[SUPPLEMENTARY MATERIALS]

Stochastic Variance Inflation Factor with Collective Information Content Pre-analysis for Detecting Linkage Disequilibrium in Indonesian Rice SNPs

Nicholas Dominic[1[0000-0003-2015-6689]] and Bens Pardamean[1,2[0000-0002-7404-9005]]

[1] Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia, 11480

[2] BINUS Graduate Program, Bina Nusantara University, Jakarta, 11480, Indonesia

bdsrc@binus.edu

Definition 1. By deriving from Eq. (2) and minimizing the residual sum of squares (RSS) or the cost function $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, the normal Ordinary Least Squares (OLS) can be formulated as

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

where β_0 and β_1 are the coefficients, act as the regression parameters.

Proof 1. The sum of all residuals is equal to zero, $\sum_{i=1}^{N}\varepsilon_i = 0$.

1. Estimate value of β_0 that minimizes the OLS:

$$\frac{\partial}{\partial \beta_0}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

$$-2\sum_{i=1}^{N} y_i - \beta_0 - \beta_1 x_i = 0$$

   Ignore the constant for a while. Since $\varepsilon_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i)$, it is proved that $\sum_{i=1}^{N}\varepsilon_i = 0$. Remember, since a partial derivative has been done w.r.t. β_0, this property applies when $\beta_0 = 0$.

2. Estimate value of β_1 that minimizes the OLS:

$$\frac{\partial}{\partial \beta_1}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

$$-2\sum_{i=1}^{N} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

Omit the constant and thus it is proved that $\sum_{i=1}^{N} x_i \varepsilon_i = 0$. Remember, since a partial derivative has been done w.r.t. β_1, this property applies when $\beta_{1,\dots,N} = 0$.

Proof 2. Residual and the predicted value are uncorrelated, $corr(\hat{y}, \varepsilon) = 0$.

$$
\begin{aligned}
Corr(\hat{y}, \varepsilon) &= \frac{Cov(\hat{y}, \varepsilon)}{\sqrt{\sigma_{\hat{y}}^2 \sigma_{\varepsilon}^2}} \\
&= \frac{N^{-1} \sum_{i=1}^{N} (\hat{y} - \mu_{\hat{y}})(\varepsilon - \mu_{\varepsilon})}{\sqrt{\sigma_{\hat{y}}^2 \sigma_{\varepsilon}^2}} \\
&= \frac{N^{-1} \sum_{i=1}^{N} \varepsilon (\hat{y} - \mu_{\hat{y}})}{\sqrt{\sigma_{\hat{y}}^2 \sigma_{\varepsilon}^2}} && \dots\ \mu_{\varepsilon} = 0 \\
&= \frac{N^{-1} \left(\sum_{i=1}^{N} \hat{y}\varepsilon - \mu_{\hat{y}} \sum_{i=1}^{N} \varepsilon\right)}{\sqrt{\sigma_{\hat{y}}^2 \sigma_{\varepsilon}^2}} && \dots\ \textstyle\sum_{i=1}^{N} \varepsilon = 0 \text{ and } \sum_{i=1}^{N} \hat{y}\varepsilon = 0, \text{ has been proved in } \textbf{\textit{Proof 1}}. \\
&= 0
\end{aligned}
$$

Hence, it is proved that $Corr(\hat{y}, \varepsilon) = 0$, as well as $Cov(\hat{y}, \varepsilon) = 0$.

Proof 3. The coefficient of determination is equivalent to the squared Pearson correlation coefficient, $R^2(y, \hat{y}) \equiv r_{y\hat{y}}^2$.

$$
\begin{aligned}
r^2(y, \hat{y}) &= \left(\frac{Cov(y, \hat{y})}{\sqrt{\sigma_y^2 \sigma_{\hat{y}}^2}}\right)^2 \\
&= \frac{Cov(y, \hat{y})\ Cov(y, \hat{y})}{\sigma_y^2 \sigma_{\hat{y}}^2}
\end{aligned}
$$

Recall that residual, $\varepsilon = y - \hat{y} \Leftrightarrow y = \hat{y} + \varepsilon$.

$$
= \frac{Cov(\hat{y} + \varepsilon, \hat{y})\ Cov(\hat{y} + \varepsilon, \hat{y})}{\sigma_y^2 \sigma_{\hat{y}}^2}
$$

This equation can be expanded since $Cov\big(a, (b + c)\big) = Cov(a, b) + Cov(a, c)$, and then can be further simplified since $Cov(a, a) = \sigma_a^2$. The cancellation of $Cov(\hat{y}, \varepsilon)$ is due to the second property of residuals, as proved in *Proof 2*.

$$
= \frac{\big(Cov(\hat{y}, \hat{y}) + Cov(\hat{y}, \varepsilon)\big)\big(Cov(\hat{y}, \hat{y}) + Cov(\hat{y}, \varepsilon)\big)}{\sigma_y^2 \sigma_{\hat{y}}^2}
$$

$$= \frac{Cov(\hat{y},\hat{y})\ Cov(\hat{y},\hat{y})}{\sigma_y^2 \sigma_{\hat{y}}^2}$$

$$= \frac{\sigma_{\hat{y}}^2 \sigma_{\hat{y}}^2}{\sigma_y^2 \sigma_{\hat{y}}^2}$$

$$r^2(y,\hat{y}) = \frac{N^{-1}\sum_{i=1}^{N}\left(\hat{y}_i - \mu_{\hat{y}_i}\right)^2}{N^{-1}\sum_{i=1}^{N}\left(y_i - \mu_{y_i}\right)^2}$$

where explained sum of squares, $ESS = N^{-1}\sum_{i=1}^{N}\left(\hat{y}_i - \mu_{\hat{y}_i}\right)^2$, total sum of squares, $TSS = N^{-1}\sum_{i=1}^{N}\left(y_i - \mu_{y_i}\right)^2$, and hence $R_{y\hat{y}}^2 = \frac{ESS}{TSS} = r^2(y,\hat{y})$.