The Significance of the Adjusted R Squared

Olivier Binette
Anthony Coache

UQÀM

The Problem

In the context of least squares linear regression, we relate n observed responses to a design matrix $X \in \mathbb{R}^{n \times p}$ through $Y = X\beta + \varepsilon$ where β is an unknown parameter and ε the associated errors.

One of many summary statistics arising from this model is the adjusted R^2 coefficient

$$R_a^2(Y, X) = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{Y}\|^2} \frac{n-1}{n-p},$$

where $\hat{\varepsilon}$ is the vector of residual errors. It is widely used as a measure of goodness of fit, as a model selection criterion and as an estimator of the squared multiple correlation coefficient ρ^2 of the parent population.

However, the literature is scarce in explanations as to what, exactly, R_a^2 adjusts for in non-trivial cases. It is not an unbiased estimator of ρ^2 and the degrees of freedom adjustment heuristic (Theil, 1971) is of limited depth.

Dozens of other adjustments of R^2 have been proposed in the literature. Why and how should we use this one? In what sense is it "unbiased"?

Previous Work

It has been shown in Cramer (1987), that if $\varepsilon \sim N(0, \sigma^2 I_n)$ and $\tilde{X}$ is an augmented design matrix such that $\text{span}(\tilde{X}) \supset \text{span}(X)$, then

$$\mathbb{E}\left[R_a^2(Y, \tilde{X})\right] = \mathbb{E}\left[R_a^2(Y, X)\right].$$

This shows a sort of "unbiasedness", but the expectation still intricately depends on the unknown parameters. The proof provided by Cramer is also relatively complicated, involving the Kummer function. This limits its pedagogical uses and falls short of the natural explanation we're after.

A test of significance for linear regression using the R_a^2 can also be found in Quinino & al. (2013), in which they focused on its didactic approach. However, their framework is adapted to evaluate the utility of a given model instead of comparing nested models.

The First Idea

For the purpose of nested model comparison, we should condition on past model performance and consider the distribution

$$I(Y, \tilde{X} \mid X) = R_a^2(Y, \tilde{X}) \mid R_a^2(Y, X).$$

There are two main justifications for this: under the model $Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I_n)$, we have

1. The distribution of $I(Y, \tilde{X} \mid X)$ is very simple*, with

$$\mathbb{E}\left[I(Y, \tilde{X} \mid X)\right] = R_a^2(Y, X).$$

2. The nested model test resulting from the use of this conditional statistic is equivalent to Fisher's classical F-test.

This explains what underlies Cramer's result: conditioning on current model performance shows how R_a^2 adjusts for irrelevant covariates and allows its use as a test statistic for model selection.

A Dual Perspective

The willing soul may take the preceding one step further, conditioning on the whole vector Y instead of only $R_a^2(Y, X)$. Inference is now restricted to the content of $\text{span}(\tilde{X}) \backslash \text{span}(X)$. Models are compared by asking the question: do the additional covariates in $\tilde{X}$ bring more information* than random covariates?

Now to formalize the idea, suppose that Y is observed (non-random). Write $\tilde{X} = [X\ W] \in \mathbb{R}^{n \times \tilde{p}}$, the concatenation of X with k new covariates $W = [W_1\ \cdots W_k]$. The null hypothesis we consider is

$$H_0 : \quad \{W_i\} \text{ is indep. of uniform directions.}$$

The following theorem shows, in particular, that the expected adjusted R squared is invariant under the addition of such random covariates.

Theorem

Under H_0 we have

$$\mathbb{E}\left[R_a^2(Y, \tilde{X})\right] = R_a^2(Y, X)$$

and $R_a^2(Y, \tilde{X})$ is distributed as

$$1 - \frac{(n-1)\|\hat{\varepsilon}\|^2}{(n-\tilde{p})\|Y - \bar{Y}\|^2} \text{Beta}\left(\tfrac{n-\tilde{p}}{2}, \tfrac{k}{2}\right).$$

Furthermore, the resulting test of H_0 is equivalent to Fisher's classical F-test of nested models.

Model Selection

We define conditional random covariate tests (CRCTs) as such tests of covariate randomness taken through a given performance metric and which are conditional on Y.

This framework is easily adapted to arbitrary models, such as GLMs, and provides exact tests of covariate randomness. The result of this test is used for model comparison.

Illustrations

We simulate datasets from

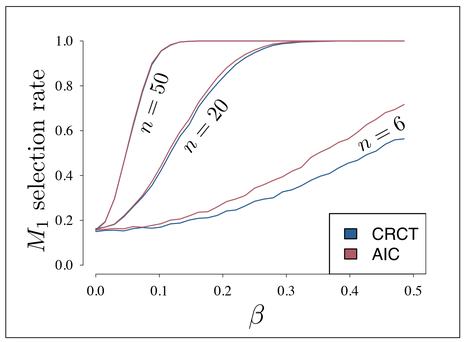
$$Y_i \sim \text{Pois}(\exp(1 + x_{1,i} + x_{2,i}\beta))$$

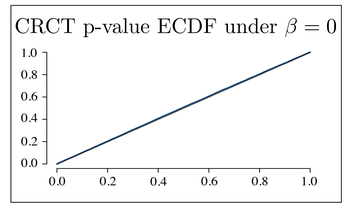
for $i \in \{1, 2, \ldots, n\}$ and $\beta \in [0, 1/2]$, where $x_{1,i} = \log(i)$ and $x_{2,i} = \mathbb{I}_{2|i}$. The Poisson GLM models given by the R formulas

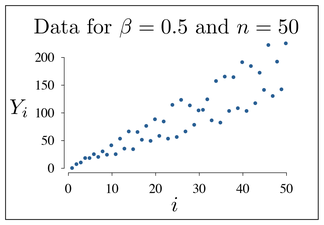
$$M_0 : Y \sim 1 + x_1$$
$$M_1 : Y \sim 1 + x_1 + x_2$$

are considered.

Our key experimental result is that, under the hypothesis $\beta = 0$, the CRCT p-value (w.r.t. the likelihood) is uniformly distributed. Furthermore, when used for model selection, it behaves similarly to the AIC criterion. This suggests CRCT statistics may provide valid alternatives to ad hoc and approximate procedures.







References

- Quinino, R. C., Reis, E. A., & Bessegato, L. F. (2013). Using the coefficient of determination R2 to test the significance of multiple linear regression. *Teaching Statistics*, 35(2), 84-88.

- Cramer, J. S. (1987). Mean and variance of R2 in small and moderate samples. *Journal of Econometrics* 35(2), 253 – 266

- Theil, H. (1971). Principles of Econometrics. J. Wiley, New York.