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; model 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).

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 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

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

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.

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 (ref here) principle and rely on the asymptotic properties of the null distribution. For testing non-linearity, one may apply the F-test (ref here) 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).

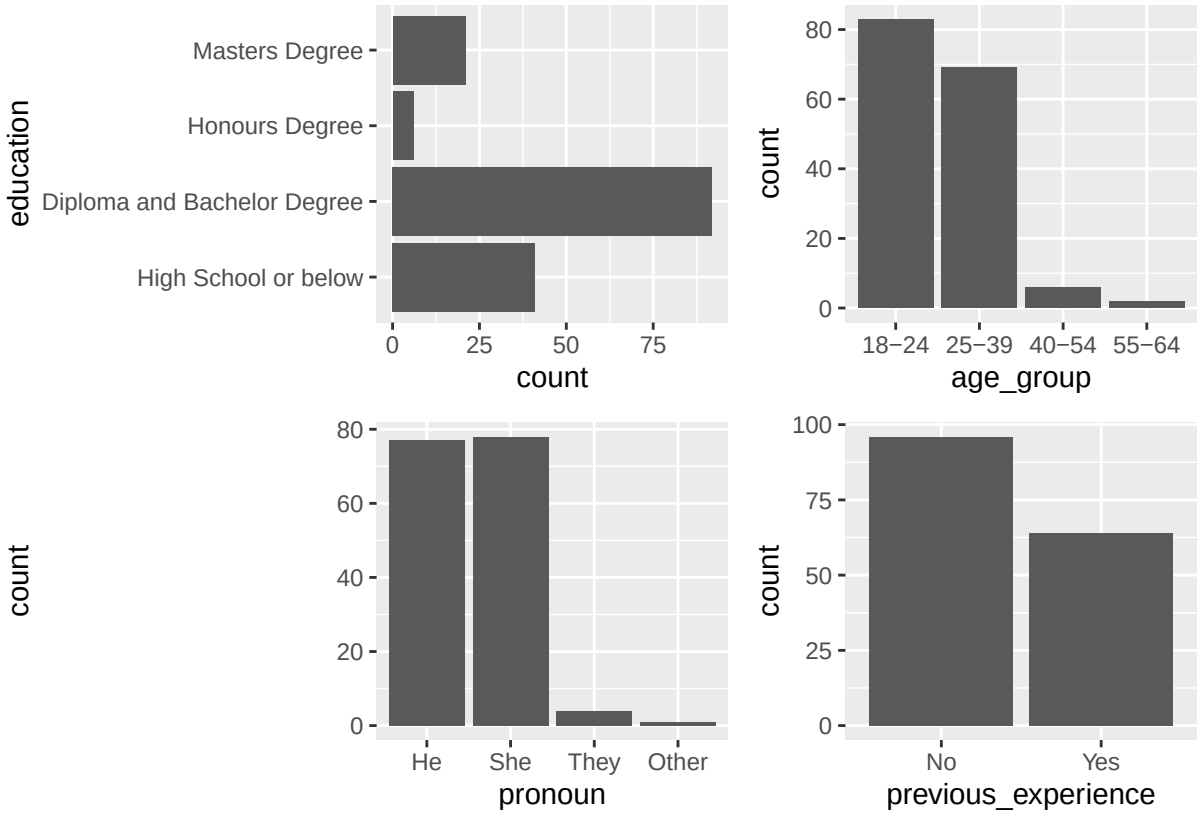
As discussed in Cook and Weisberg (1982), most residual-based tests for a particular type of departures from model assumptions are sensitive to other types of departures. It is likely the null hypothesis is correctly rejected but for the wrong reason, which is also known as the "Type III error". Additionally, outliers will often incorrectly trigger the rejection of 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, 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 statistic 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 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 is any. A good regression diagnostic practice should be a balanced combination of both methods.

Diagnostic plots are

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.

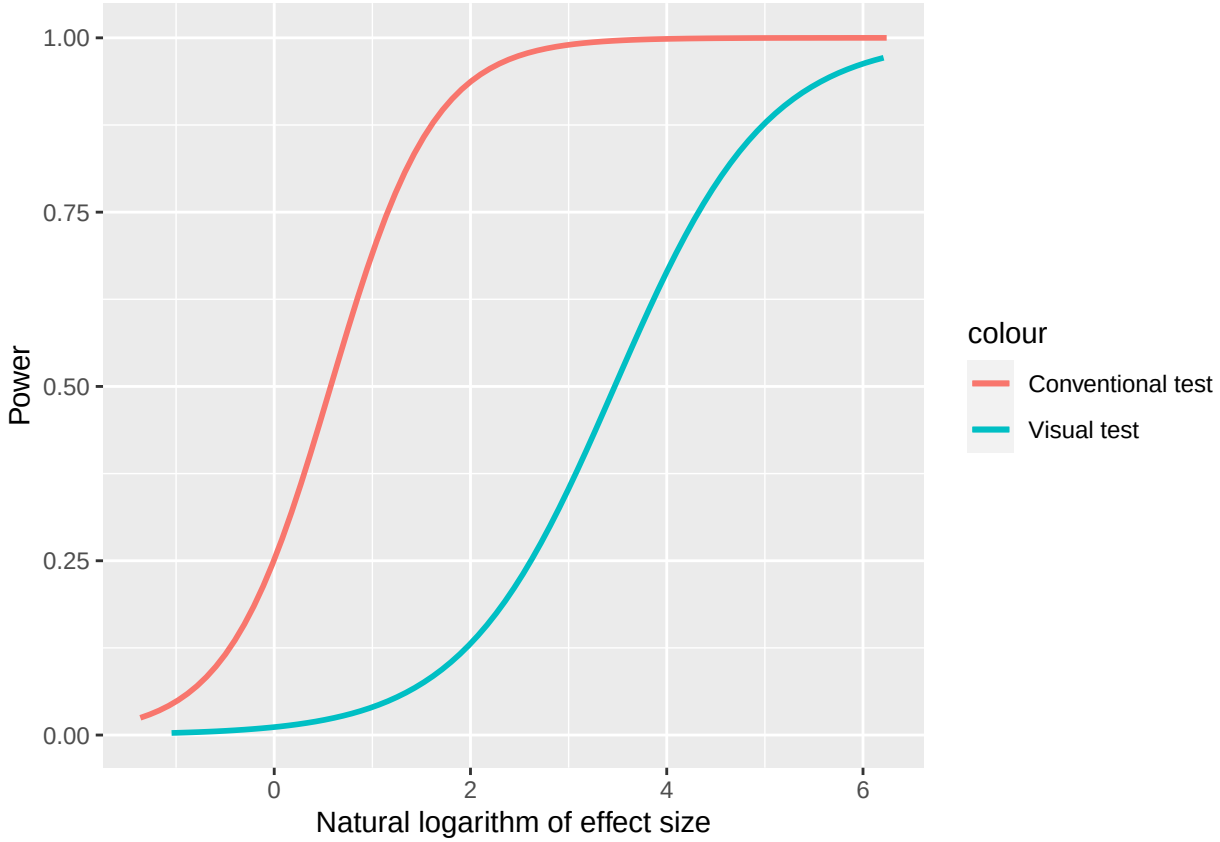
As shown in Figure **??**, both F-test and Breusch–Pagan test generally possessed greater power than visual test. A visual tests is a collection of test against any alternatives that would create visual discoverable features, while a conventional test is usually targeting at a pre-specified alternative. 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.

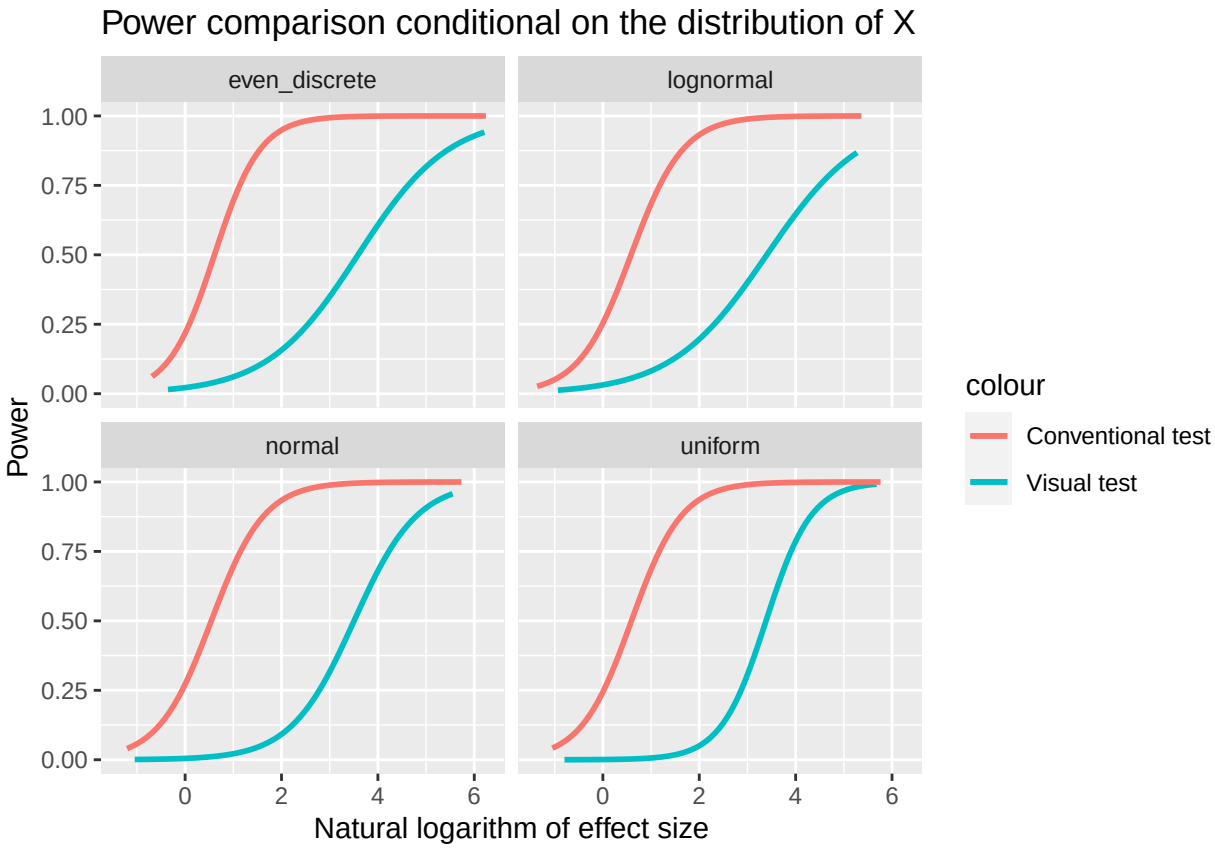
It was also found that there was a noticeable gap between curves of the conventional test and the visual test at around $log(\text{effect size}) = 0$ for the cubic model and $log(\text{effect size}) = 2.5$ for the heteroskedasticity model, where the differences in power were greater than 0.6. We further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

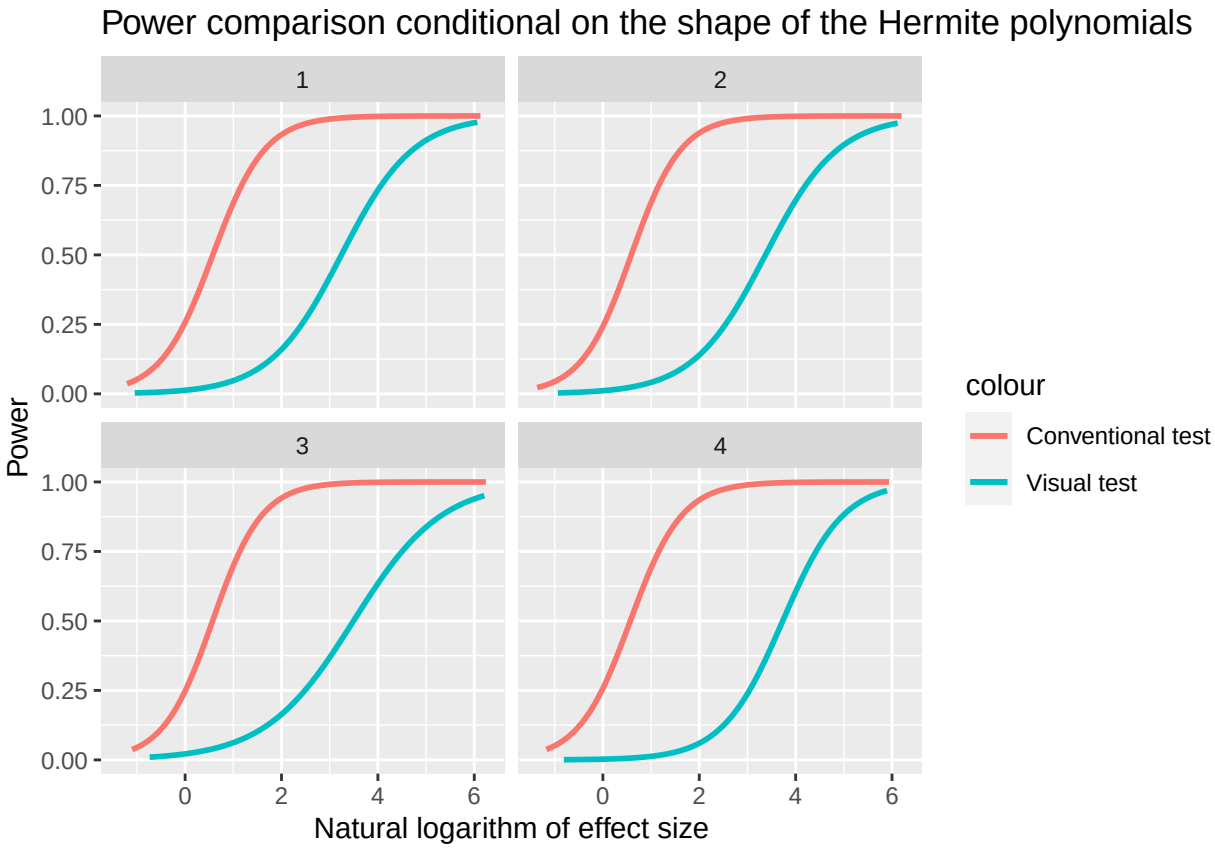
The previous section focuses on the change of effect size relative to the power of the visual test. However, effect size is only a one dimensional summarisation of parameters used in data simulation. Individual factor embedded in the simulation process should also be analysed.

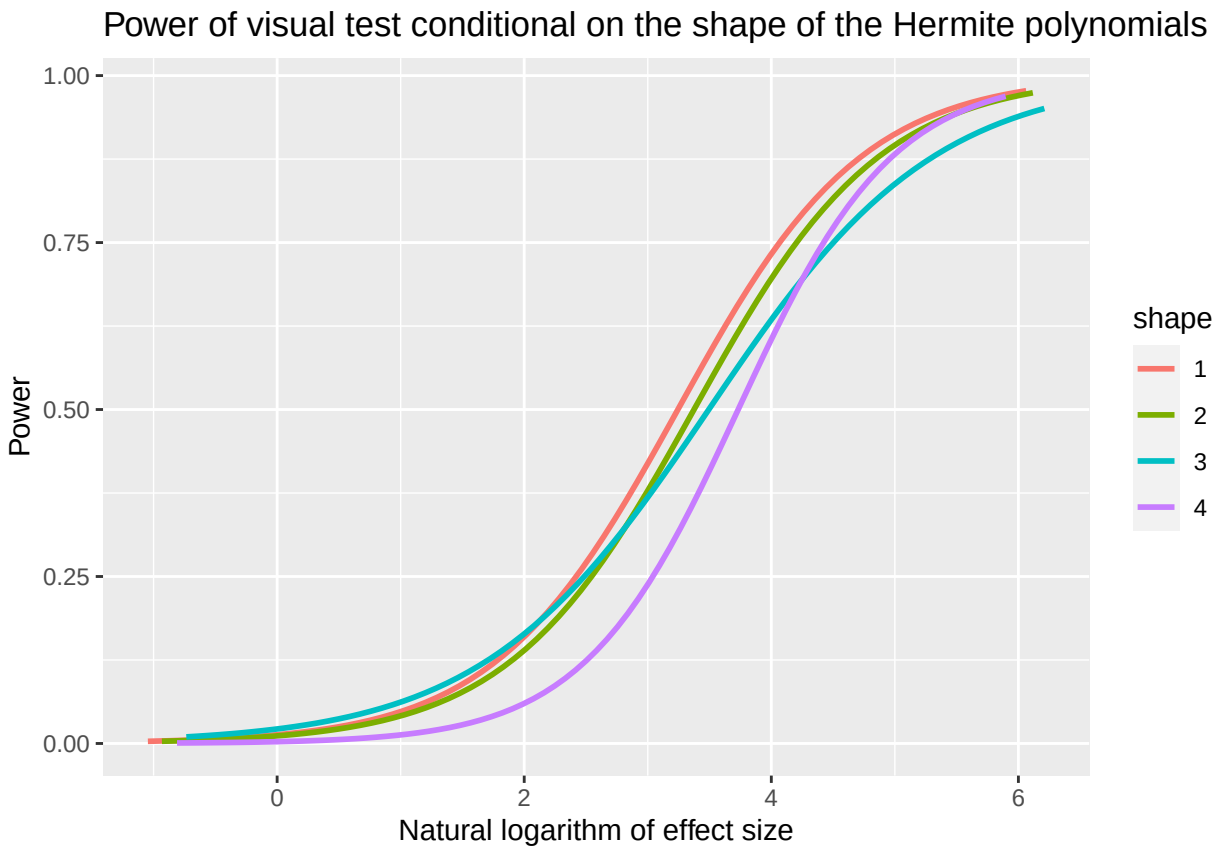
In cubic model, two major factors that influencing the strength of the signal are a and b. Figure **??** and **??** illustrates 30 different logistic regressions fit for different number of evaluations and different number of observations n. The regressor used in these models was $|a|/\sigma$ since the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

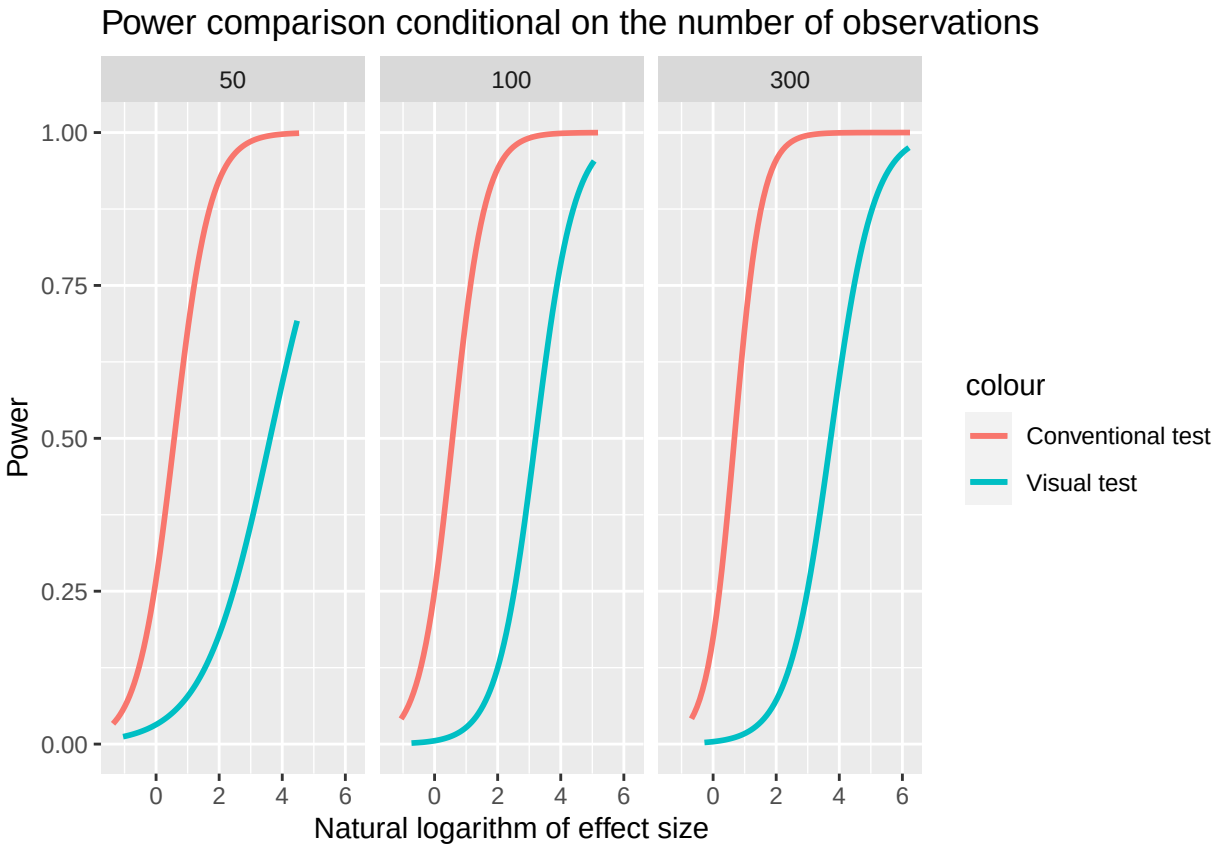


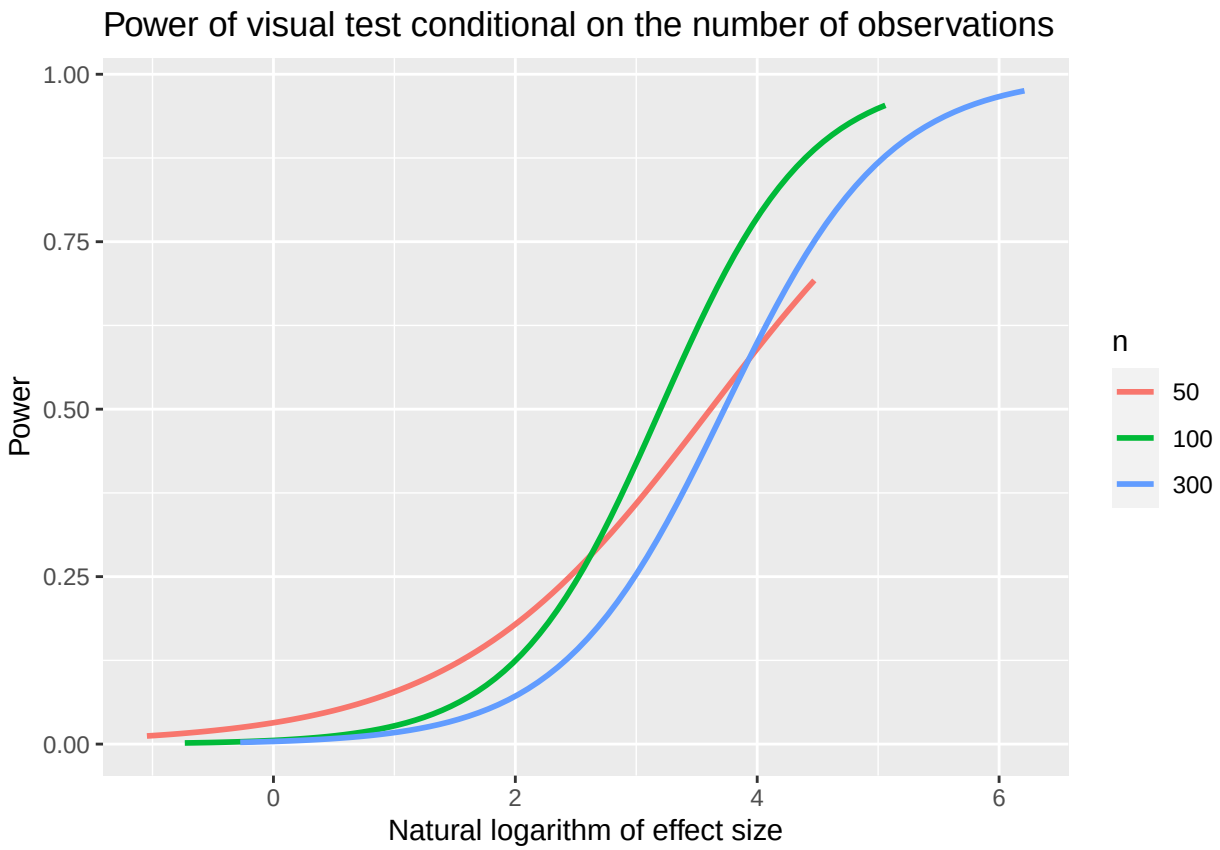

Power comparison conditional on the distribution of X
even_discrete
lognormal
normal
uniform
Power
1.00
0.75
0.50
0.25
0.00
0
2
4
6
Natural logarithm of effect size
colour
Conventional test
Visual test


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. Accessed 2022-04-24. `http://www.tandfonline.com/doi/abs/10.1080/00401706.1963.10490071`.

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. Publisher: [Wiley, Econometric Society], Accessed 2022-04-26. `http://www.jstor.org/stable/1911963`.

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

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.

Gunst, Richard F, and Robert L Mason. 2018. *Regression analysis and its application: a data-oriented approach*. CRC Press.

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. Publisher: [Royal Statistical Society, Wiley], Accessed 2022-04-26. http://www.jstor.org/stable/2984219.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (4): 817–838. Publisher: [Wiley, Econometric Society], Accessed 2022-04-26. http://www.jstor.org/stable/1912934.