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

ABSTRACT
Abstract to fill.

problem: residual plot diagnostics conventional test: too sensitive
background:

1. residual plot for model diagnostics

a. residual is widely used
b. what are the types of residual plots
c. comparison

2. conventional test: F, BP
3. visual test: lineup, theory

desc of experiment: 1. simulation setup 2. experimental design 3. result
comparison of conventional tests: 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)
conclusion: 1. too sensitive, visual test is needed/preferable 2. visual test is infeasible in large scale (expensive) 3. future work (role of computer vision)

1. Introduction

Regression diagnostics conventionally involve evaluating the fitness of the proposed model, detecting the presence of influential observations and outliers, checking the validity of model assumptions and many more. Common diagnostic techniques including summary statistics, hypothesis testing, and data plots are essential tools for a

systematic and detailed examination of the regression model (Mansfield and Conerly 1987).

1.1. *Diagnostic plots*

Regression analysis is a field of study with at least a hundred years of history. 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 of the book, one will find the importance and usefulness of diagnostic plots being emphasized again and again. Checking diagnostic plots is also the recommended starting point for validating model assumptions like 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 refereed to as standard residual plots. They are commonly used to identify patterns which are indicative of nonconstant error variance or non-linearity (Cook and Weisberg 1982). 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 always present. While raw residuals are the most common output of computer regression software package, by applying a scaling factor, the ability of revealing 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 position of the 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.

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 relatively large sample size, Cleveland and Kleiner (1975) suggests the use of robust moving statistics as reference lines to give aids to 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, researcher may also perform formal tests for detecting model defects. Depends on the alternative, variety of tests can be applied. For example,

for testing heteroskedasticity, one may use the White test (White 1980) or the Breusch-Pagan test (Breusch and Pagan 1979). And for testing non-linearity, there are RESET test (Ramsey 1969) and F-test.

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. Especially, 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. Furthermore, Montgomery, Peck, and Vining (2012) stated that based on their experience, statistical tests are not widely used in regression diagnostics. Most importantly, the same or even larger amount of information can be provided by diagnostic plots than the corresponding tests in most empirical studies. But still, the effectiveness of statistical tests shall not be disrespected. Statistical tests have chance to provide analysts with unique information. There are also 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 combination of both methods.

1.3. *Visual inference*

Diagnostic plots are

2. Experimental design

3. Data processing

4. Results

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.

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)

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. To

explore the difference in power between conventional test and visual test, ten logistic regression models are fit to the data, with natural logarithm of the effect size as the only fixed effect, and whether the test successfully rejects the null hypothesis as the response variable. Table 2 and 2 summarizes the results. The natural logarithm of the effect size is positive and significant across all models. Figure 1 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 same model setting used in two experiments and 500000 samples are drawn for both cubic and heteroskedasticity model.

From Figure 1, 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 curves of visual tests with evaluations greater than two were always above the curves of visual test with evaluations smaller than two. However, this only held for large enough effect size. For small effect size, visual tests with fewer evaluations might have greater power. Note that, the expected power of visual test derived by **?** followed by the assumption that human has the ability to select the plot with the highest t-statistic from a lineup under the classical linear model regression setting showed similar properties, where visual tests with fewer evaluations were expected to perform better when the parameter values close to the null hypothesis.

For cubic model, the separation between curves was small. Fitted power of visual test 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 when the effect size is large enough.

As shown in Figure @ref(fig:power-com), both F-test and Breusch–Pagan test generally possessed greater power than visual test. It was also found that there was a huge gap between the conventional test curves and the visual test curves at around log(effect size) $= 0$ for the cubic model and log(effect size) $= 2.5$ for the heteroskedasticity model.

We further analysed the lineup with the corresponding effect size.

Table 1. Logistic regressions with rejection as response variable and natural logarithm of the effect size as regressor fitted on the data collected from experiment I and II for cubic model. Five regression were fitted for different number of evaluations.

	Dependent variable:				
			reject		
	num_eval: 1	num_eval:2	num_eval: 3	num_eval: 4	num_eval:5
log_effect_size	0.929***	1.258***	1.096***	1.113***	1.244***
	(0.073)	(0.076)	(0.054)	(0.056)	(0.077)
Constant	−1.969***	−2.534***	−1.549***	−1.334***	−1.404***
	(0.169)	(0.173)	(0.113)	(0.122)	(0.185)
Observations	630	1,135	1,660	1,904	1,879
Log Likelihood	−220.882	−314.394	−471.152	−396.979	−206.296
Akaike Inf. Crit.	445.764	632.788	946.305	797.958	416.592

Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2. Logistic regressions with rejection as response variable and natural logarithm of the effect size as regressor fitted on the data collected from experiment I and II for heteroskedasticity model. Five regression were fitted for different number of evaluations.

			Dependent variable:		
			reject		
	num_eval: 1	num_eval:2	num_eval: 3	num_eval: 4	num_eval:5
log_effect_size	0.821***	1.104***	1.157***	1.387***	1.651***
	(0.086)	(0.079)	(0.066)	(0.076)	(0.110)
Constant	−4.472***	−5.887***	−5.431***	−6.154***	−7.123***
	(0.434)	(0.415)	(0.344)	(0.400)	(0.591)
Observations	630	1,138	1,596	1,662	1,416
Log Likelihood	−333.119	−631.519	−835.317	−650.570	−342.058
Akaike Inf. Crit.	670.238	1,267.038	1,674.634	1,305.140	688.116

Note: *p<0.1; **p<0.05; ***p<0.01

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.
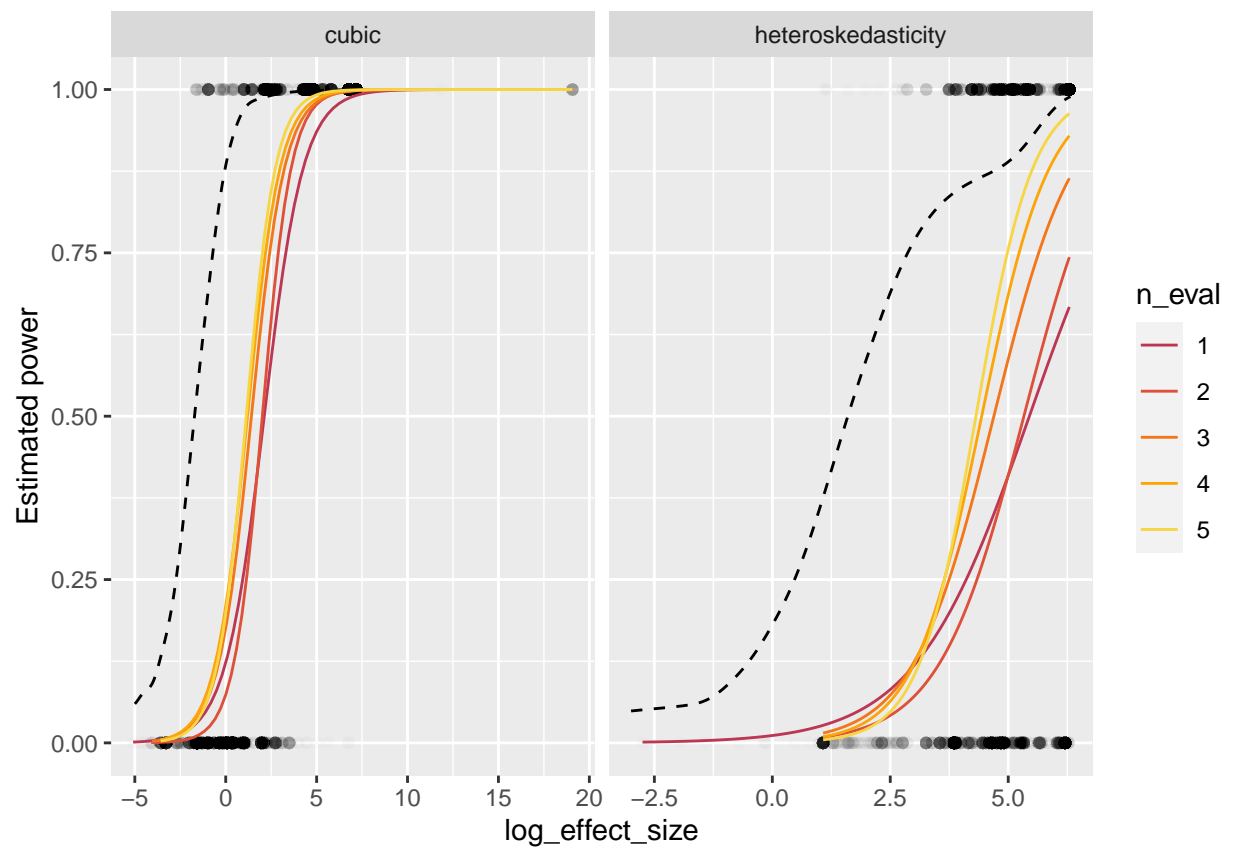
Figure 1. power com