ARTICLE TEMPLATE

Why shouldn't you use numerical tests to diagnose the linear regression models?

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

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

ARTICLE HISTORY
Compiled July 7, 2022

ABSTRACT
Abstract to fill.

KEYWORDS
visual inference; regression diagnostics;

1. Introduction

Regression analysis is a field of study with at least a hundred years of history, and regression diagnostics is one of the essential steps in regression analysis. 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 (2014), Montgomery, Peck, and Vining (2012), 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

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

identify patterns that are indicative of nonconstant error variance or non-linearity. Raw residuals and studentized residuals are the two most frequently used residuals 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 2018). 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. Therefore, 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 a 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*

In addition to diagnostic plots, analysts may perform formal hypothesis testing for detecting model defects. Depends on the alternative hypothesis that being focused on, 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 both derived from the Lagrange multiplier test (Silvey 1959) principle and rely on the asymptotic properties of the null distribution. For testing non-linearity, one may apply the F-test to examine the significance of certain 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 also known as the "Type III error". Additionally, outliers will often incorrectly trigger the rejection of the null hypothesis despite the residuals being 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, Peck, and Vining (2012) stated 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". The outcome "reject" still 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. There are situations where no suitable diagnostic plots can be found for a particular violation of the assumptions, or excessive diagnostic plots need to be checked. One will have no choice but to rely on statistical tests if there are 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. 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 report 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

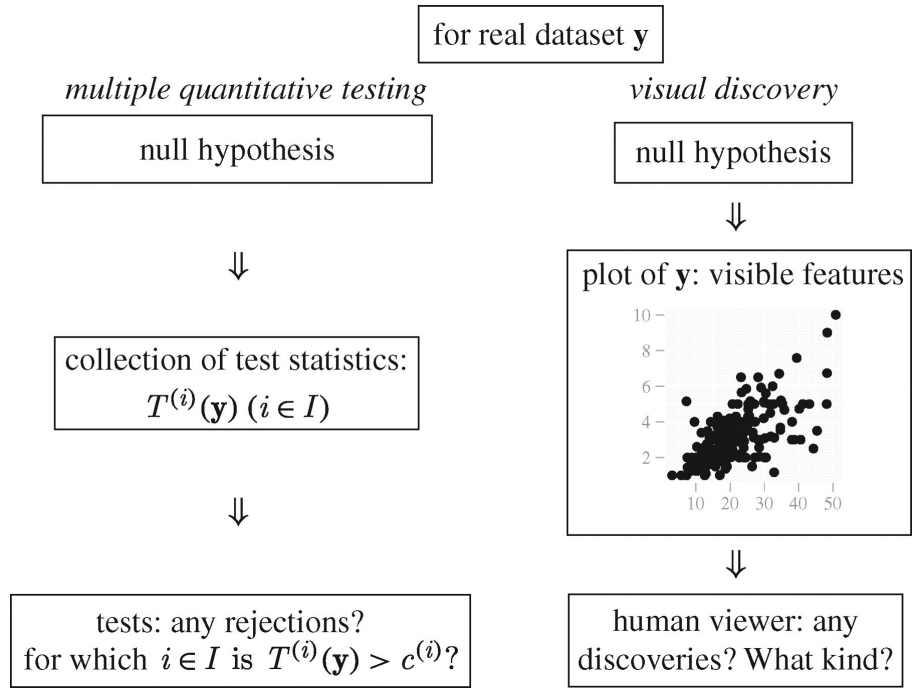


Figure 1. Parallelism between multiple quantitative testing and visual discovery (Buja et al. 2009). Visible features in a plot are viewed as a collection of test statistics $T^{(i)}(\mathbf{y})$ $(i \in I)$, and any visual discoveries that are inconsistent with the null hypothesis are treated as evidence 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.

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 @ref(fig:residual-plot-cubic-heter).

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

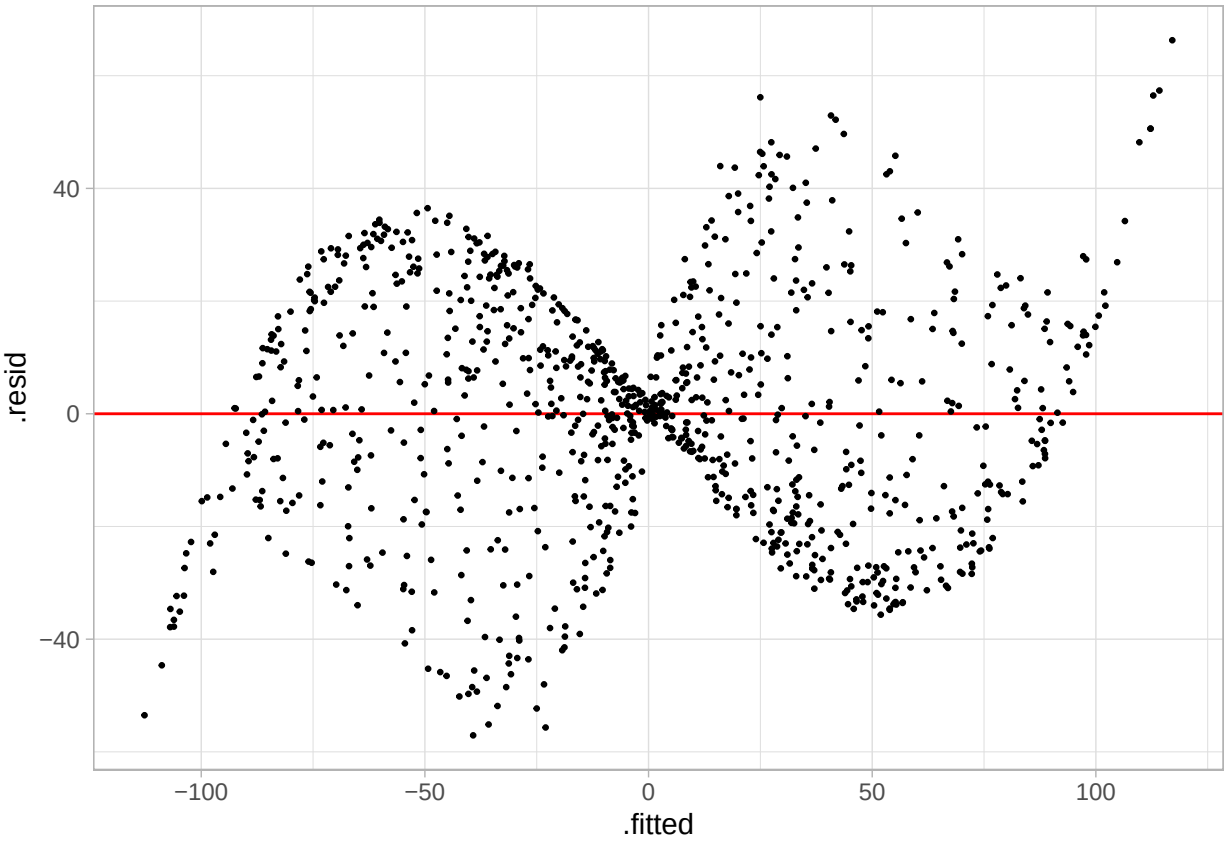


Figure 2. Residuals vs. fitted values plot for a classical linear regression model. The residuals are produced by fitting a two-predictor multiple linear regression model with data generated from a cubic linear model. From the residual plot, "butterfly shape" can be observed which generally would be 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.

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). 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 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 $P(\text{there exists } i \in I : T^{(i)}(\mathbf{y}) > c^{(i)}|\mathbf{y}) \leq \alpha$, 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, if we involve K independent observers in a visual test, and let X be a random variable denoting the number of observers correctly picking the actual data plot. Then, under the null hypothesis $X \sim \text{Binom}_{K,1/m}$, and 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.

Compared to the conventional test, whose power only depends on the parameter of interest θ, several studies (see Hofmann et al. 2012; Majumder, Hofmann, and Cook 2013; **?**; Roy Chowdhury et al. 2015; Loy, Follett, and Hofmann 2016) have shown the power of the visual test is subject-specific. Thus, to be able to account for individual's ability, an individual is required to evaluate multiple lineups (Majumder, Hofmann, and Cook 2013).

Suppose individuals have the same ability and a lineup has been evaluated by multiple individuals, under the alternative hypothesis, the estimated power for a lineup can be expressed as $\hat{p} = x/K$, the estimated probability of identifying the actual data plot from the lineup. If the individual skill needs to be taken into account, and L lineups have been evaluated by K individuals, Majumder, Hofmann, and Cook (2013) suggests that mixed effects logistic regression model can be fit as:

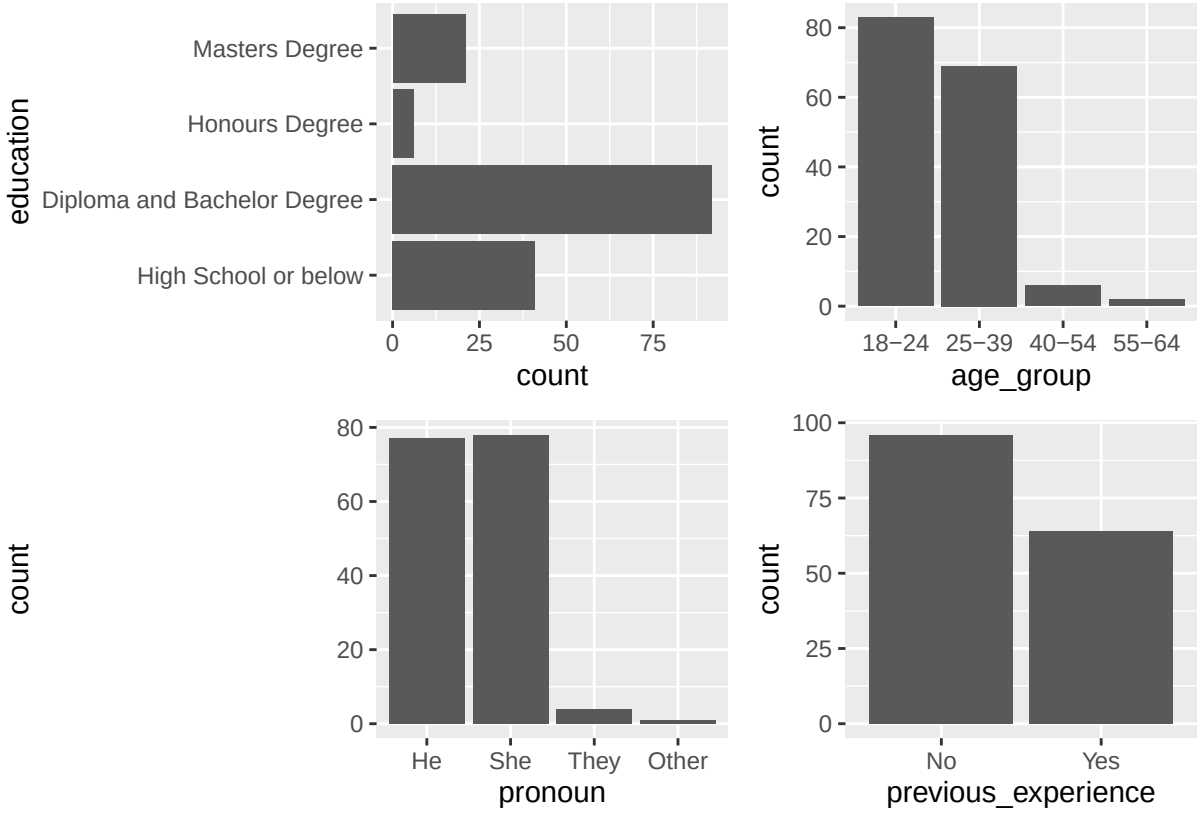
$$g(p_{li}) = W_{li}\delta + Z_{li}\tau_{li},$$

where $g(.)$ is the logit link function $g(p) = log(p) - log(1-p)$; $0 \leq p \leq 1$. W_{li}, $1 \leq i \leq K$, $1 \leq l \leq L$, is the covariate matrix including lineup-specific elements and demographic information of individuals, and δ is a vector of parameters. Z is the random effects matrix, and τ is a vector of variables follow $N(\mathbf{0}, \sigma_\tau \boldsymbol{I}_{KL\times KL})$.

Then, the estimated power for lineup l and individual i can be calculated as $\hat{p}_{li} = g^{-1}(W_{li}\hat{\delta} + Z_{li}\hat{\tau}_{li})$ (Majumder, Hofmann, and Cook 2013).

2. Experimental design

3. Demographic summary



4. Data processing

5. Results

5.1. *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.

5.2. *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.

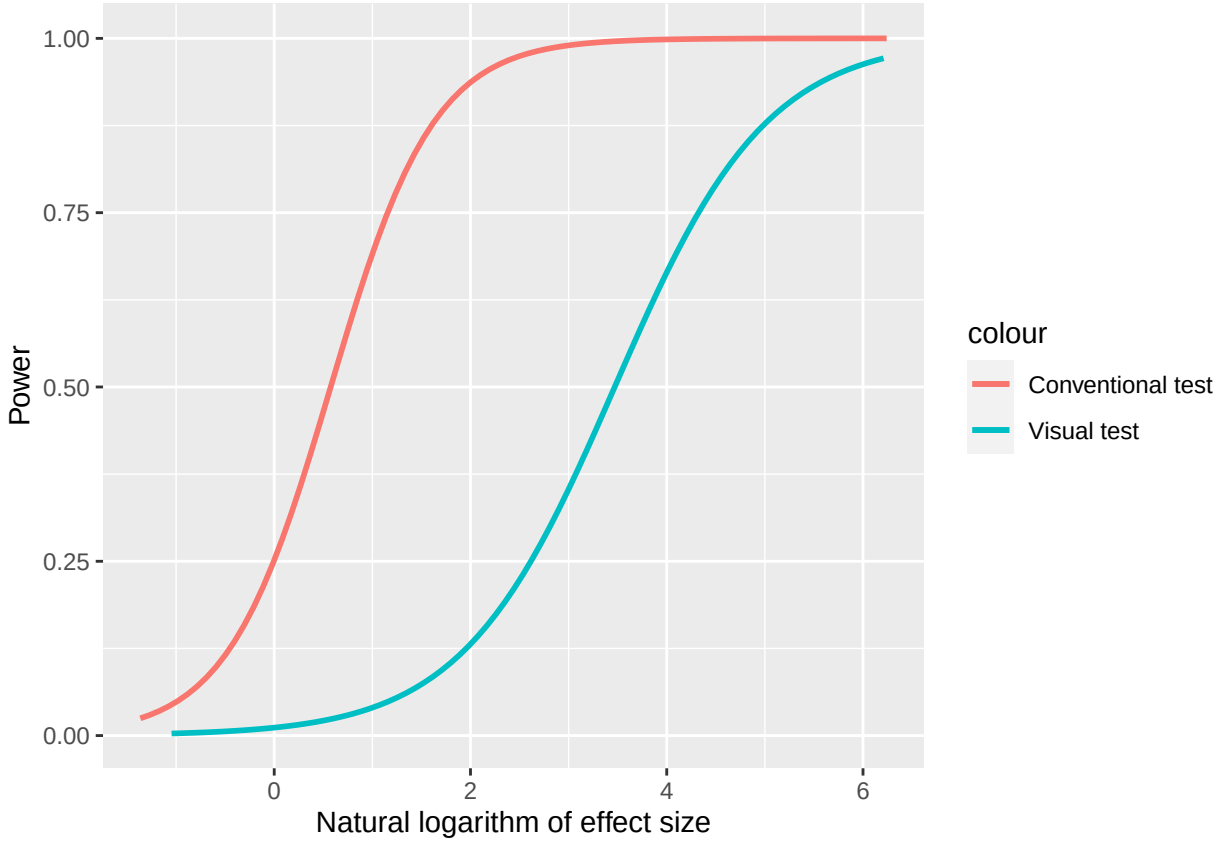
It was also found that there was a noticeable gap between curves of the conventional test and the visual test at around $log(\text{effect size}) = 0$ for the cubic model and $log(\text{effect size}) = 2.5$ for the heteroskedasticity model, where the differences in power were greater than 0.6. We further 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.

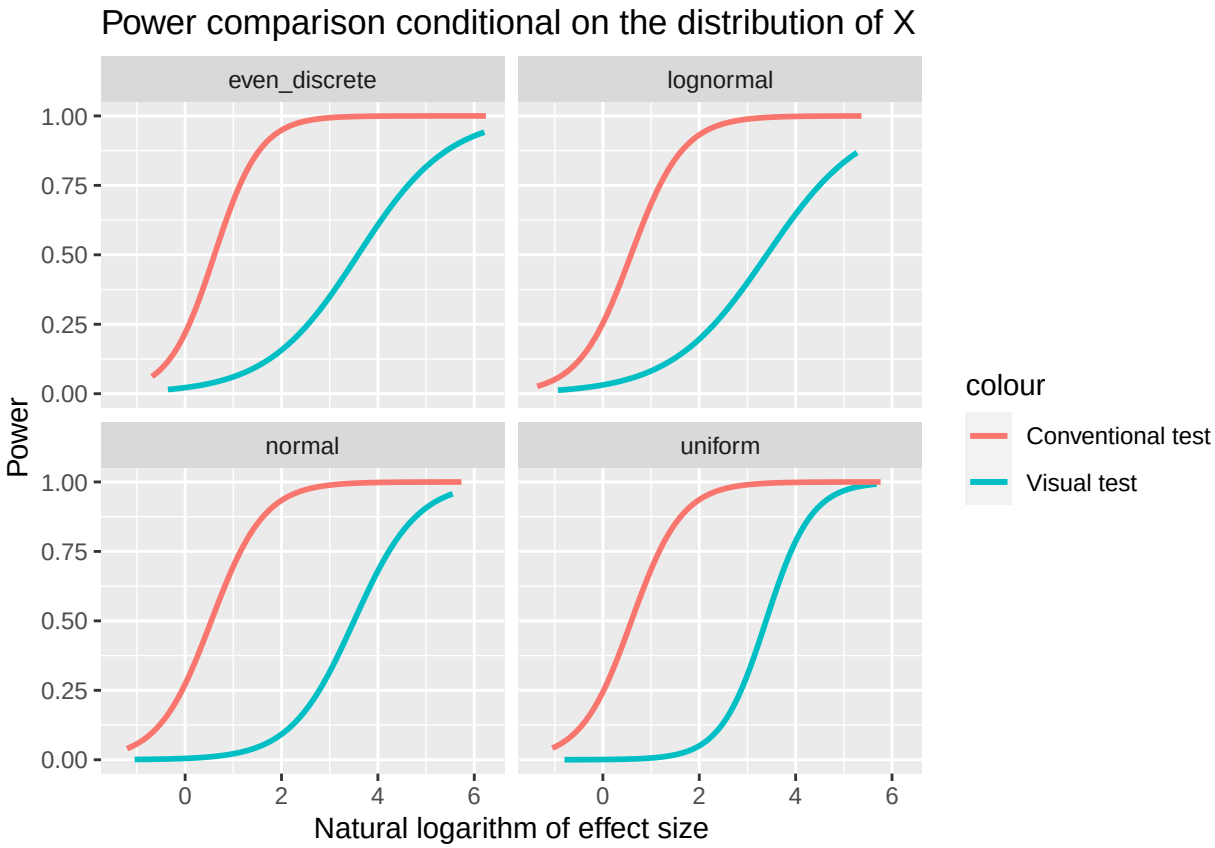
5.3. *Effect of parameters on power of the visual test*

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.

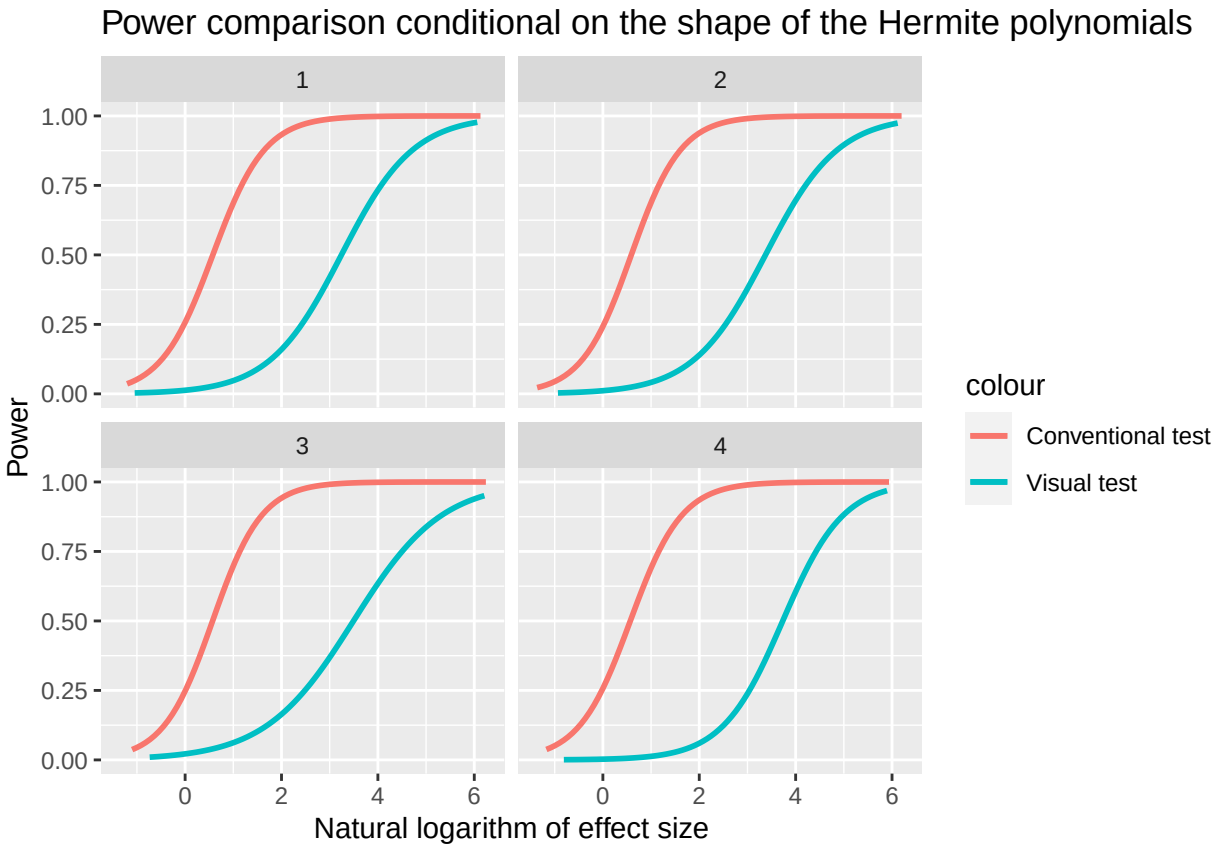
In cubic model, two major factors that influencing the strength of the signal are a and b. Figure **??** and **??** illustrates 30 different logistic regressions fit for different

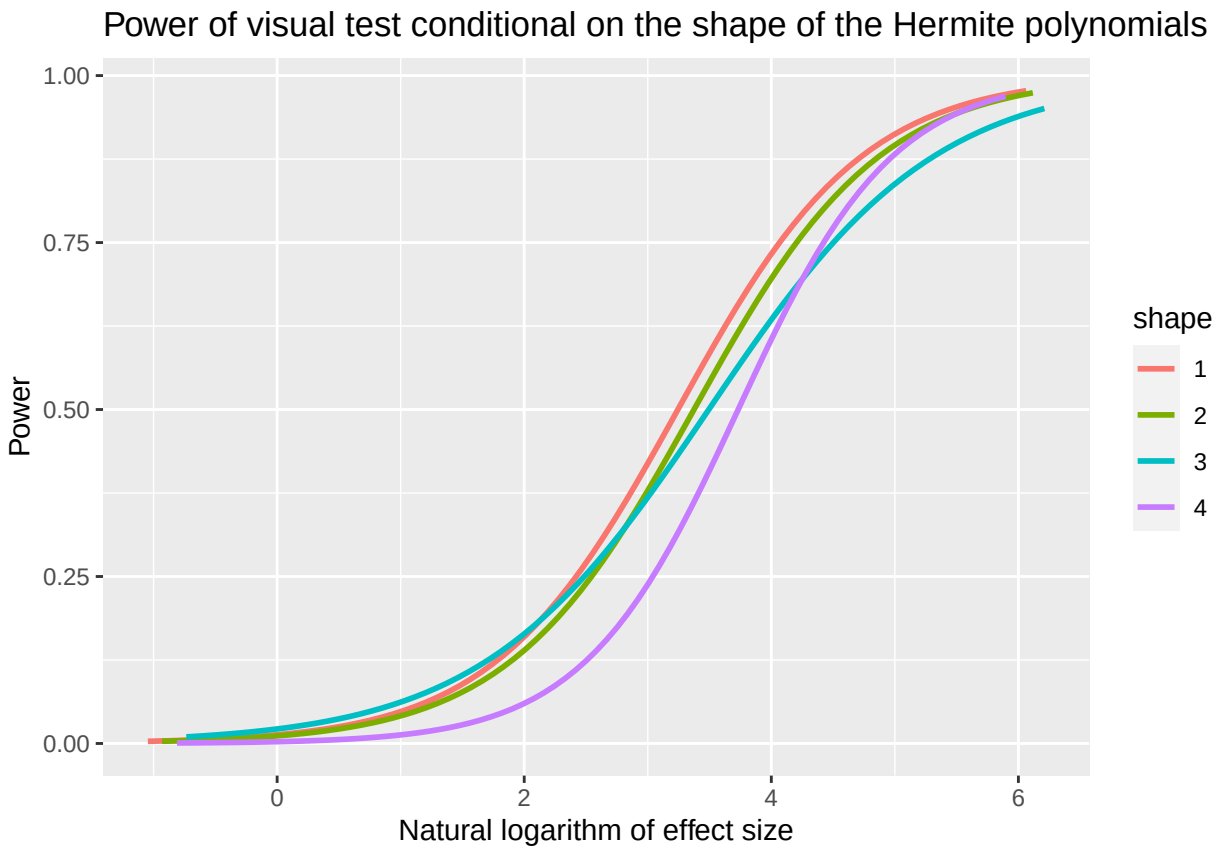
number of evaluations and different number of observations n. The regressor used in these models was $|a|/\sigma$ since the noise level σ needed to be taken into account. From the figures, we can observe ...

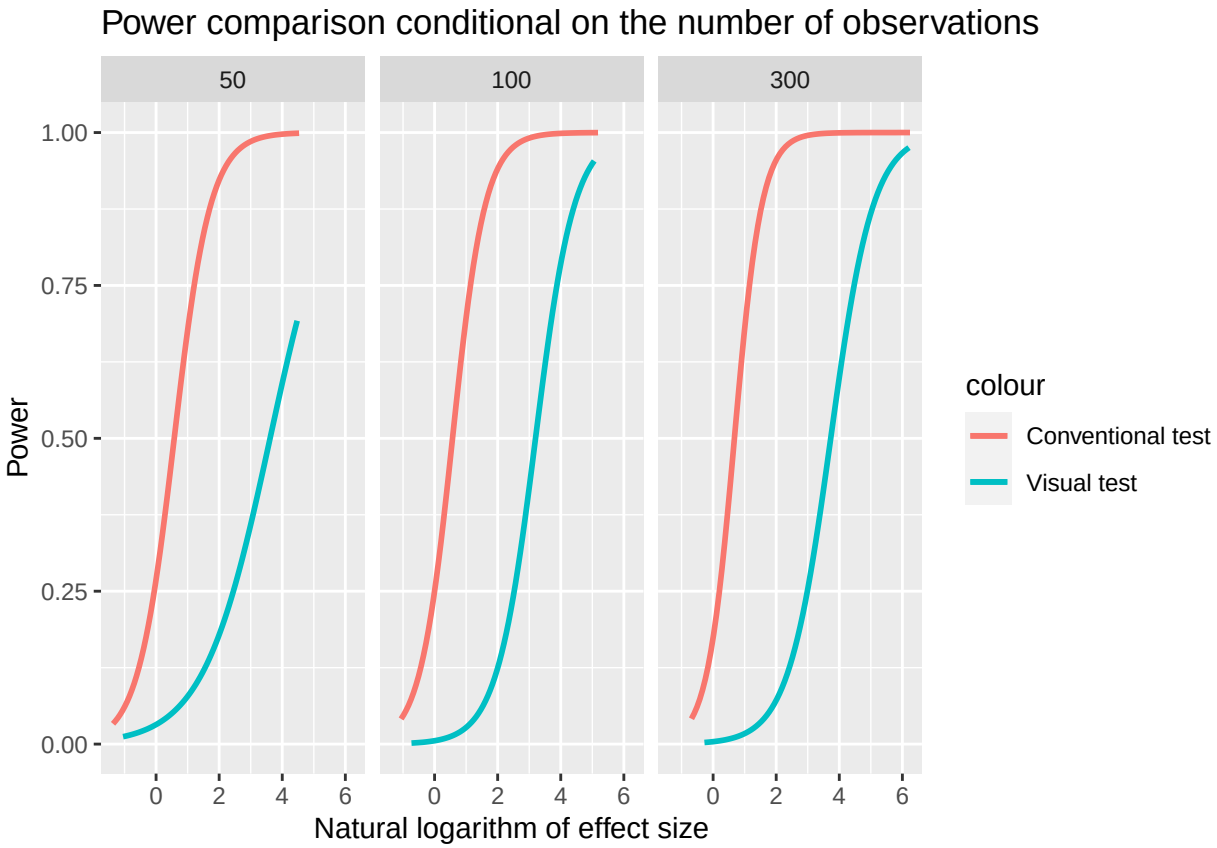


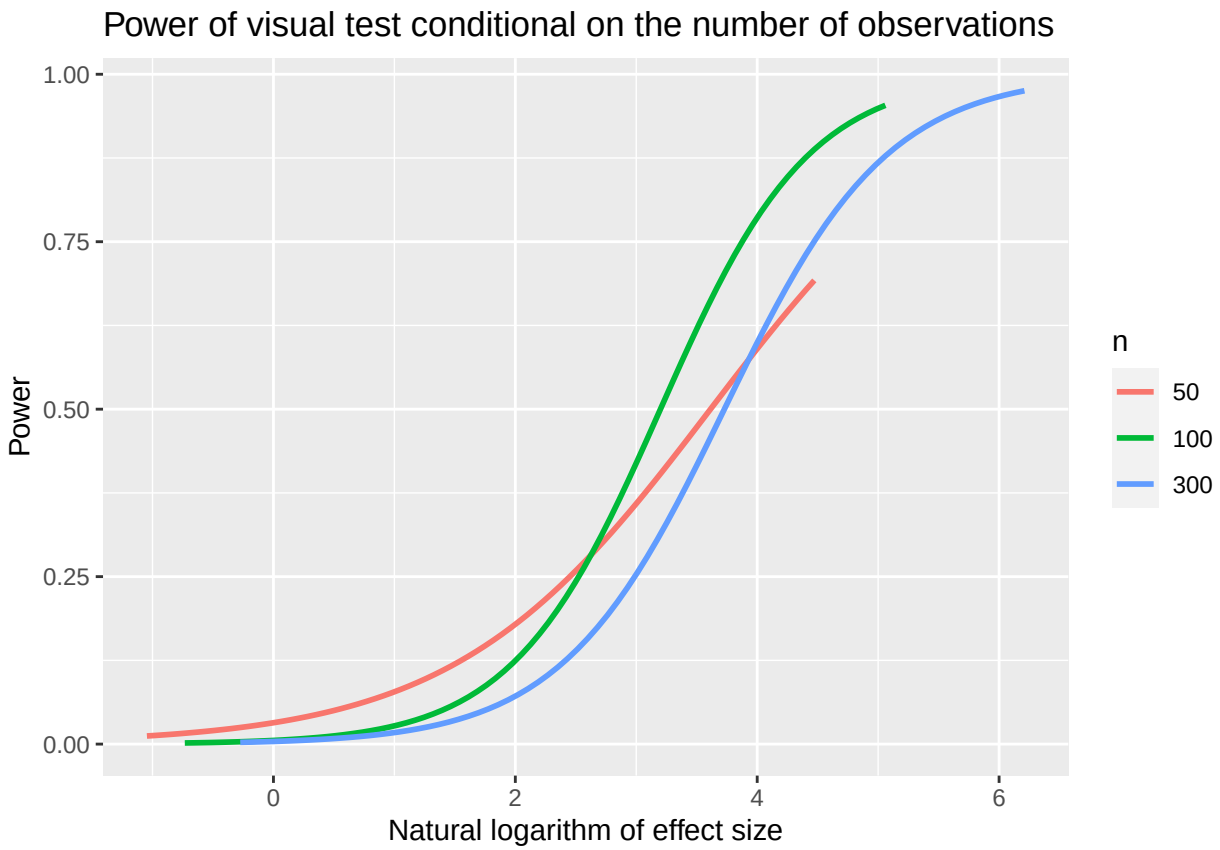

Power comparison conditional on the distribution of X
even_discrete
lognormal
normal
uniform
Power
1.00
0.75
0.50
0.25
0.00
0
2
4
6
Natural logarithm of effect size
colour
Conventional test
Visual test


Power of visual test conditional on the distribution of X
Power
1.00
0.75
0.50
0.25
0.00
0
2
4
6
Natural logarithm of effect size
x_dist
even_discrete
lognormal
normal
uniform


Power comparison conditional on the shape of the Hermite polynomials
1
2
3
4
1.00
0.75
0.50
0.25
0.00
Power
0
2
4
6
Natural logarithm of effect size
colour
Conventional test
Visual test


Power of visual test conditional on the shape of the Hermite polynomials
1.00
0.75
0.50
0.25
0.00
Power
0
2
4
6
Natural logarithm of effect size
shape
1
2
3
4


Power comparison conditional on the number of observations
50
100
300
1.00
0.75
0.50
0.25
0.00
Power
0
2
4
6
Natural logarithm of effect size
colour
Conventional test
Visual test



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. 2014. *Applied regression analysis*. 3rd ed. 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. 2018. *Regression analysis and its application: a data-oriented approach*. CRC Press.

Hofmann, Heike, Lendie Follett, Mahbubul Majumder, and Dianne Cook. 2012. "Graphical Tests for Power Comparison of Competing Designs." *IEEE Transactions on Visualization and Computer Graphics* 18 (12): 2441–2448.

Loy, Adam, Lendie Follett, and Heike Hofmann. 2016. "Variations of Q–Q Plots: The Power of Our Eyes!" *The American Statistician* 70 (2): 202–214.

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, Douglas C, Elizabeth A Peck, and G Geoffrey Vining. 2012. *Introduction to linear regression analysis*. 5th ed. John Wiley & Sons.

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.

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