Notes on R-SQUARED AND GOODNESS OF FIT

Zhou Qiao

2016-04-13

R-SQUARED AND GOODNESS OF FIT

ANOVA Decomposition

Given the simple regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

We'll decompose its total variance into two components using **ANOVA Decomposition** (ANOVA stands for Analysis of Variance).

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{e}_i^2$$

Or more succinctly as

$$SS(Total) = SS(Regression) + SS(Residual)$$

Proof of ANOVA Decomposition:

We first prove a relationship that will be useful later on:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$$

To prove this, we rewrite the LHS:

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=\sum_{i=1}^{n}e_i\left(\hat{\alpha}+\hat{\beta}x_i-\bar{y}\right)=\sum_{i=1}^{n}e_i\left(\hat{\alpha}+\hat{\beta}x_i-\hat{\alpha}-\hat{\beta}\bar{x}\right)=\hat{\beta}\sum_{i=1}^{n}e_i(x_i-\bar{x})$$
$$=\hat{\beta}\sum_{i=1}^{n}e_ix_i-\hat{\beta}\bar{x}\sum_{i=1}^{n}e_i=0$$

Then,

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2$$
$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$
$$=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}\hat{e}_i^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$
$$=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}\hat{e}_i^2$$

Therefore, the total variations in the data can be decomposed as the sum of the variations of the regression and the variations in the residuals.

R-SQUARED

R-squared is known as the coefficient of determination. It is closely linked to the ANOVA Decomposition and is defined as

$$r^2=\frac{SS(Regression)}{SS(Total)}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}\hat{e}_i^{\,2}}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sigma_e^2}{\sigma_y^2}$$

The value $1-r^2$ is known as the unexplained variation.

The adjusted R^2 is defined as

$$\bar{R}^2=R^2-(1-R^2)\frac{k}{n-k-1}$$

Because of the ANOVA decomposition, the R-squared is guaranteed to be smaller than 1. Adding regressors into a model always increases R-squared and adjusted R-squared is adjusted for this effect.

R-squared can also be interpreted as the square of the sample Pearson correlation between the observed dependent variable (y_i) and the predicted values ($\hat{y}_i$) (proof can be found in the **Appendix A** section). By this argument, high R^2 is often (but not always) associated with high predictive power for a regression model.

Standard Deviation of the Residuals

The standard deviation of the residuals can be computed as

$$\hat{\sigma}_e = \sqrt{\frac{\sum_{i=1}^{n} \hat{e_i}^2}{n-k}}$$

Where k is the number of parameters. It is a goodness of fit measure, and it is in the same unit as y_i. About 95% of the data will line within $2\hat{\sigma}_e$ of the regression line. For future prediction, the 95% confidence interval can be constructed by

$$\bar{y} \pm 2\hat{\sigma}_e$$

Appendix

A. Proof: R-Squared as Correlation Squared

R-squared can be interpreted as the square of the sample Pearson correlation between the observed dependent variable (y_i) and the predicted values ($\hat{y}_i$), $\rho_{y_i,\hat{y}_i}$. For the regression model $y_i = \beta' x_i + e_i = \hat{y}_i + e_i$, this conclusion requires the assumption that

$$E[\hat{y}_i e_i] = Cov(\hat{y}, e) = 0$$

Which will be satisfied if error **exogeneity** is assumed[1]. It then follows that

$$\begin{aligned}
\rho^2_{y_i,\hat{y}_i} &= \left(\frac{cov(y_i, \hat{y}_i)}{\sqrt{var(y_i)var(\hat{y}_i)}}\right)^2 \\
&= \frac{cov(y_i, \hat{y}_i)}{var(y_i)} \frac{cov(y_i, \hat{y}_i)}{var(\hat{y}_i)} \\
&= \frac{cov(\hat{y}_i + e_i, \hat{y}_i)}{var(y_i)} \frac{cov(\hat{y}_i + e_i, \hat{y}_i)}{var(\hat{y}_i)} \\
&= \frac{cov(\hat{y}_i, \hat{y}_i) + cov(e_i, \hat{y}_i)}{var(y_i)} \frac{cov(\hat{y}_i, \hat{y}_i) + cov(e_i, \hat{y}_i)}{var(\hat{y}_i)} \\
&= \frac{var(\hat{y}_i)}{var(y_i)} \frac{var(\hat{y}_i)}{var(\hat{y}_i)} = \frac{var(\hat{y}_i)}{var(y_i)} \\
&= \frac{SS(Regression)}{SS(Total)} \\
&= R^2
\end{aligned}$$

[1] Under the classical OLS assumptions, **exogeneity** ($i.e., E[e|X] = 0$) implies that $E[eX] = 0$, $E(x_j e_i) = 0\ \forall i, j$ and unbiasedness of the coefficient estimation: $E[\hat{\beta}] = \beta$.

Reference

1. Simon Jackman, ‘r-squared and goodness of fit’, Retrieved from http://jackman.stanford.edu/classes/ssmart/2011/rsquared.pdf