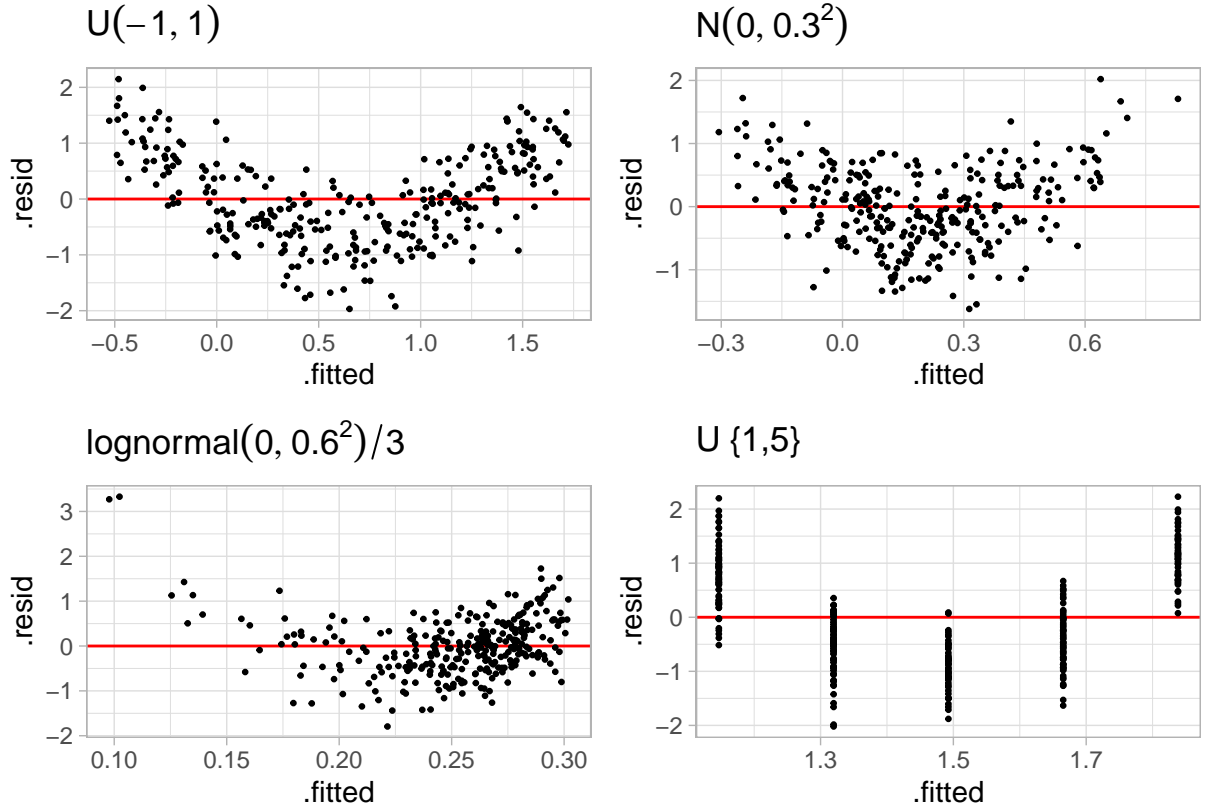
Three replications are made for each of the parameter values shown in Table 1 resulting in 192 different lineups. For each lineup, the actual data plot was drawn as standard residual plot of the null model with raw residuals on the y-axis and fitted values on the x-axis. The 19 null datasets were generated 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 is four. All five subjects correctly identified the actual data plot for this lineup.

In addition, each lineup is 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 experimental design described above.
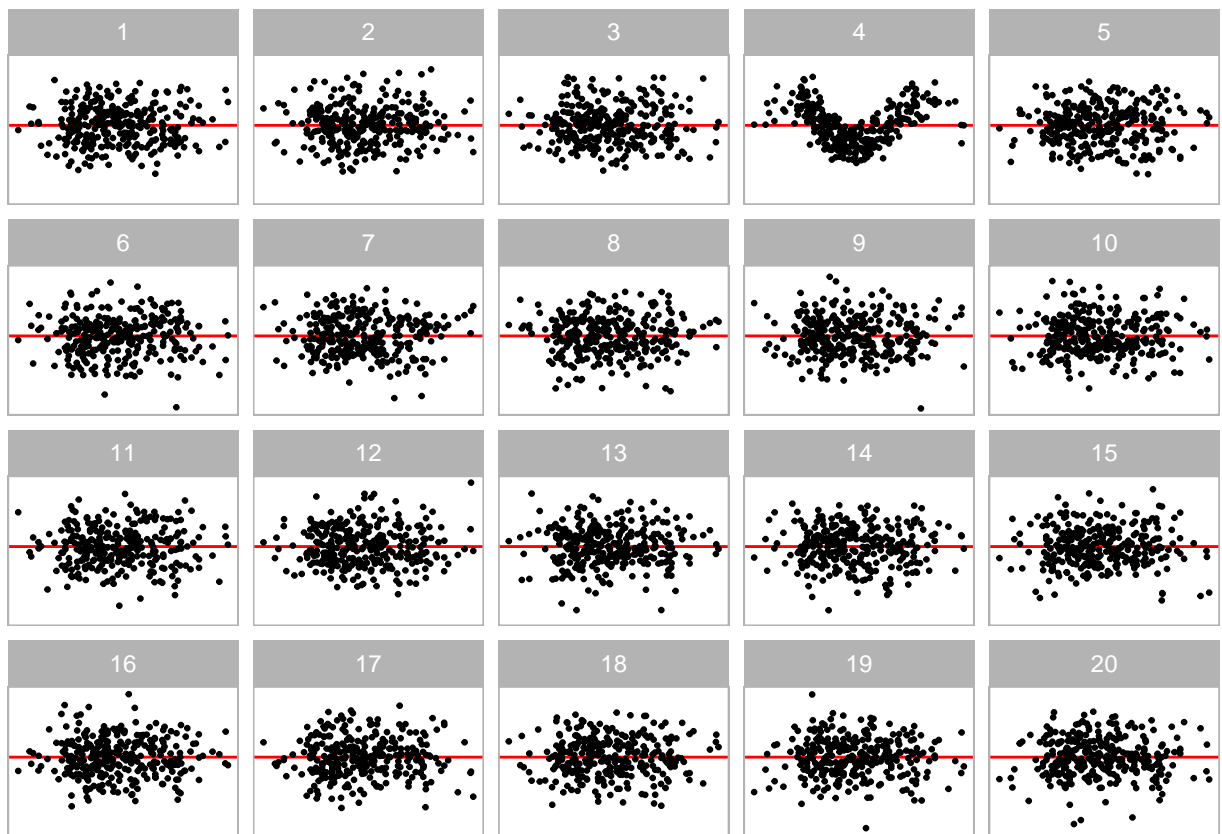
Figure 3. Example lineup

3. Results

3.1. *Data preprocessing*

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, which literally means there was no visble difference between plots, but the simulated 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 experiment I, where 320 lineup evaluations were attention checks and were not used in the following analysis. The collated dataset is provided in `polynomials` of the `visage` R package.

3.2. *Demographic summary*

Table 2 tabulates the number of subjects, preferred pronouns, education backgrounds, age groups, and previous experience in visual experiment. Figure 4 visulizes 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.

3.3. *Model fitting*

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

$$E = \frac{1}{2\sigma^2} \boldsymbol{X}'_b \boldsymbol{R}'_a (diag(\boldsymbol{R}_a))^{-1} \boldsymbol{R}_a \boldsymbol{X}_b, \tag{13}$$

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:

$$Pr(\text{reject } H_0|H_1, E) = \Lambda(\beta_0 + \beta_1 log_e(\boldsymbol{E})), \tag{14}$$

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

Table 2. Summary of demographic information

| Preferred pronoun	Education	Age group	Previous experience	Subject
She	Diploma and Bachelor Degree	18-24	No	17
He	Diploma and Bachelor Degree	18-24	Yes	14
She	Diploma and Bachelor Degree	18-24	Yes	12
He	Diploma and Bachelor Degree	18-24	No	11
He	Diploma and Bachelor Degree	25-39	No	10
She	Diploma and Bachelor Degree	25-39	No	10
She	Diploma and Bachelor Degree	25-39	Yes	9
He	High School or below	18-24	No	8
She	High School or below	18-24	No	7
He	High School or below	18-24	Yes	6
He	Diploma and Bachelor Degree	25-39	Yes	5
He	Masters Degree	25-39	No	5
He	Masters Degree	25-39	Yes	5
He	High School or below	25-39	No	4
She	High School or below	25-39	No	4
She	Masters Degree	25-39	Yes	4
He	High School or below	25-39	Yes	3
She	High School or below	18-24	Yes	3
She	Masters Degree	25-39	No	3
He	Honours Degree	25-39	No	2
She	Diploma and Bachelor Degree	40-54	No	2
She	Honours Degree	18-24	No	2
He	High School or below	40-54	Yes	1
He	High School or below	55-64	No	1
He	Masters Degree	40-54	Yes	1
He	Masters Degree	55-64	No	1
Other	High School or below	25-39	No	1
She	High School or below	40-54	No	1
She	Honours Degree	25-39	No	1
She	Honours Degree	25-39	Yes	1
She	Masters Degree	18-24	No	1
She	Masters Degree	40-54	No	1
They	Diploma and Bachelor Degree	18-24	No	1
They	Diploma and Bachelor Degree	25-39	No	1
They	High School or below	18-24	No	1
They	High School or below	25-39	No	1
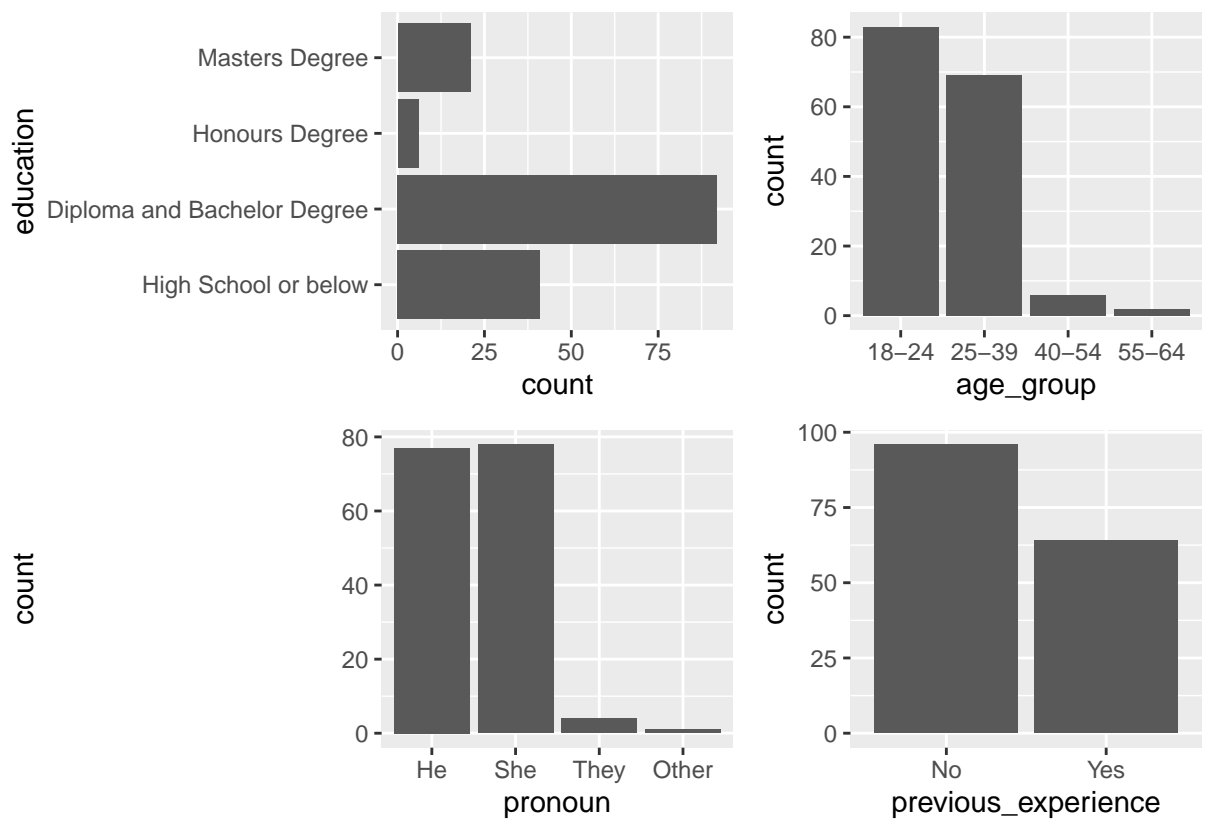
Figure 4. Summary of demographic information.

The same model is fit on different subsets of the collated data to study the effect of different factors have on the power of the visual test.

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

3.4. *Power comparison*

Figure 5 shows an overview of estimated power of visual test against natural logarithm of the effect. Without further concenring different component contributing to the effect,
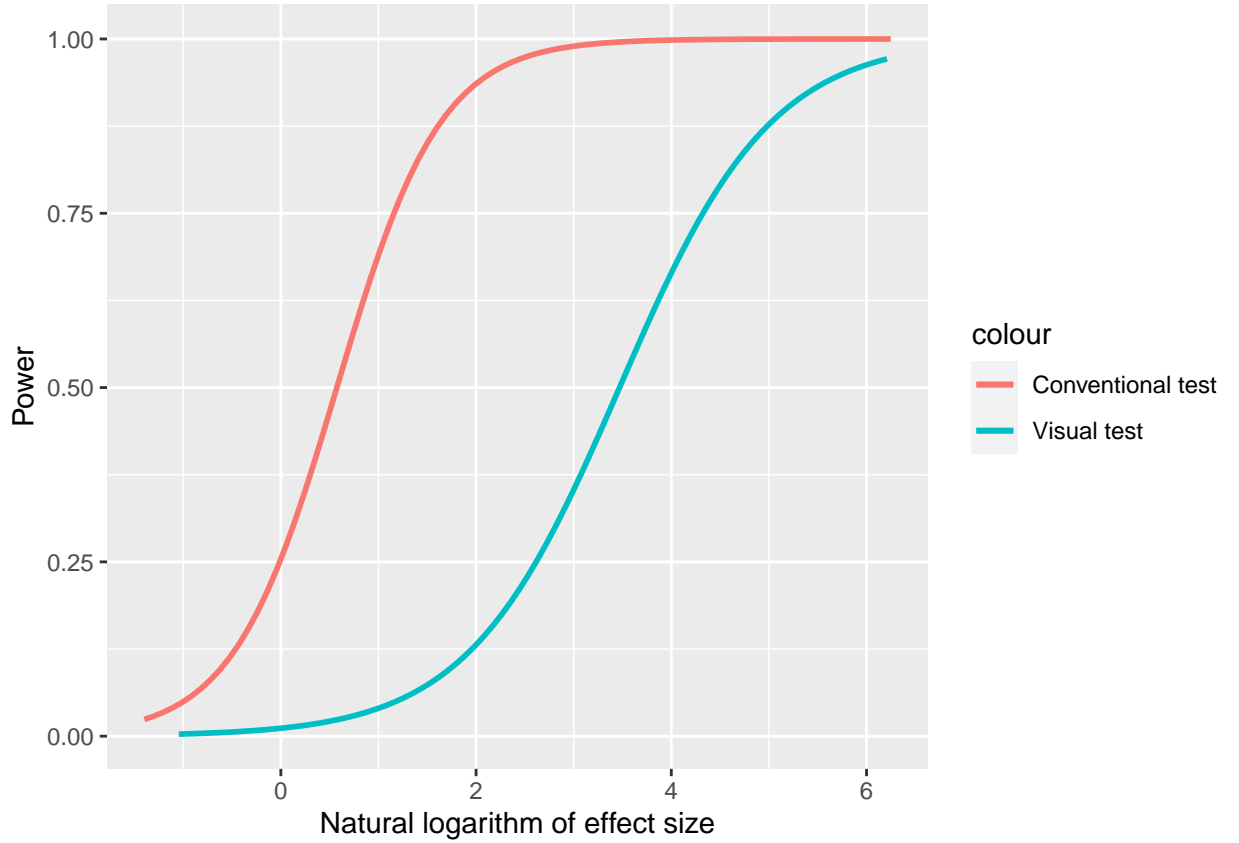
4. **Old**

4.1. *Overview of the Data*

We collected 400 lineup evaluations made by 20 participants in experiment I and 880 lineup evaluations made by 44 participants in experiment II. In total, 442 unique lineups were evaluated by 64 subjects. In experiment I, one of the participants skipped all 20 lineups. Hence, the submission was rejected and removed from the dataset. In experiment II, there was a participant failed one of the two attention checks, but there was no further evidence of low-effort throughout the experiment. Therefore, the submission was kept.
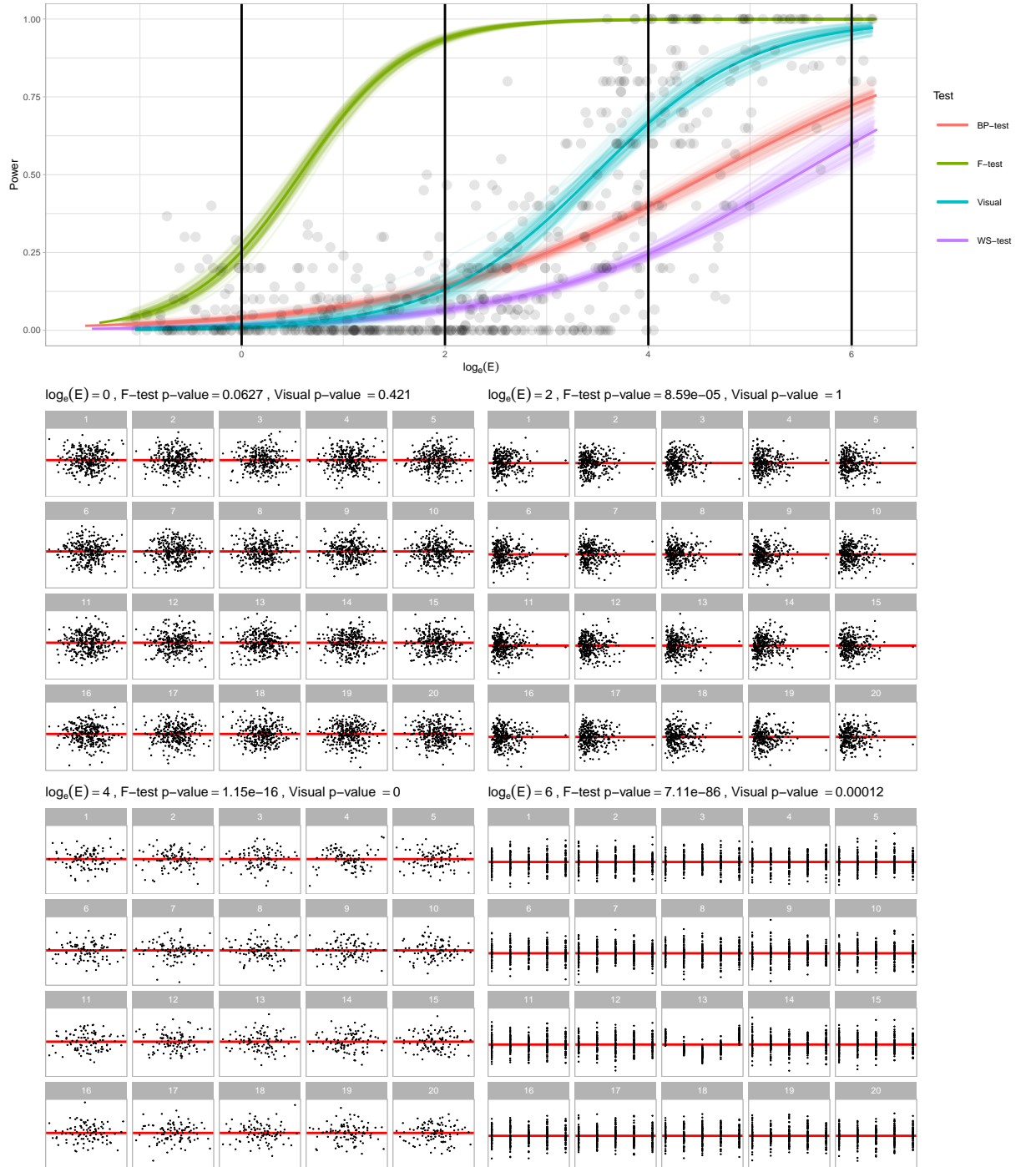
Figure 5. Power overview.

4.2. *Power comparision*

1. power (visual test vs. conventional test) (visual test most different one (everything test, any departure)) plot figure in a paper, desc, exp
2. investigate the difference (gap), give examples
3. conventional is too sensitive
4. make conventional less sensitive (vary alpha)

To model the power of visual test, 10 logistic regression were fit for different number of evaluations ranged from one to five and two different types of simulation setting. All 10 models used natural logarithm of the effect size as the only fixed effect, and whether the test successfully rejects the null hypothesis as the response variable. Given the way we define the effect size, it was expected that with larger effect size, both conventional test and visual test will have higher probability in rejecting the null hypothesis when it is not true. The modelling result summarized in **??** and **??** aligned with the expectation as the coefficients of natural logarithm of the effect size are positive and significant across all 10 models.

Figure **??** illustrates the fitted models, while providing the local constant estimate of the power of F-test and Breusch–Pagan test for comparison. Data for the conventional test is simulated under the model setting described in section . . . and 5000000 samples are drawn for both cubic and heteroskedasticity model. From Figure **??**, it can be observed that the fitted power of visual test increased as the number of evaluations increased for both cubic and heteroskedasticity model.

For heteroskedasticity model, this phenomenon was more obvious as the power of visual tests with evaluations greater than two were always greater than those with evaluations smaller than two.

For cubic model, the separation between curves was small. The estimated power of visual tests with three to five evaluations were almost identical to each other in regards of effect size. In addition, all five curves peaked at one as effect size increased, suggesting that identification of non-linearity as a visual task can be completed reliably by human as long as the departure from null hypothesis is large enough.

As shown in Figure **??**, both F-test and Breusch–Pagan test generally possessed greater power than visual test. A visual tests is a collection of test against any alternatives that would create visual discoverable features, while a conventional test is usually targeting at a pre-specified alternative. Considering the data generating process of the model defect was known and controlled in this research, where all other alternatives have been eliminated except the one we concerned, the result was suggested that conventional tests were more sensitive to violations of linearity and homoscedasticity assumption than visual tests.

It was also found that there was a noticeable gap between curves of the conventional test and the visual test at around log(effect size) $= 0$ for the cubic model and log(effect size) $= 2.5$ for the heteroskedasticity model, where the differences in power were greater than 0.6. We further analysed the lineups with correspoding effect sizes. Figure **??** and **??** showed that human was indeed hard to identify the patterns at this level of difficulty. The visual difference between the true data plot and null plots were almost unnoticeable.

4.3. *Effect of parameters on power of the visual test*

The previous section focuses on the change of effect size relative to the power of the visual test. However, effect size is only a one dimensional summarisation of parameters used in data simulation. Individual factor embedded in the simulation process should also be analysed.

In cubic model, two major factors that influencing the strength of the signal are a and b. Figure **??** and **??** illustrates 30 different logistic regressions fit for different number of evaluations and different number of observations n. The regressor used in these models was $|a|/\sigma$ since the noise level σ needed to be taken into account. From the figures, we can observe . . .
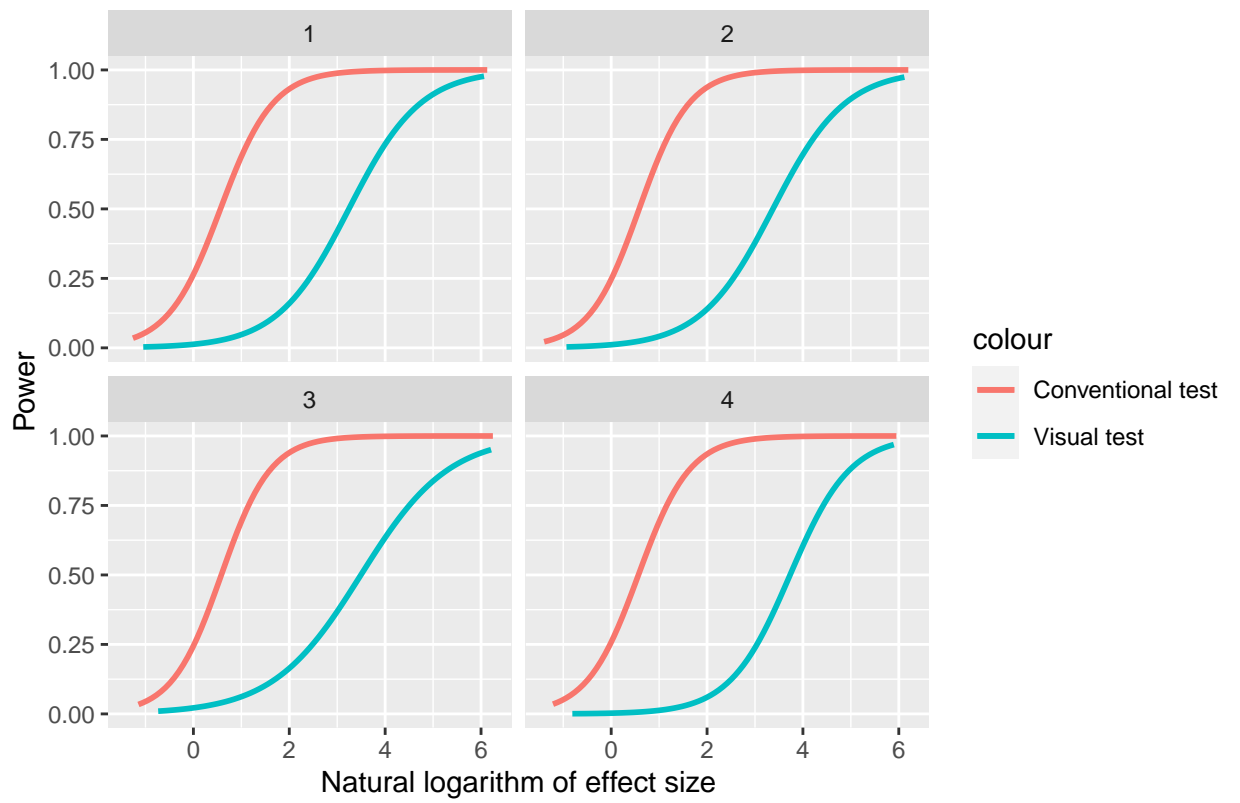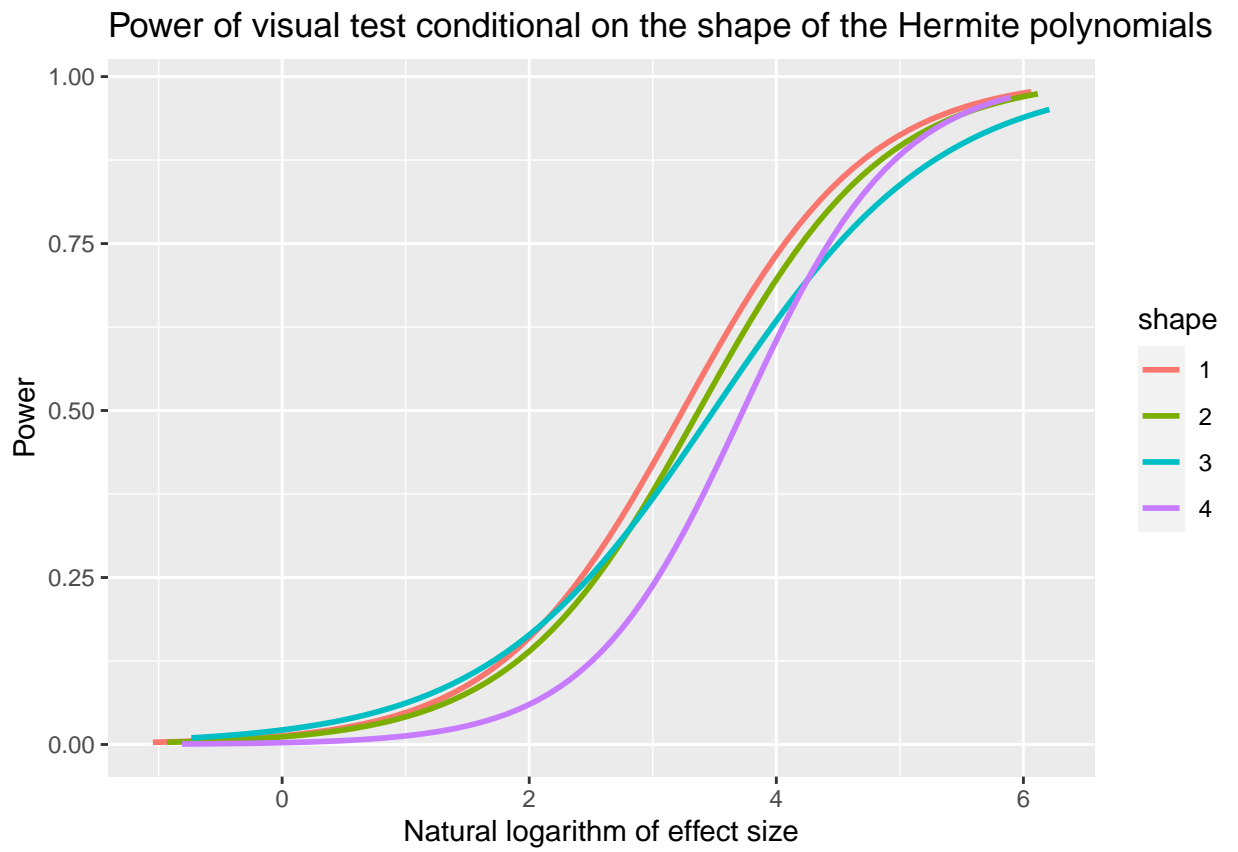
Power comparison conditional on the distribution of X
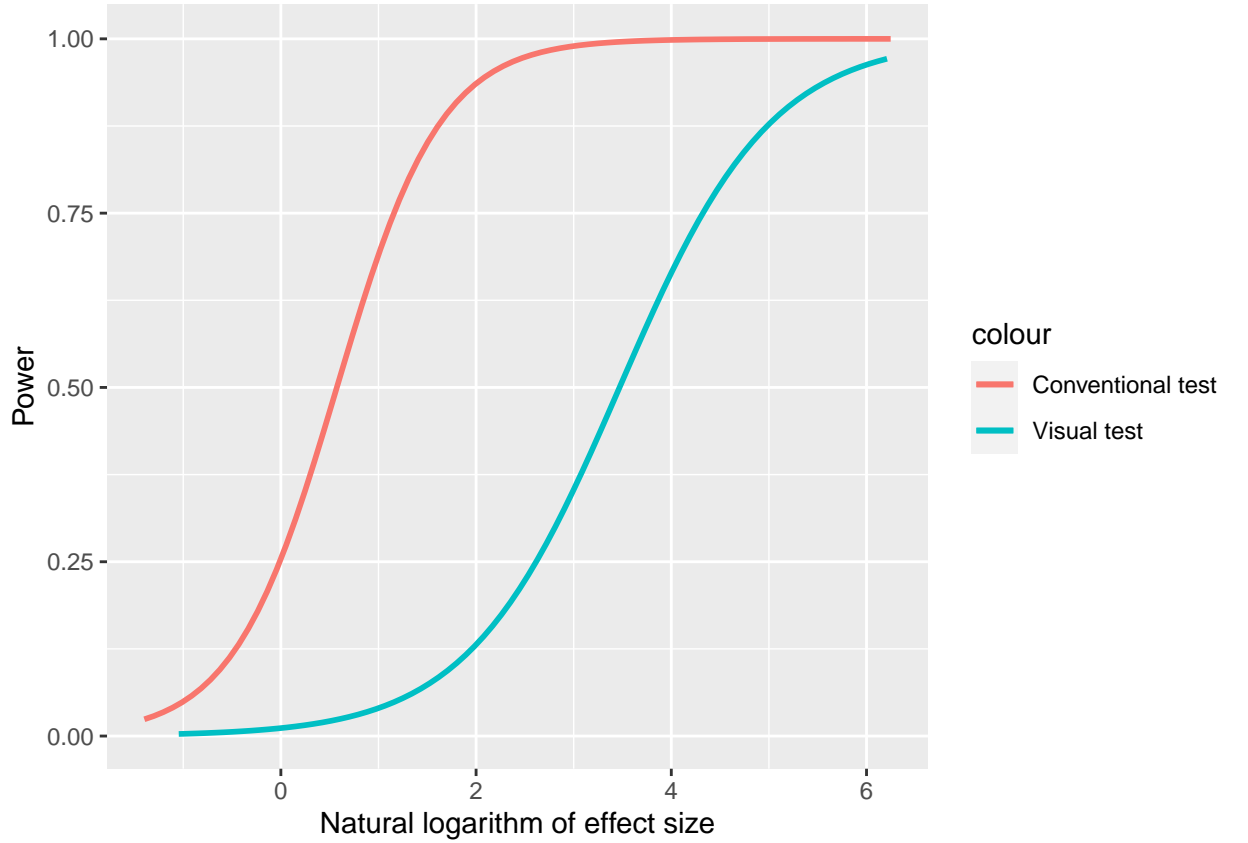
Power of visual test conditional on the distribution of X

Power comparison conditional on the shape of the Hermite polynomials
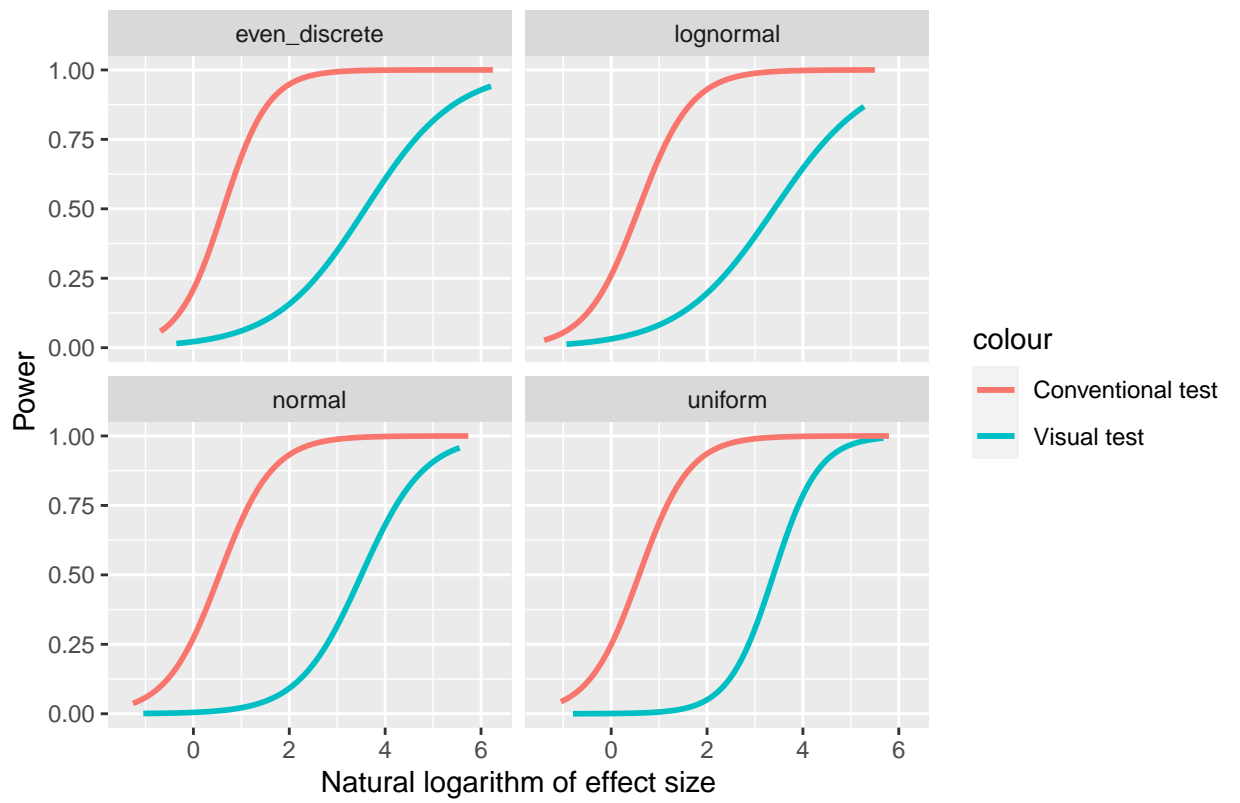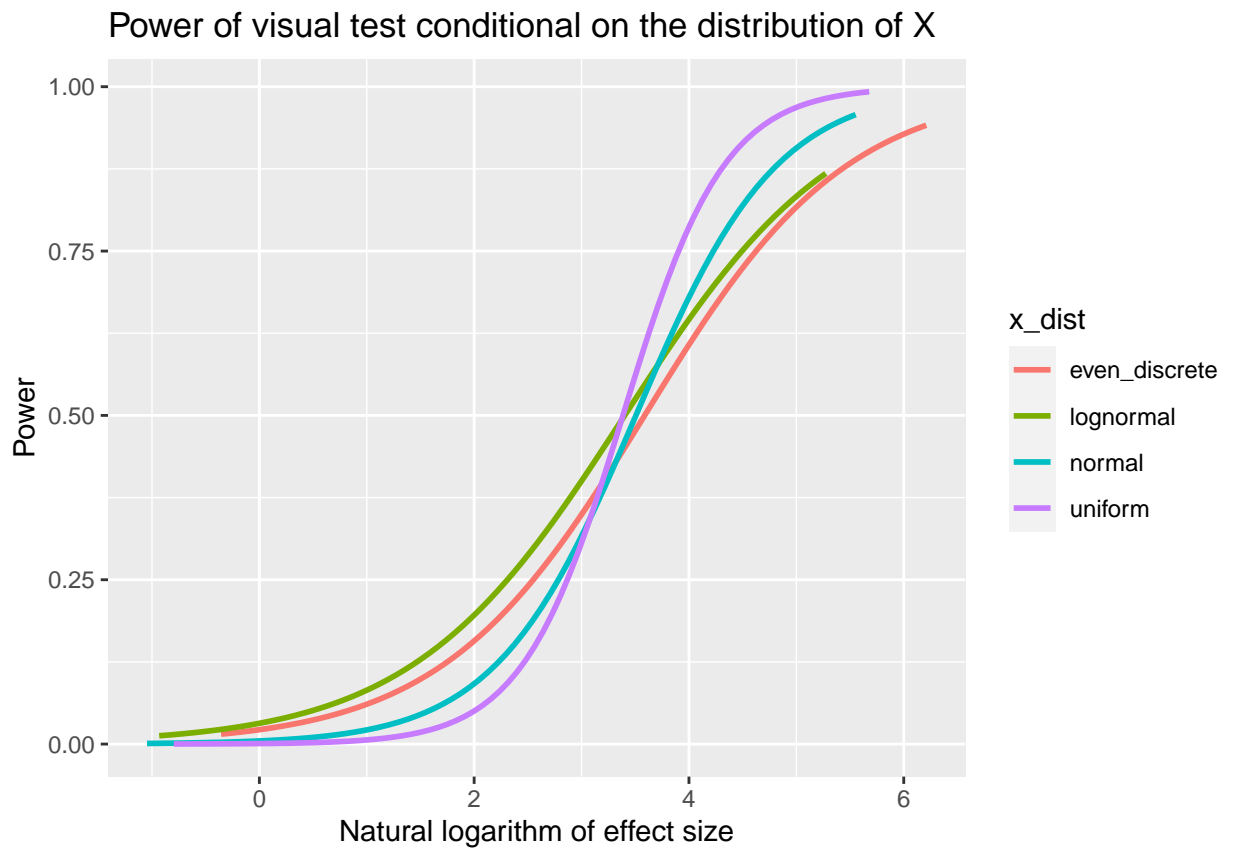
Power of visual test conditional on the shape of the Hermite polynomials

Power comparison conditional on the number of observations

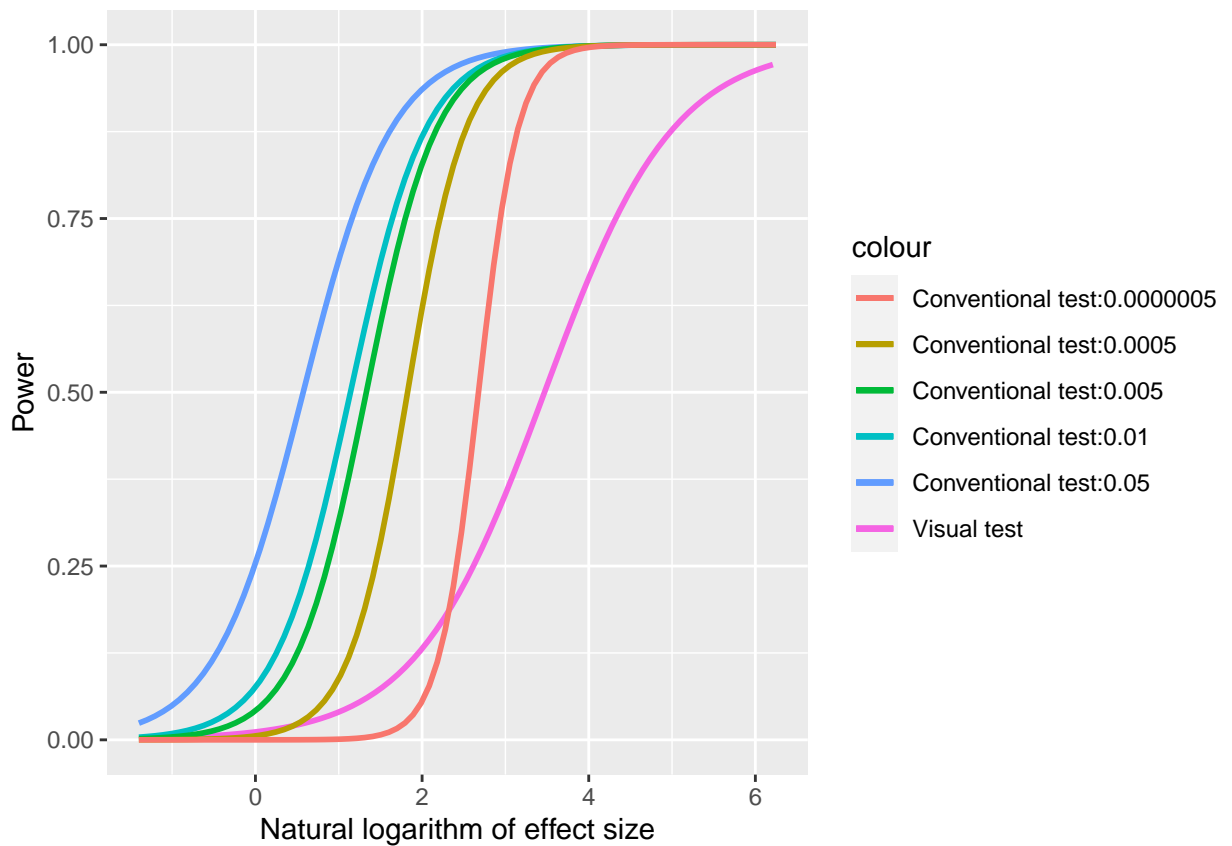Power of visual test conditional on the number of observations
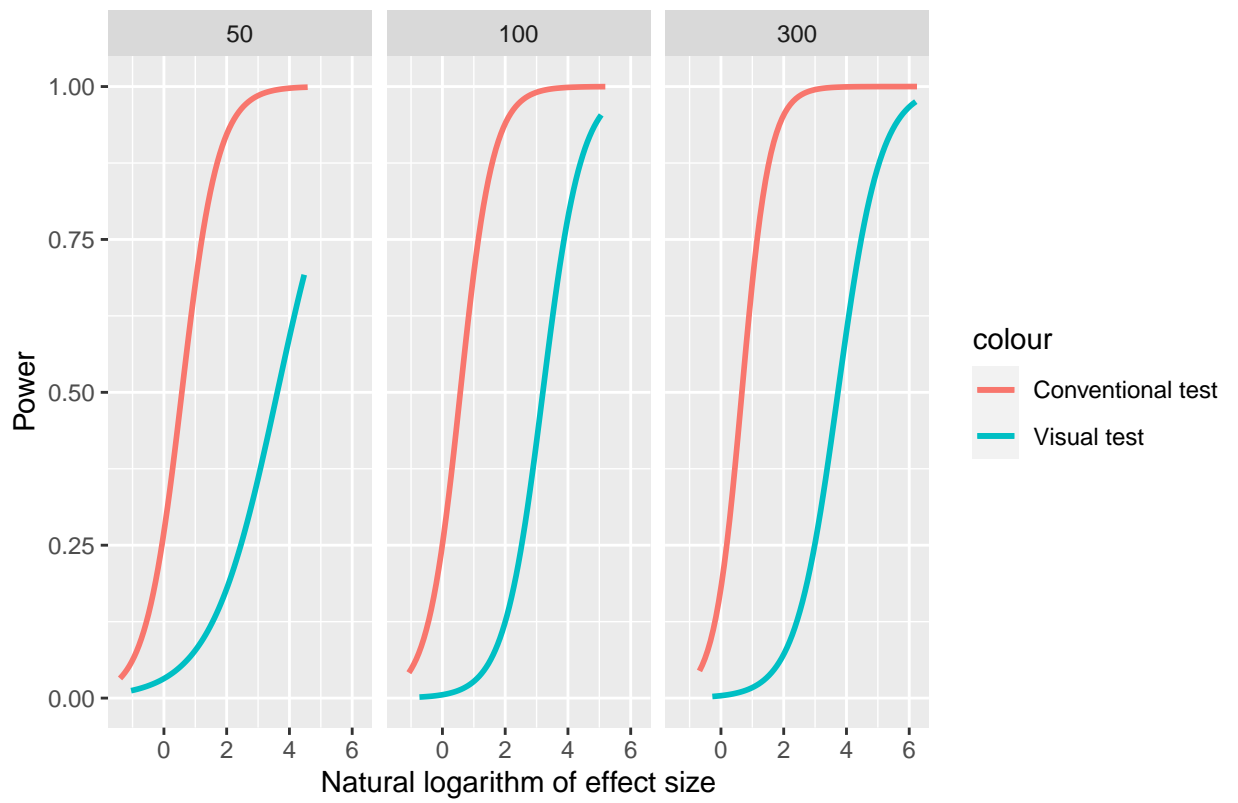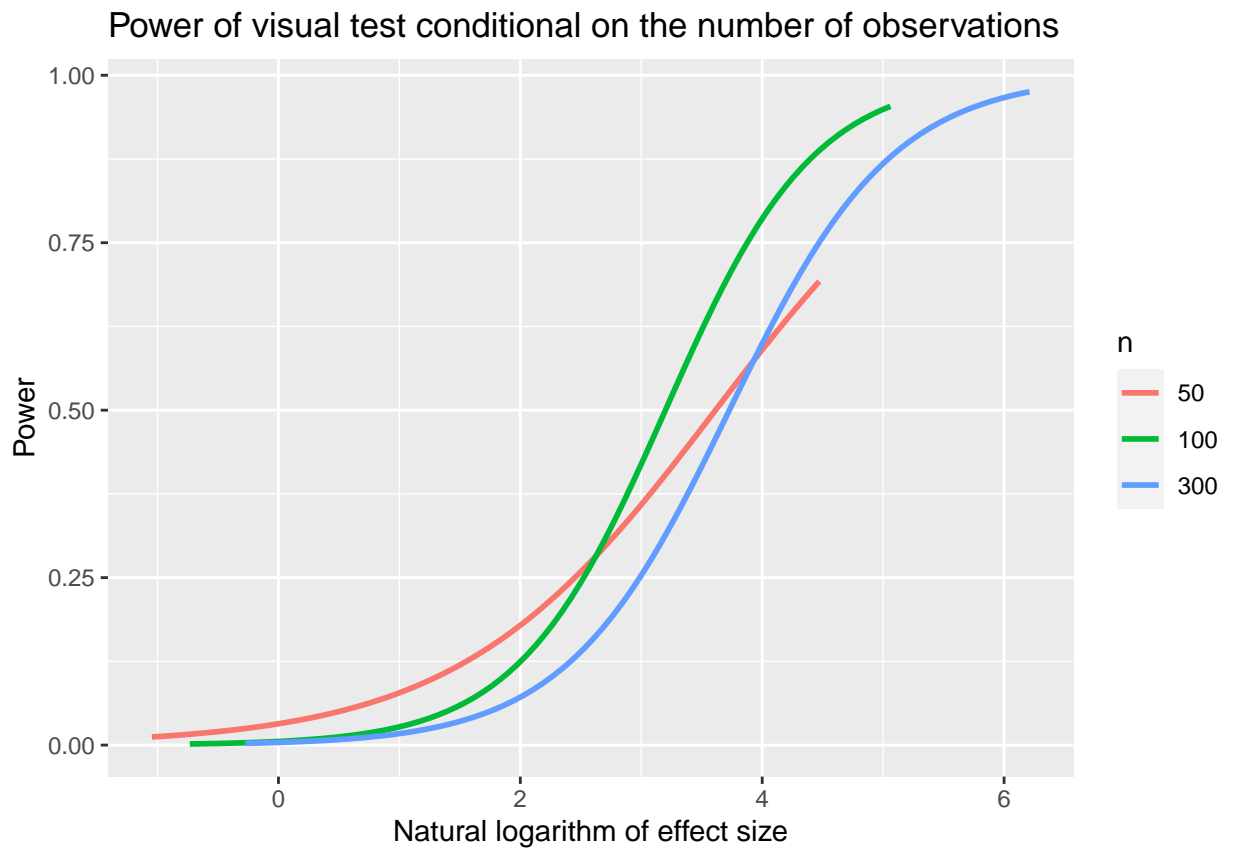
References

Anscombe, F. J., and John W. Tukey. 1963. "The Examination and Analysis of Residuals." *Technometrics* 5 (2): 141–160.

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

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 Beat Kleiner. 1975. "A graphical technique for enhancing scatterplots with moving statistics." *Technometrics* 17 (4): 447–454.

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.

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.

Gunst, Richard F, and Robert L Mason. 1980. *Regression Analysis and its Application: A Data-Oriented Approach.* Vol. 34. CRC Press.

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.

Mansfield, Edward R, and Michael D Conerly. 1987. "Diagnostic value of residual and partial residual plots." *The American Statistician* 41 (2): 107–116.

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.

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.