Confidence intervals for R^2

Andrey Konovalov

MIPT

April 14, 2014

Coefficient of Determination

Given a data set $\{y_i, x_{1i}, \ldots, x_{ki}\}_{i=1}^{n}$

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2 \quad \text{Total Sum of Squares}$$

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \quad \text{Explained Sum of Squares}$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{Residual Sum of Squares}$$

Coefficient of Determination:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Coefficient of Determination (cont'd)

- R^2 takes on values between 0 and 1.
- The higher the R^2, the more useful the model.
- Interpretation: R^2 tells us how much better we do by using the regression equation rather than just the mean $\bar{y}$ to predict y.
- Despite the interpretation the value of R^2 doesn't mean much by itself.
- The value of R^2 can be small, but your regression is perhaps still better than doing nothing.
- R^2 might be interesting in some rare cases, like comparing two models on the same dataset.

Pearson's correlation coefficient

For a population:

$$\rho_{x_1 x_2} = \frac{\mathrm{cov}(x_1, x_2)}{\sigma_1 \sigma_2} = \frac{E[(x_1 - \mu_1)(x_2 - \mu_2)]}{\sigma_1 \sigma_2}$$

$$\rho_{x_1(x_2 \ldots x_k)} = 1 - \frac{|\hat{R}|}{\hat{R}_{11}}, \quad \hat{R} = |\rho_{x_i x_j}|, \quad \hat{R}_{11} \text{ is cofactor of } \rho_{x_1 x_1}$$

For a sample:

$$r_{x_1 x_2} = \frac{\sum_{i=1}^{n} (x_{1i} - \overline{x_1})(x_{2i} - \overline{x_2})}{\sum_{i=1}^{n} (x_{1i} - \overline{x_1})^2 \sum_{i=1}^{n} (x_{2i} - \overline{x_2})^2}$$

$$r_{x_1(x_2 \ldots x_k)} = 1 - \frac{|\hat{R}|}{\hat{R}_{11}}, \quad \hat{R} = |r_{x_i x_j}|, \quad \hat{R}_{11} \text{ is cofactor of } r_{x_1 x_1}$$

Turns out $R^2 = r^2_{y(x_1 \ldots x_k)}$

Olkin and Finn models

- Model A: Determining whether an additional variable provides an improvement in predicting the criterion: $\rho^2_{0(12)} - \rho^2_{01}$. This comparison shows whether an additional variable x_2 provides an improvement over x_1 alone in predicting $y = x_0$.

- Model B: Deciding which of two variables adds more to the prediction of the criterion: $\rho^2_{0(12)} - \rho^2_{0(13)}$ This comparison shows whether the pair of predictors x_1, x_2 or the pair x_1, x_3 is more effective in predicting criterion $y = x_0$.

- Model E: Deciding if a given set of predictors performs equally well in two separate populations: $\rho^2_{I} - \rho^2_{II}$. This comparison shows whether a given set of predictors $(x_1, x_2, ..., x_k)$ performs equally well in two independent samples of data.

General procedure form

- r_A and r_B - two sample correlations to be compared
- ρ_A and ρ_B - their correspondng population values

The large sample distribution for the comparison:

$$[(r_A - r_B) - (\rho_A - \rho_B)] \sim N(0, \sigma^2_\infty)$$
$$\sigma^2_\infty = \mathrm{var}(r_A) + \mathrm{var}(r_B) - 2\mathrm{cov}(r_A, r_B)$$

A $100(1-\alpha)\%$ confidence interval:

$$r_A - r_B \pm c\hat{\sigma}_\infty$$

where

- c is the standard normal deviate $z_{\alpha/2}$
- $\hat{\sigma}_\infty$ is an estimate of σ_∞ in which sample values replace population values

General procedure form (cont'd)

The general form of variance of function of a set of correlations:

$$\mathrm{var}_\infty f(r_{12}, r_{13}, r_{23}) = \mathbf{a\Phi a}'$$

$$\boldsymbol{a} = \left(\frac{\partial f}{\partial r_{12}}, \frac{\partial f}{\partial r_{13}}, \frac{\partial f}{\partial r_{23}}\right)$$

$$\boldsymbol{\Phi} = \begin{pmatrix} \mathrm{var}(r_{12}) & \mathrm{cov}(r_{12}, r_{13}) & \mathrm{cov}(r_{12}, r_{23}) \\ & \mathrm{var}(r_{13}) & \mathrm{cov}(r_{13}, r_{23}) \\ & & \mathrm{var}(r_{23}) \end{pmatrix}$$

The variances and covariances of correlations:

$$\begin{aligned} \mathrm{var}(r_{ij}) &= ((1-\rho_{ij}^2)^2)/n \\ \mathrm{cov}(r_{ij}, r_{jk}) &= ((2\rho_{jk} - \rho_{ij}\rho_{ik})(1-\rho_{ij}^2-\rho_{ik}^2-\rho_{jk}^2)/2 + \rho_{jk}^3)/n \\ \mathrm{cov}(r_{ij}, r_{kl}) &= [\rho_{ij}\rho_{kl}(\rho_{ik}^2+\rho_{il}^2+\rho_{jk}^2+\rho_{jl}^2)/2 + \rho_{ik}\rho_{jl} + \rho_{il}\rho_{jk} \\ &\quad -(\rho_{ij}\rho_{ik}\rho_{il} + \rho_{ji}\rho_{jk}\rho_{jl} + \rho_{ki}\rho_{kj}\rho_{kl} + \rho_{li}\rho_{lj}\rho_{lk})]/n \end{aligned}$$

Data: A Study of Teenage Use of Abusable Substances

- The data were collected as part of a study of the use of alcohol, cigarettes, and marijuana among urban school children.
- An abusable substance score (USE, ranging from 0 to 3) was created by summing the number of substances (cigarettes, alcohol, or marijuana) that the individual had tried.
- Perceived friends' use (FRIENDS, ranging from 0 to 12) was assessed by questions that asked students to indicate the number of friends, who were using alcohol, cigarettes, or marijuana.
- Perceived family use (FAMILY) is the number of abusable substances, out of three, that were used by any member of the student's family.

Model A illustration

- Determining whether an additional variable provides an improvement in predicting the criterion.

- The variables are $x_0 = \text{USE}$, $x_1 = \text{FRIENDS}$, $x_2 = \text{FAMILY}$.

- The procedure compares $\rho^2_{0(12)}$ with ρ^2_{01} using estimates $r^2_{0(12)}$, r^2_{01} and $\hat{\sigma}^2_\infty = \mathrm{var}(r^2_{0(12)} - r^2_{01})$

Model A illustration (cont'd)

Following the procedure,

$$\mathrm{var}(f(r_{01}, r_{02}, r_{12})) = \mathrm{var}(r_{0(12)}^2 - r_{01}^2) = \mathbf{a}\boldsymbol{\Phi}\mathbf{a}'$$

$$\mathbf{a} = \left(\frac{\partial f}{\partial r_{01}}, \frac{\partial f}{\partial r_{02}}, \frac{\partial f}{\partial r_{12}} \right) = (a_1, a_2, a_3)$$

$$a_1 = \frac{2\rho_{12}}{1-\rho_{12}^2}(\rho_{01}\rho_{12} - \rho_{02}), \quad a_2 = \frac{2}{1-\rho_{12}^2}(\rho_{02} - \rho_{01}\rho_{12})$$

$$a_3 = \frac{2}{(1-\rho_{12}^2)^2}(\rho_{12}\rho_{01}^2 + \rho_{12}\rho_{02}^2 - \rho_{01}\rho_{02} - \rho_{01}\rho_{02}\rho_{12}^2)$$

$$\boldsymbol{\Phi} = \begin{pmatrix} \mathrm{var}(r_{01}) & \mathrm{cov}(r_{01}, r_{02}) & \mathrm{cov}(r_{01}, r_{12}) \\ & \mathrm{var}(r_{02}) & \mathrm{cov}(r_{02}, r_{12}) \\ & & \mathrm{var}(r_{12}) \end{pmatrix}$$

Model A illustration (cont'd)

The sample correlation matrix (obtained from the data):

$$R = \begin{pmatrix} r_{00} & r_{01} & r_{02} \\ & r_{11} & r_{12} \\ & & r_{22} \end{pmatrix} = \begin{pmatrix} 1.000 & 0.433 & 0.199 \\ & 1.000 & 0.178 \\ & & 1.000 \end{pmatrix}$$

The estimate of ρ_{01}^2 is $r_{01}^2 = 0.188$. The estimate of $\rho_{0(12)}^2$ is

$$r_{0(12)}^2 = \frac{r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12}}{1 - r_{12}^2} = 0.203$$

The difference is $r_{0(12)}^2 - r_{01}^2 = 0.015$

Model A illustration (cont'd)

The sample values in R are substitued in the expressions for a_1, a_2 and a_3:

$$\hat{\mathbf{a}} = (-0.447, 0.2511, -0.1032)$$

The variance-covariance matrix:

$$\hat{\mathbf{\Phi}} = \frac{1}{1.415} \begin{pmatrix} 0.6598 & 0.1056 & 0.1265 \\ & 0.9226 & 0.3893 \\ & & 0.9377 \end{pmatrix}$$

Consequently,

$$\hat{\sigma}_\infty = \sqrt{\hat{\mathbf{a}}\hat{\mathbf{\Phi}}\hat{\mathbf{a}}'} = \sqrt{\frac{0.0481}{1.415}} = 0.0058$$
$$r^2_{0(12)} - r^2_{01} \pm c\hat{\sigma}_\infty = 0.015 \pm (1.96)(0.0058) = [0.004, 0.027]$$

The family's use of abusable substances contributes to explaining use in school, above and beyond the effects of friends.

References

- Olkin, I., & Finn, J. D. (1995). Correlations Redux.
- java.dzone.com/articles/damn-r-squared
- en.wikipedia.org/wiki/Coefficient_of_determination
- en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient