KEN4258: Computational Statistics

Assignment 1

Aurélien Bertrand Bart van Gool Gaspar Kuper
Ignacio Cadarso Quevedo Nikola Prianikov

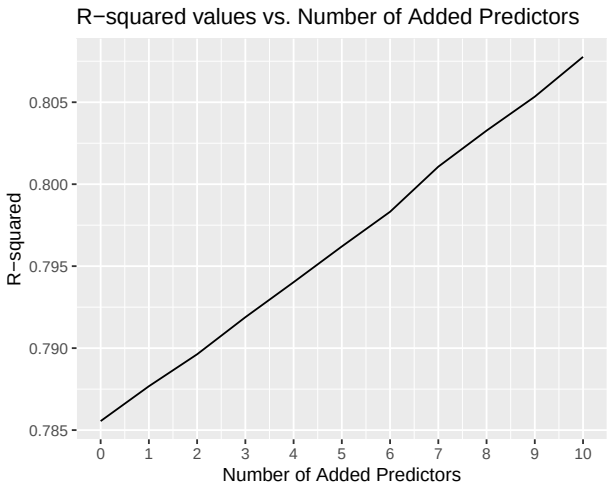
February 2024

Assignment 1

Link to our GitHub repository: https://github.com/nprianikov/compstats

1) Create a Monte Carlo simulation to illustrate the problem

The Monte Carlo simulation below is intended to show that the R^2 of a model will never decrease, even if predictors are added which don't have anything to do with the data. First data is generated by using 5 predictors, and a linear model is fit to it. After this, 10 random predictors are generated and added to the model. The plot at the bottom of Section 1 shows how the R^2 value continues to grow even though the predictors bare no actual relation to the data.



We can see that the R^2 increases with the number of predictors even though they have no relationship with the generated data. Although these random predictors can slightly increase the accuracy on the training data, they make make the true performance of the model worse. This suggests that R^2 does not reflect the true goodness (accuracy) of the fit.

2) Provide a mathematical proof showing that the problem really exists.

The R^2 score is given by

$$R^2 = \frac{\text{Var}(X\hat{\beta})}{\text{Var}(Y)} = \frac{\text{Var}(Y) - \text{Var}(e)}{\text{Var}(Y)} = 1 - \frac{\text{Var}(e)}{\text{Var}(Y)}$$

In an unbiased mode, $\text{Var}(e)$ depends on the the sum of squared residuals. As the sum of squared residuals grows, so does $\text{Var}(e)$.

Claim: the sum of squared residuals can either decrease or stay the same when adding parameters to the models. To illustrate this we define a simple model $Y = \alpha + X^{(1)}\beta_1 + \epsilon$ with a single parameter,

and $Y' = \alpha + X^{(1)}\beta_1 + X^{(2)}\beta_2 + \epsilon$ that extend first model with the second parameter. The residuals are defined respectively as

$$e_i^{(1)} = Y_i - \alpha - X_i^{(1)}\beta_1$$

$$e_i^{(2)} = Y_i - \alpha - X_i^{(1)}\beta_1 - X_i^{(2)}\beta_2$$

Due to the fact that OLS minimizes the criterion $\sum_i e_i^2$, it aims to find the set of parameters that minimizes the sum of squared residuals. In the model with a single parameter, OLS will find an optimal β_1 that minimizes the sum of squared residuals given the constraint of the model structure. Adding another predictor $X^{(2)}$ allows the model to choose a value for β_2 which could allow the sum of squared residuals to become even lower. There are two cases.

Case 1: $X^{(2)}$ can help the model If this is the case, the model will choose a value for β_2 which lowers the sum of squared residuals when compared to the model with only one predictor.

Case 2: $X^{(2)}$ can not help the model In this case, the model will simply set the parameter β_2 to 0, essentially ignoring the new predictor. As this only leaves a single predictor, the same one as in the original model, the value for β_1 will also be the same as it was before. The sum of squared residuals will therefore remain the same.

In both cases the sum of squared residuals does not increase, and as a result R^2 will either increase or stay the same.

3) Propose a solution to address the problem.

As shown in part 2, the problem with R^2 is that it is highly influenced by the number of predictors used in the model. In particular, as the number of predictors grows, the R^2 score of the model can only increase or stay the same but not decrease. The issue is then that predictors can be included in the model which don't actually have any relationship to the data, but allow the model to gain some accuracy anyway. Although adding such predictors to the model may seem like a good idea when looking at the R^2 score, it can cause the model to overfit to the training data.

This drawback of the R^2 metric has been well documented, and a solution to it has been proposed in the adjusted R^2. The formula for the adjusted R^2 metric is the following:

$$\text{Adjusted } R^2 = 1 - (1 - R^2) \times \frac{n-1}{n-p-1} = 1 - \frac{\text{Var}(e)}{\text{Var}(Y)} \times \frac{n-1}{n-p-1}$$

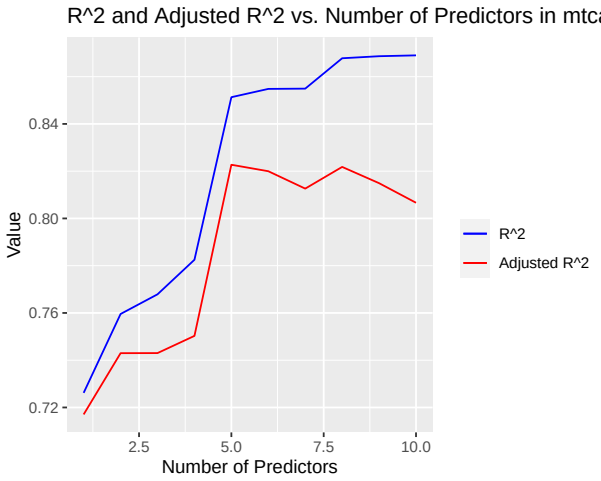
In the adjusted R^2 an extra term is introduced at the end which is a function of n and p. Since n is a constant, this term is only influenced by p. As p gets larger, the denominator gets smaller, causing this term to increase. This means that as new predictors are introduced to the model, the adjusted R^2 metric will apply a penalty. Therefore, if a predictor does not add enough predictive power to the model to counteract this penalty, the adjusted R^2 score will go down.

The adjusted R^2 score, by applying this penalty to unneeded predictors, discourages overfitting of models. This can come in use, for example when comparing models trained using different amounts of predictors. If model A is able to achieve the same accuracy as model B, but it uses less predictors to do so, then model A would be preferred by the adjusted R^2 while the regular R^2 would give both models the same score.

4) Find a real dataset to illustrate the problem and your fix.

Here, we used the mtcars dataset, which contains 32 observations and 11 numeric variables. We fit a linear regression model starting with 1 predictor, then added predictors until all the variables were used. The goal was to investigate the impact of adding predictors to R^2 and adjusted R^2.

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```



As we can see on the plot above, whenever adding a predictor, we notice either an increase or no change to R^2. However, adjusted R^2 sometimes decreases when a new predictor is added. This comes from the fact that adjusted R^2, unlike R^2, penalizes the model for including parameters that don't help increase the accuracy enough.

Using these two metrics together, we can argue that using 5 predictors for the example above is sufficient, since adding the rest of the predictors does not improve the model sufficiently much, even though R^2 increases.