Maastricht University

Assignment 1

Faculty of Science and Engineering

Computational Statistics

KEN4258

Leonardo Ercolani

Lisa Incollingo

Chiara Magrone

27 February 2024

1 Task 1: Create a Monte Carlo simulation to illustrate the problem

In this analysis[1], we address the issue of inflated R^2 values in models where the number of predictors p nearly equals the number of observations n.
We utilized a Monte Carlo simulation, setting observations n at 100 and predictors p at 98. This choice was made to illustrate the extreme scenario where the number of predictors is almost the same as the number of observations.
The simulation iterates 1000 times, each time generating a new dataset with 100 observations and 98 predictors, where both the predictors X and the response variable y are generated from a normal distribution. A linear model is then fitted to each generated dataset, and the R^2 value is calculated.
The distribution of these R^2 values is visualized using a histogram in Figure(1).

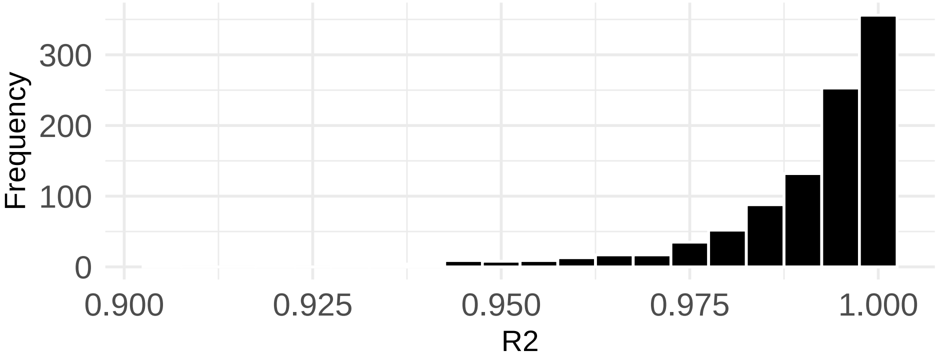

Figure 1: R² Distribution

In the analysis, the graph reveals a concerning pattern where R^2 values approach 1 across numerous iterations, with all observed values falling between 0.9 and 1. This trend signifies model saturation, potentially leading to overfitting. While such high R^2 values may initially appear a good result, they do not guarantee the model's effectiveness on new, unseen data. To further explore the R^2 metric's sensitivity, a second Monte Carlo simulation have been done, always with 1000 iterations and always with 100 observations, but changhing the number of predictors across four scenarios: 5, 20, 50, and 98.

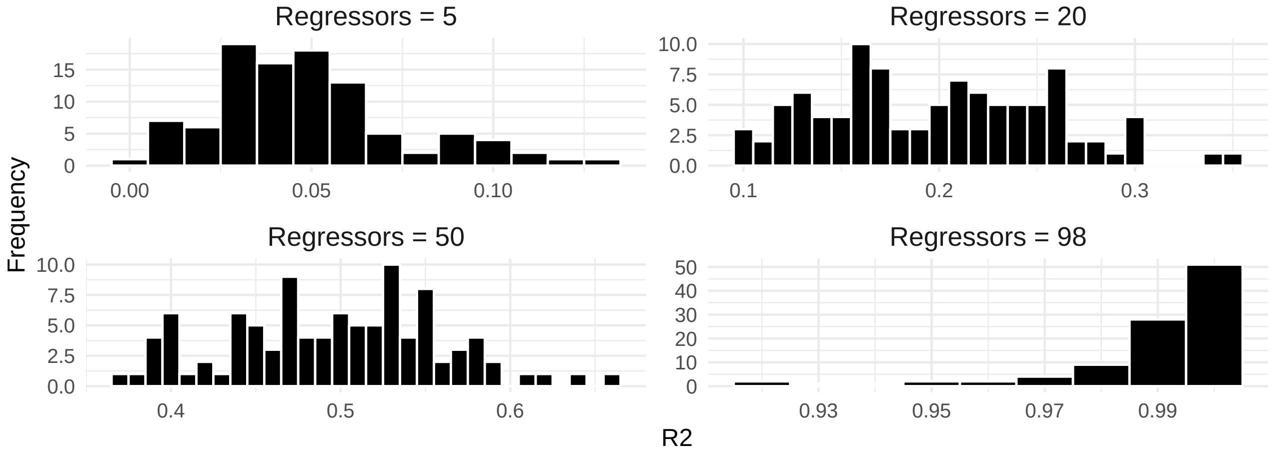

Figure 2: R² Distribution by Number of Regressors

Figure (2) clearly shows R^2's dependence on the number of predictors: R^2 fluctuates between 0 and 0.15 for 5 predictors, expanding to 0 to 0.7 with 50 predictors.

[1] https://github.com/LisaIncollingo24/Computational-Statistics.git

2 Task 2: Provide a mathematical proof showing that the problem really exists

In the context of multivariate linear regression, data can be described by the following matrix model: $Y = X\beta + \epsilon$ where Y represents the vector of responses, X is the matrix of independent variables, i.e., observations, β is the vector of coefficients to be estimated, and ϵ denotes the vector of errors, assumed distributed with zero mean and constant variance. After estimating the parameters through the Ordinary Least Squares method, we obtain a fitted regression model, expressed as $Y = X\hat{\beta} + \hat{\epsilon}$, where $\hat{\beta}$ are the estimated coefficients.

At this point we can introduce the term R^2, the coefficient of determination, is a useful measure to assess the goodness of fit to the data of a model. This represents the proportion of the variance of the dependent variable that is predictable from the independent variables:

$$R^2 = 1 - \frac{\sigma^2_{\text{res}}}{\sigma^2_y} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

where:

- σ^2_{res} is the residual variance, that is, the variance of y that is not explained by the model. It thus represents the sum of the squares of the residual errors. Specifically, the residual variance is calculated: $\sigma^2_{\text{res}} = \sigma^2_y - \sum_{i=1}^k \hat{\beta}^2_i \cdot \sigma^2_{x_i}$. Where $\sum_{i=1}^k \hat{\beta}_i \sigma^2_{x_i}$ is the explained variance, and each addend of the summation represents the covariance between regressed and individual regressors.
- σ^2_y is the total variance of the dependent variable y before any regression model is applied. It represents the total variance of Y with respect to its mean, where $\bar{y}$ is the mean of the observed values of Y. It is actually the sum of the explained variance and the residual variance.

Thus, the R^2 measures the fraction of the total variance of Y that is explained by the model. As we add more regressors to our model, the variance explained tends to increase. This, in turn, reduces the residual variance because more information is incorporated into the model to explain Y. As a result, R^2 tends to increase because a larger share of the total variance of Y is "captured" by the model.

Although the addition of new regressors may reduce the residual variance by increasing the complexity of the model, this may not correspond to a real improvement in the predictive ability of the model. In fact, adding noninformative variables may lead to an artificial increase in R^2, due to the nature of this metric that favors more complex models without penalty for adding such variables. This phenomenon is known as overfitting: the model becomes overly complex, fitting the noise of the training data rather than the underlying relationship between independent and dependent variables. Overfitting manifests itself with high R^2 on training data but poor predictive ability on unseen data.

So how do we tell if the improvement in the goodness of fit of the model is fictitious as the number of regressors increases? What is the effect attributable to each regressor? To answer these questions, we introduce the adjusted R^2.

3 Task 3: Propose a solution to address the problem

The question we ask is: how much is the additional contribution of the new explanatory variable?

$$R^2_{\text{ADJ}} = 1 - \left[(1 - R^2)\left(\frac{n-1}{n-k-1}\right)\right]$$

where:

- n represents the number of observations.
- k represents the number of independent variables, i.e., the regressors.

- $\frac{n-1}{n-k-1}$ is the adjustment ratio that takes into account the number of observations and the number of regressors; this term acts as an adjustment factor.

Accordingly, we note that $R^2 \geq R^2_{\text{ADJ}}$ always applies because we are introducing a penalty in the adjustment index if the model includes too many regressors. The adjustment term slightly penalizes the goodness of fit to the data if the number of observations and regressors is low. On the other hand, if the number of observations is small and the number of regressors is large, the adjustment term can significantly penalize R^2, potentially leading it to be negative.

In conclusion, using the adjusted R^2, we can compare with the unadjusted R^2 to assess whether the proposed model is structurally adequate, that is, whether we have included too many regressors, risking overfitting. A large discrepancy between the two measures suggests that we may have too many regressors in the model.

4 Task 4: Find a real dataset to illustrate the problem and your fix

To illustrate the issue with R^2's tendency to increase with more variables, and how adjusted R^2 can address this, we used the “attitude” dataset from R. It consists of 30 observation and 7 variables from a survey of clerical employees at a large organization, where the response variable is the overall rating, and the predictors include aspects like complaints, privileges, and learning.

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.78708	11.58926	0.931	0.361634
complaints	0.61319	0.16098	3.809	0.000903 ***
privileges	-0.07305	0.13572	-0.538	0.595594
learning	0.32033	0.16852	1.901	0.069925 .
raises	0.08173	0.22148	0.369	0.715480
critical	0.03838	0.14700	0.261	0.796334
advance	-0.21706	0.17821	-1.218	0.235577
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error: 7.068 on 23 degrees of freedom				
Multiple R-squared: 0.7326, Adjusted R-squared: 0.6628				
F-statistic: 10.5 on 6 and 23 DF, p-value: 1.24e-05				

Table 1: Summary of Linear Model

The model summary, which incorporates all variables, shows a notably high R^2 at 0.73, suggesting a strong fit. However, the adjusted R^2 stands at 0.66, being more conservative. This difference of 7 percentage points highlights the adjusted R^2's effectiveness in offering a more accurate measure of model fit by considering the number of predictors. The summary also reveals several variables that significantly contribute to inflating the R^2 unnecessarily, indicating their limited relevance in the model.