appendix

2023-01-22

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 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

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

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

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

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

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

p-value scatter plot

A collection interesting lineups (unusual results)

why unusual? what is the possible explanations?

targe journal

JRSSB: Journal of the Royal Statistical Society Series B (Statistical Methodology) Deadline: Jan 1, Apr 1

Reading: style of writing, author guideline (https://rss.onlinelibrary.wiley.com/hub/journal/14679868/author-guidelines)

JCGS JDSS