appendix

2023-01-22

Appendix

Effect size

Derivation

Effect size is usually defined as the difference of a parameter or a statistic derived from a sample. 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 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 @kullback1951information, 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

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

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

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

Caculation

The effect size can be calculated using Equation 1 for each lineup by plugging in the data generated from the simulated model. However, since the term $\boldsymbol{R}_a$ will differ for each sample, the effect size will be different even if the parameter values are the same. To overcome this undesired property, for the same set of parameter values, we simulate a sufficient large number of samples, and take the average of the effect size as the final effect size.

Experiment setup

Mapping of subjects to experimental factors

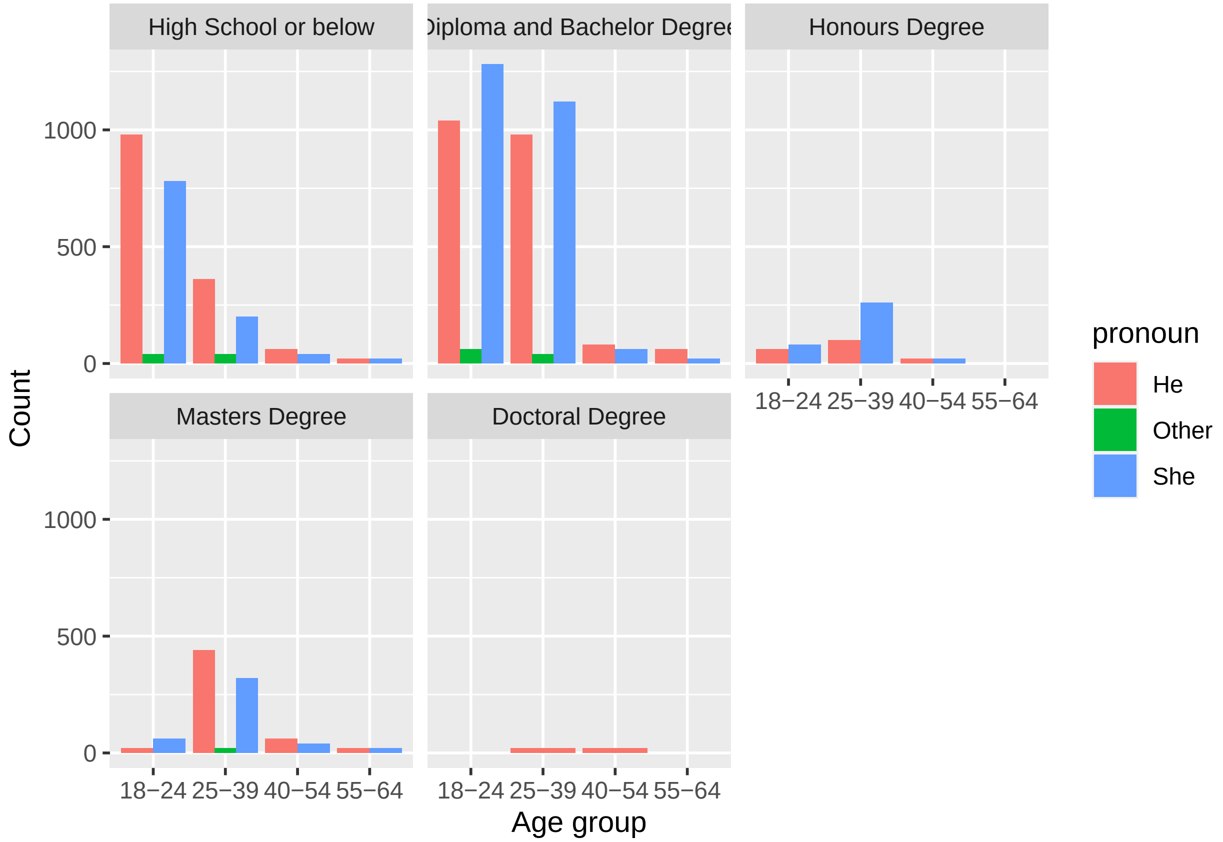
Mapping of subjects to experimental factors is an important part of experiment design. Essentially, we want to maximum the difference in factors exposed to a subject. For this purpose, we design an algorithm to conduct subject allocation. Let L be a set of available lineups and S be a set of available subjects. According to the experimental design, available means a lineup can be allocated to at most five different subjects (for lineups with uniform fitted value distribution, this value is 11) and a subject can view at most 18 different lineups.

The algorithm starts from picking a random subject s from S with the minimum number of allocated lineups. It then computes a metric for every lineup l in L indicating the distance D between lineup l to the lineups already allocated to this subject $l_{allocated}$. Lineup with the highest distance D will be allocated to the subject. The algorithm will repeatedly pick a subject, compute metrics, and allocate lineups until there is no available lineups or no available subjects.

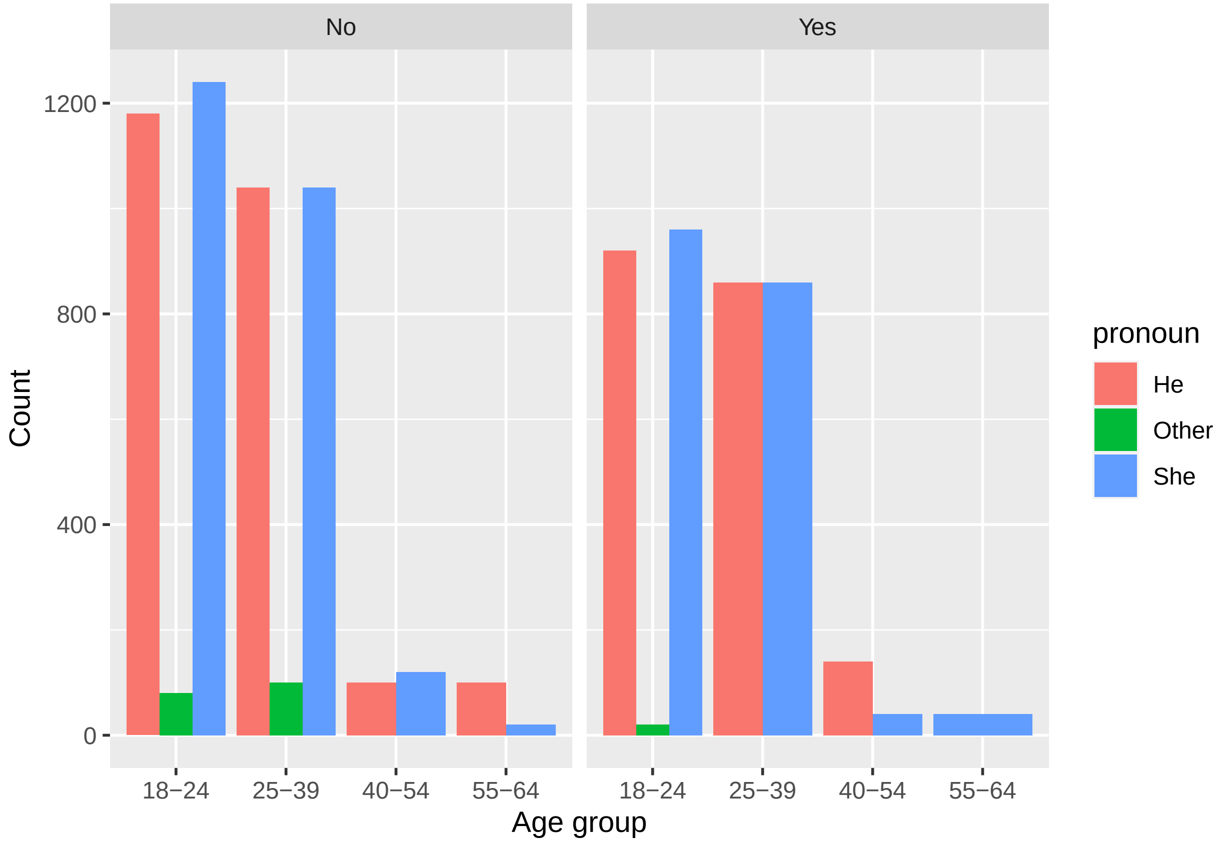
The distance D is defined using the factor values used to generate lineup l and lineups $l_{allocated}$. If there does not exist $l_{allocated}$, then the distance $D = 0$. Otherwise, distance D is roughly the number of factors minus the number of factor values of lineup l exists in a union of factor values of lineups $l_{allocated}$. The distance D is further adjusted by the existence of n-factor combinations, where n could be two to the number of factors. In this way, all factors values are exposed to every subject, and some even ensure all 2-factor combinations are exposed to every subject.

Demographics

The education background of most participants have Diploma or Bachelor degrees, followed by High school or below. The survey data is gender balanced. Most of the participants are between 18 to 39.



Number of participants who have previous experience in experiment about reading data plots are not very different from the number of participants 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 **??** shows the weighed proportion of detect over different data collection period. The 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.

Sensitivity analysis for α

The parameter α 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 @vanderplas2021statistical and @buja_statistical_2009. Therefore, we conduct a sensitivity analysis in this section to examine the impact of the variation of the estimator α on our primary findings.

The analysis is conducted by setting up two relatively extreme scenarios, where the α 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 **??**. It can be observed that p-value changes very little. It is very unlikely the findings will be affected because of the estimate of α.

Other results

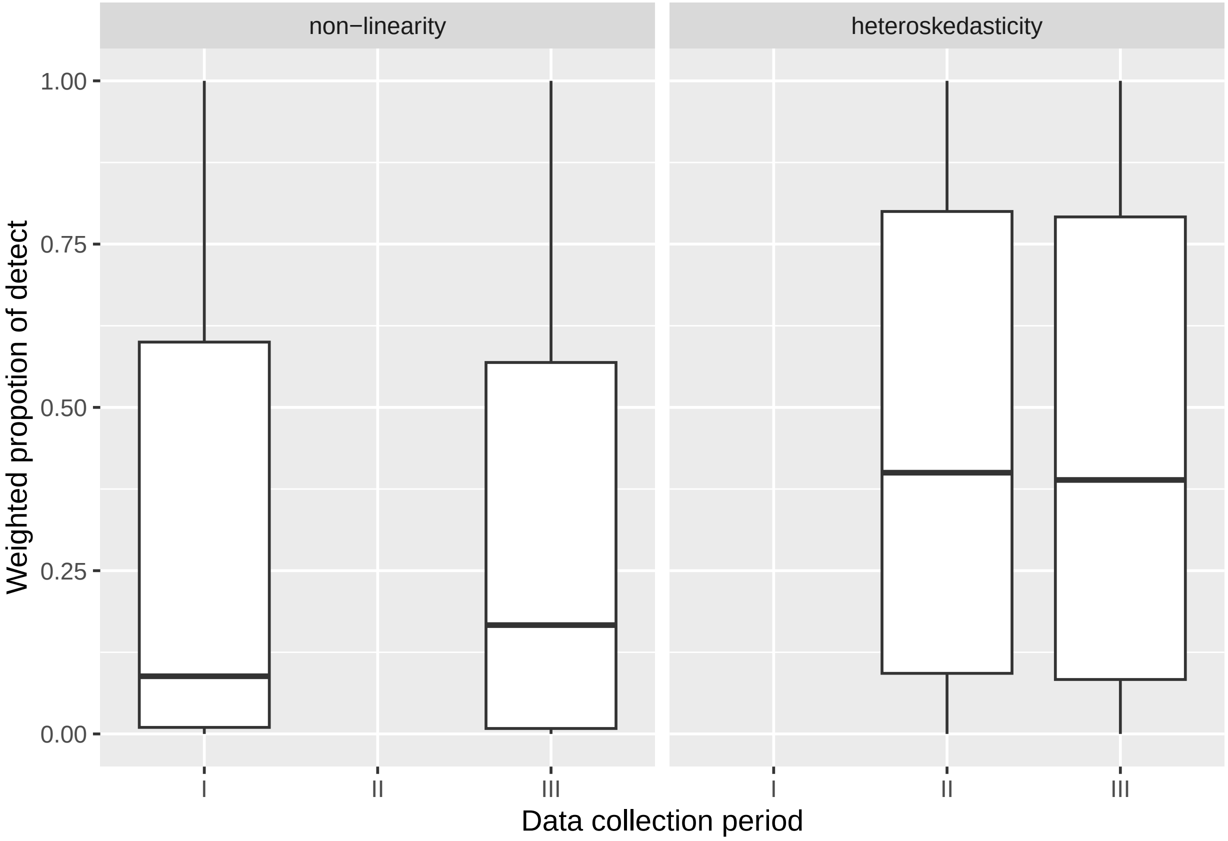


Figure 1: Boxplot of weighted propotion of detect for lineups over different data collection periods.

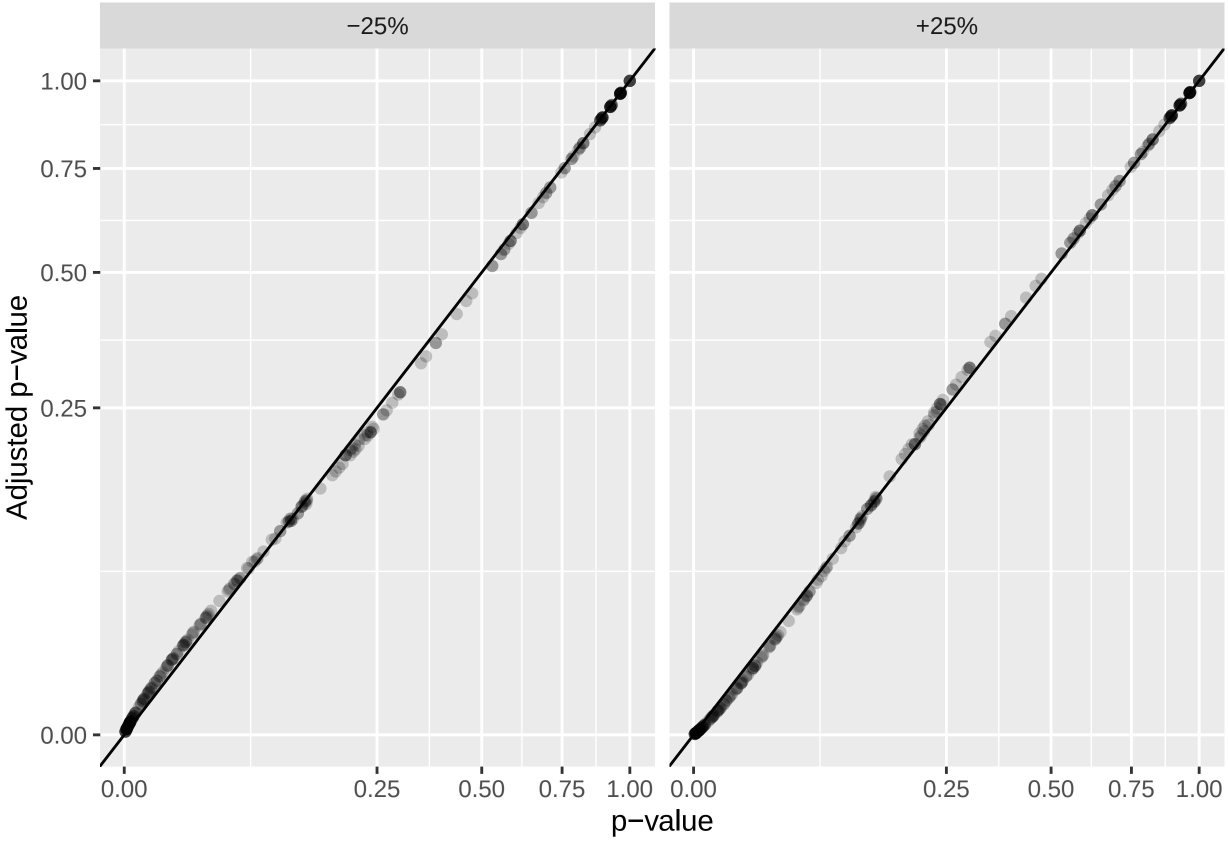


Figure 2: Change of p-values with $\hat{\alpha}$ inflated by 25% and deflated by 25%.