# Supplmentary material - A plot is worth a thousand tests: assessing residual diagnostics with visual inference

Weihao Li, Dianne Cook, Emi Tanaka and Susan VanderPlas

## Appendix

### Effect size derivation

Effect size can be defined as the difference of a parameter for a particular model or distribution, or a statistic derived from a sample. Importantly, it needs to reflect the treatment we try to measure. Centred on a conventional statistical test, we usually can deduce the effect size from the test statistic by substituting the null parameter value. When considering the diagnostics of residual departures, there exist many possibilities of test statistics for a variety of model assumptions. Meanwhile, diagnostic plots such as the residual plot have no general agreement on measuring how strong a model violation pattern is. To build a bridge between various residual-based tests, and the visual test, we focus on the shared information embedded in the testing procedures, which is the distribution of residuals. When comes to comparison of distribution, Kullback-Leibler divergence (@ Kullback and Leibler 1951) is a classical way to represent the information loss or entropy increase caused by the approximation to the true distribution, which in our case, the inefficiency due to the use of false model assumptions.

Following the terminology introduced by Kullback and Leibler (1951), $P$ represents the measured probability distribution, and $Q$ represents the assumed probability distribution. The Kullback-Leibler divergence is defined as $\int_{-\infty}^{\infty} log(p(x)/q(x))p(x)dx$, where $p(.)$ and $q(.)$ denote probability densities of $P$ and $Q$.

Let $\boldsymbol{X}_a = (\boldsymbol{1}, \boldsymbol{X})$ denotes the $p$ regressors with $n$ observations, $\boldsymbol{R}_a = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ denotes the residual operator, and let $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ denotes the error. Using the Frisch–Waugh–Lovell theorem, residuals $\boldsymbol{e} = \boldsymbol{R}_a\boldsymbol{\varepsilon}$. Because $rank(\boldsymbol{R}_a) = n - p < n$, $e$ follows a degenerate multivariate normal distribution and does not have a density. Since the Kullback-Leibler divergence requires a proper density function, we need to simplify the covariance matrix of $\boldsymbol{e}$ by setting all the off-diagonal elements to 0. Then, the residuals will assumed to follow $N(\boldsymbol{0}, diag(\boldsymbol{R}_a\sigma^2))$ under the null hypothesis that the model is correctly specified. If the model is however misspecified due to omitted variables $\boldsymbol{Z}$, or a non-constant variance $\boldsymbol{V}$, the distribution of residuals can be derived as $N(\boldsymbol{R}_a\boldsymbol{Z}\boldsymbol{\beta}_z, diag(\boldsymbol{R}_a\sigma^2))$ and $N(\boldsymbol{0}, diag(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a'))$ respectively.

By assuming both $P$ and $Q$ are multivariate normal density functions, the Kullback-Leibler divergence can be rewritten as

$$KL = \frac{1}{2}\left(log\frac{|\Sigma_p|}{|\Sigma_q|} - n + tr(\Sigma_p^{-1}\Sigma_q) + (\mu_p - \mu_q)'\Sigma_p^{-1}(\mu_p - \mu_q)\right).$$

Then, we can combine the two residual departures into one formula

$$KL = \frac{1}{2}\left(log\frac{|diag(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a')|}{|diag(\boldsymbol{R}_a\sigma^2)|} - n + tr(diag(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a')^{-1}diag(\boldsymbol{R}_a\sigma^2)) + \boldsymbol{\mu}_z^T(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a')^{-1}\boldsymbol{\mu}_z\right). \qquad (1)$$

When there are omitted variables but constant error variance, the formula can be reduced to

$$KL = \frac{1}{2}\left(\boldsymbol{\mu}_z^T(diag(\boldsymbol{R}_a\sigma^2))^{-1}\boldsymbol{\mu}_z\right).$$

And when the model equation is correctly specified but the error variance is non-constant, the formula can be reduced to

$$KL = \frac{1}{2}\left(log\frac{|diag(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a')|}{|diag(\boldsymbol{R}_a\sigma^2)|} - n + tr(diag(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a')^{-1}diag(\boldsymbol{R}_a\sigma^2))\right).$$

Since we assume $\sigma = 1$ for the heteroskedasticity model, the final form of the formula is

$$KL = \frac{1}{2}\left(log\frac{|diag(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a')|}{|diag(\boldsymbol{R}_a)|} - n + tr(diag(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a')^{-1}diag(\boldsymbol{R}_a))\right).$$

## Experiment setup

### Mapping of subjects to experimental factors

Mapping of subjects to experimental factors is an important part of experiment design. Essentially, we want to maximum the difference in factors exposed to a subject. For this purpose, we design an algorithm to conduct subject allocation. Let $L$ be a set of available lineups and $S$ be a set of available subjects. According to the experimental design, the availability of a lineup is associated with the number of subjects it can assign to. For lineups with uniform fitted value distribution, this value is 11. And other lineups can be allocated to at most five different subjects. The availability of a subject is associated with the number of lineups that being allocated to this subject. A subject can view at most 18 different lineups.

The algorithm starts from picking a random subject $s \in S$ with the minimum number of allocated lineups. It then tries to find a lineup $l \in L$ that can maximise the distance metric $D$ and allocate it to subject $s$. Set $L$ and $S$ will be updated and the picking process will be repeated until there is no available lineups or subjects.

Let $F_1, ..., F_q$ be $q$ experimental factors, and $f_1, ..., f_q$ be the corresponding factor values. We say $f_i$ exists in $L_s$ if any lineup in $L_s$ has this factor value. Similarly, $f_if_j$ exists in $L_s$ if any lineup in $L_s$ has this pair of factor values. And $f_if_jf_k$ exists in $L_s$ if any lineup in $L_s$ has this trio of factor values. The distance metric $D$ is defined between a lineup $l$ and a set of lineups $L_s$ allocated to a subject $s$ if $L_s$ is non-empty:

$$D = C - \sum_{1\leq i\leq q} I(f_i \text{ exists in } L_s) - \sum_{\substack{1\leq i\leq q-1\\ i<j\leq q}} I(f_if_j \text{ exists in } L_s) - \sum_{\substack{1\leq i\leq q-2\\ i<j\leq q-1\\ j<k\leq q}} I(f_if_jf_k \text{ exists in } L_s)$$

where $C$ is a sufficiently large constant such that $D > 0$. If $L_s$ is empty, we define $D = 0$.

The distance measures how different a lineup is from the set of lineups allocated to the subject in terms of factor values. Thus, the algorithm will try to allocate the most different lineup to a subject at each step.

### Demographics

Along with the responses to lineups, we have collected a series of demographic information including age, pronoun, education background and previous experience in studies involved data visualization. Figure 1 and 2 provide summary of the demographic data.

As shown in Figure 1, most participants have Diploma or Bachelor degrees, followed by High school or below. The survey data is gender balanced and the majority of participants are between 18 to 39 years old.

Figure 2 shows that the number of participants who have previous experience is not very different from the number of those who haven't. Age distributions are also similar for these two groups.

### Data collection interface

## Batch effect size

We have the same type of model collected over different data collection periods, that may lead to unexpected batch effect. Figure 3 shows the weighed proportion of detect over different data collection period. The

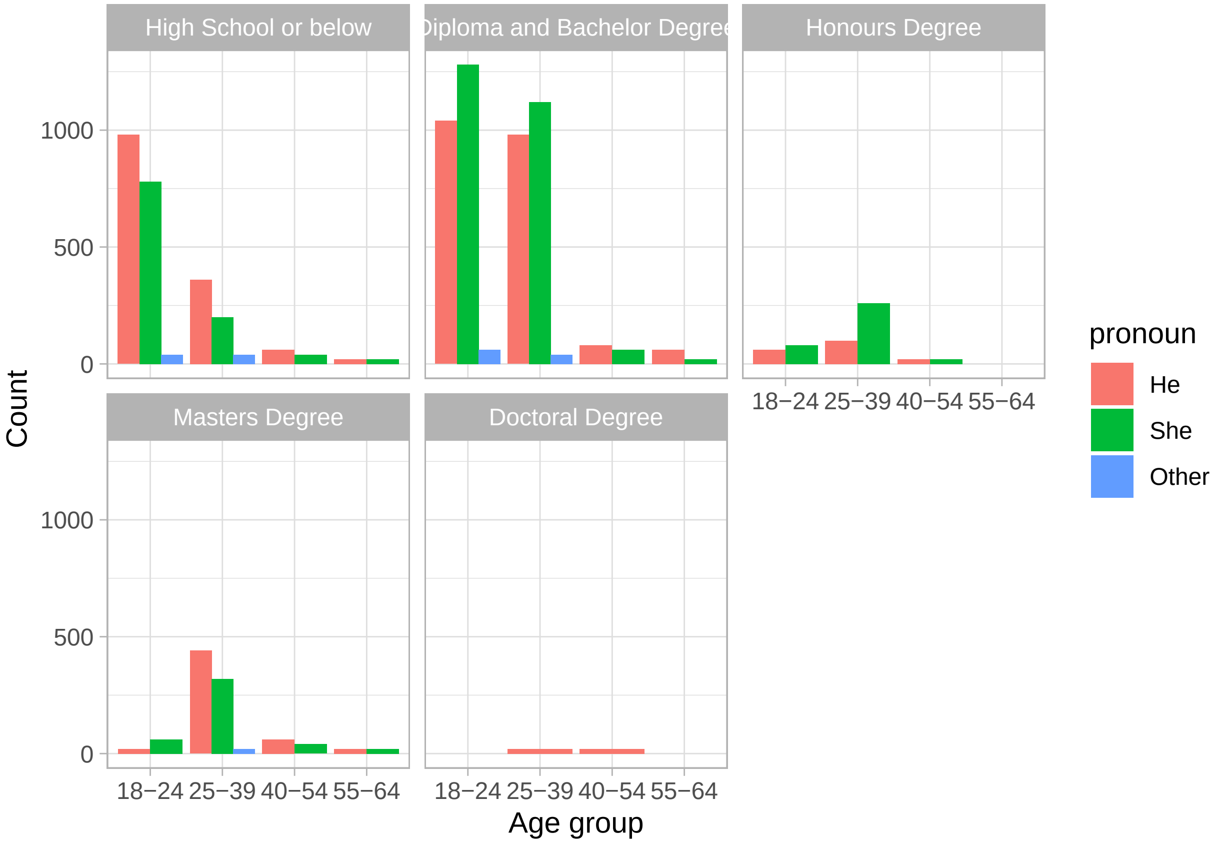

Figure 1: Education distribution of subjects recuritted in this study group by age and pronoun.

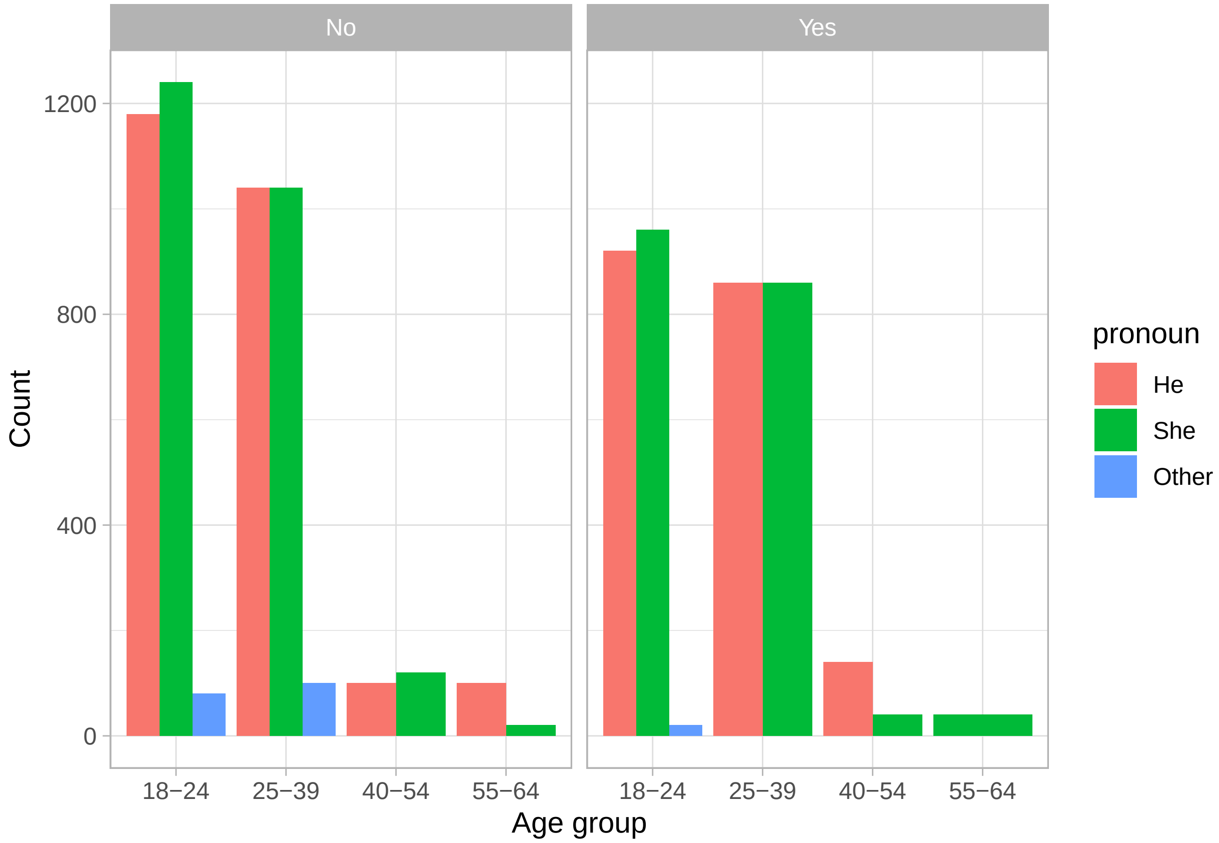

Figure 2: Status of pervious experience of subjects recuritted in this study group by age and pronoun.

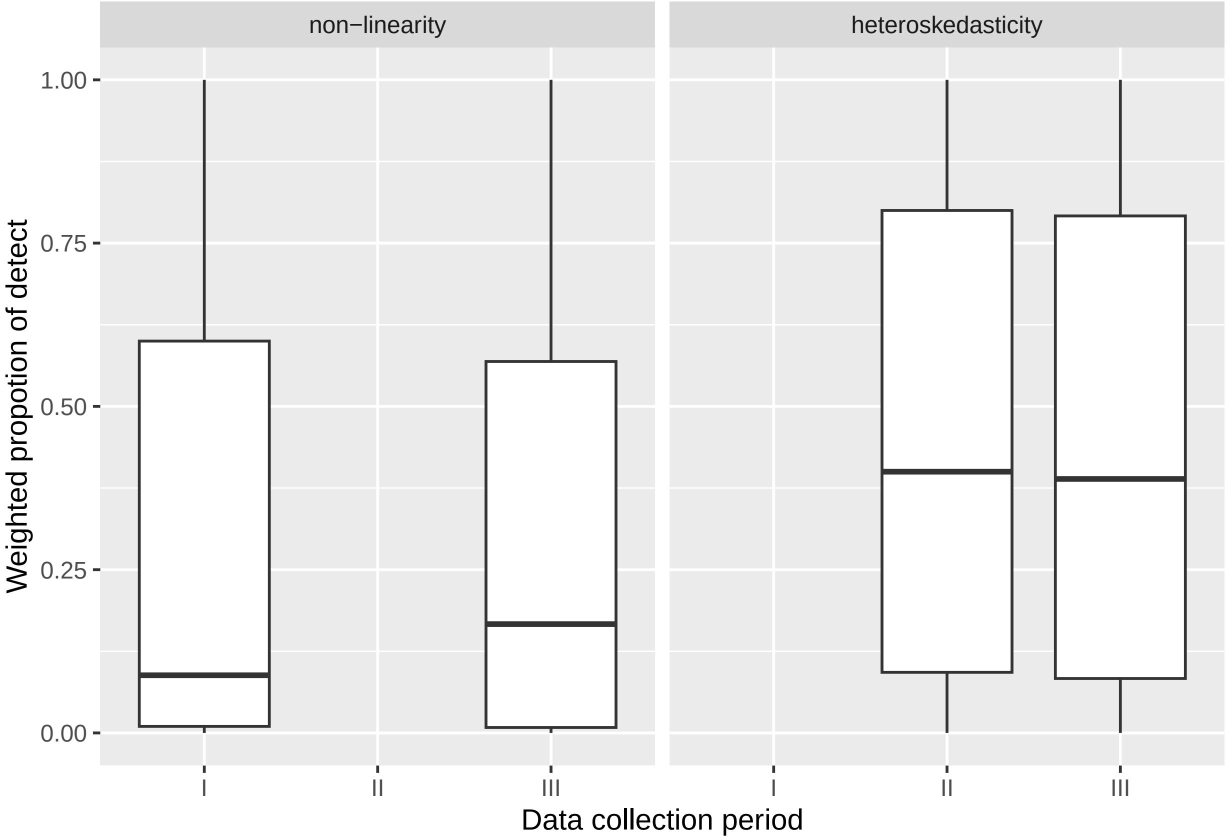

Figure 3: Boxplot of weighted propotion of detect for lineups over different data collection periods.

weighted proportion of detect is calculated by taking the average of $c_i$ of a lineup over a data collection period. According to the graph, the distribution for the heteroskedasticity patterns are almost identical in period II and III, while the distribution for the non-linearity patterns have a shift in the median. However, this difference is not statistically significance. The $p$-values of the t-test for testing the mean difference across data periods are 0.46 and 0.98 for non-linearity and heteroskedasticity patterns respectively. Both are much greater than 5%. Therefore, there is no clear evidence of batch effect.

(use LV plot `lvplot` `geom_lv`)

## Sensitivity analysis for $\alpha$

The parameter $\alpha$ used for the $p$-value calculation needs to be estimated from responses to null lineups. However, The way we generate Rorschach lineup is not strictly the same as what suggested in VanderPlas et al. (2021) and Buja et al. (2009). Therefore, we conduct a sensitivity analysis in this section to examine the impact of the variation of the estimator $\alpha$ on our primary findings.

The analysis is conducted by setting up two relatively extreme scenarios, where the $\alpha$ is under or overestimated by 25%. Using the inflated and deflated $\hat{\alpha}$, we recalculate the $p$-value for every lineup and show the results in Figure 4. It can be observed that $p$-value changes very little. It is very unlikely the findings will be affected because of the estimate of $\alpha$.

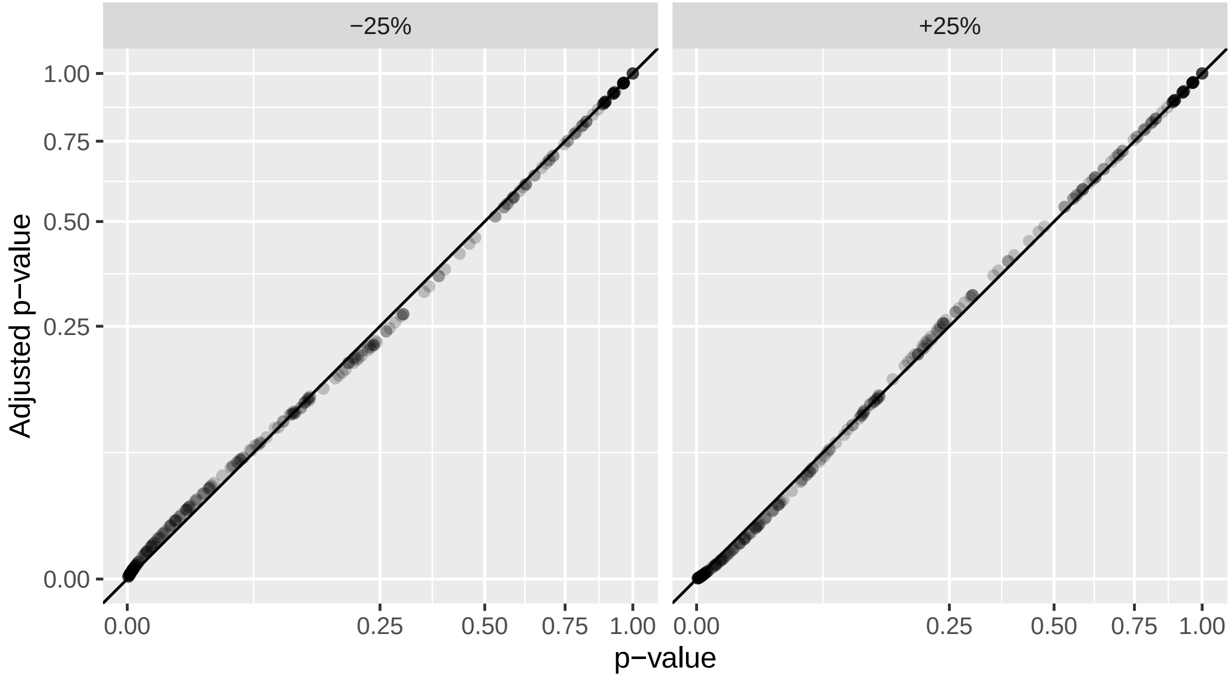

Figure 4: Change of $p$-values with $\hat{\alpha}$ inflated by 25% and deflated by 25%.

## Curious effect of predictor distribution on conventional test power

## Other results

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): 4361–83. https://doi.org/10.1098/rsta.2009.0120.

Kullback, Solomon, and Richard A Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22 (1): 79–86.

VanderPlas, Susan, Christian Röttger, Dianne Cook, and Heike Hofmann. 2021. "Statistical Significance Calculations for Scenarios in Visual Inference." *Stat* 10 (1): e337.