Statistics. HW1. R-Squared problem

Changalidi Anton
(Dated: February 29, 2024)

R_2 has a problem with datasets that have a large number of predictors with respect to the number of observations. Let's investigate it here!

PART 1. PROBLEM STATEMENT

Methods

Several experiments were made using Monte-Carlo method: each parameter set was launched for 500 iterations, therefore the empirical (experimental) estimation will be close to the real theoretical value.

Let:

- Y - target variable, amount of observations: n_{obs}; $dim(Y)$=$(n_{obs}, 1)$;
- X - all variables on which Y actually depends, distribution: $N(0, 1)$, it's amount: n_{true}; $dim(X)$=(n_{obs}, n_{true});
- Z - noise variables (variables on which Y does not depend), distribution: $N(0, 1)$, it's amount: n_{noise}; $dim(Z)$=(n_{obs}, n_{noise}).
- ϵ - error, $N(0, 1)$.

The real dependency will be: $Y = \sum_{i=1}^{n_{true}} X_i + \epsilon = \sum_{i=1}^{n_{true}} 1 \cdot X_i + \sum_{i=1}^{n_{noise}} 0 \cdot Z_i + \epsilon$ (all β_i near X are equal to 1, and near Z (noise variables) are equal to 0).

The prediction model will be: $Y = \sum_{i=1}^{n_{true}} \beta_i \cdot X_i + \sum_{i=1}^{n_{noise}} \beta_{n_{true}+i} \cdot Z_i + \epsilon$. **(*)**

The experiments will be:

- R^2 **number of noise variables**: $n_{obs} = 500$, $n_{true} \in [1, 10]$, $n_{noise} \in [0, 1, 10, 50, 100, 200, 500]$. In some experiments amount of noise variables 10-100 times more, than amount of real predictors, that will perfectly depict R^2 distribution in such a situations.
- R^2 **number of noise variables and amount of observations**: $n_{obs} \in [10, 50, 100, 500, 1000, 5000]$, $n_{true} = 1$, $n_{noise} \in [0, 1, 10, 50, 100, 200, 500]$. Here true variable (just one) is fixed, and there will be a lot of noised variables. The distribution of r^2 depending on amount of observations will be measured, leading to understanding the importance of the proper amount of samples.

Results

For the first experiment (Fig. 1A and 1B) amount of noise variables increases the R^2: for the first experiment (only one true-influence variable): from 0.5 (0 noise variables) to 0.75 for (500 noise variables), which is a significant increase (variation is insignificant compared to the growth rate)! Even though R^2 was already high in the second experiment (due to the large number of true variables), the increase in R^2 with increasing noisy variables unrelated to the target is still clearly highlighted.

The second experiment is perfectly depicted in Fig. 1D: R^2 grows as the ratio of the number of samples to the number of noisy variables increases, and grows till $\frac{n_{noise}}{n_{obs}} = 10^0 = 1$, which is exactly when there are fewer data than variables in the analysis. Fig. 1C shows the same in a different way: the growth of R^2 is from 0.5 to 1.

PART 2. MATHEMATICAL PROOF

Note: In the real situations we do not know which variables are noise, and which have true influence on the target. That is why despite the fact, that here we operate with x and z, n_{true} and n_{obs}, in reality we will just have x and n (some of variables will be "true", some "noise"). But we separate it in the proof without loss of generality.

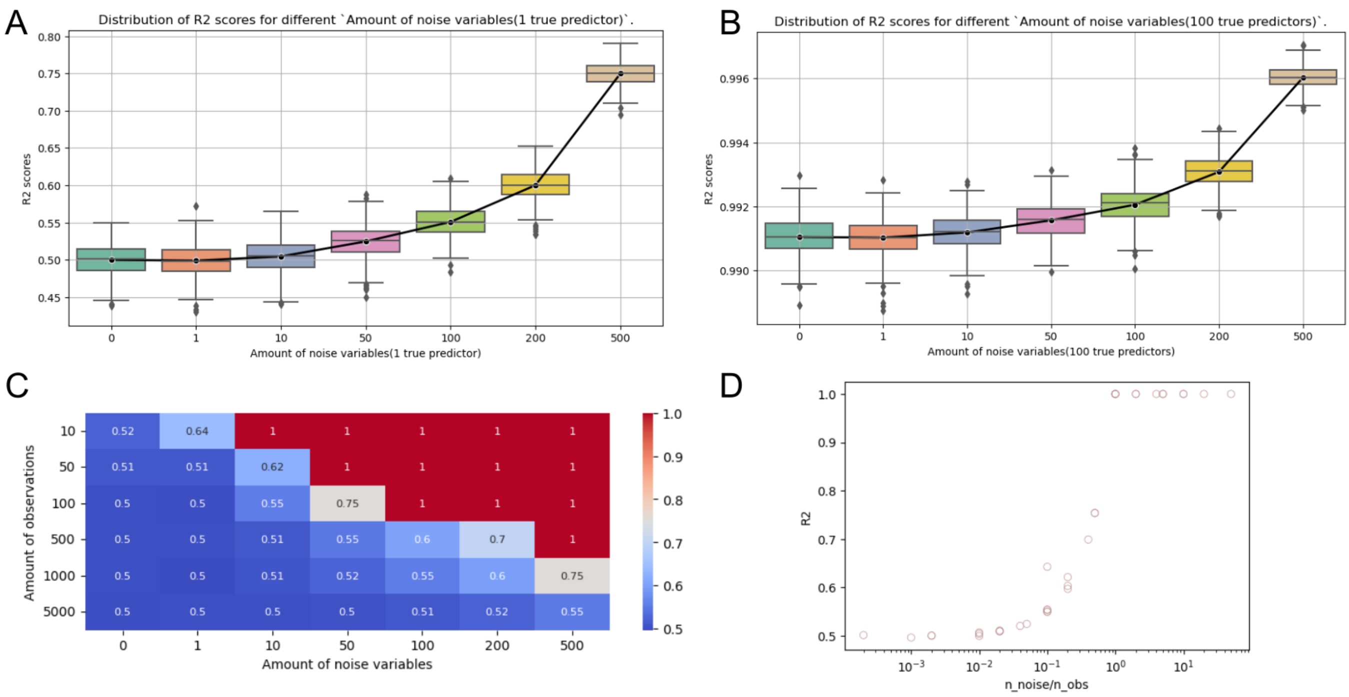

FIG. 1: A and B: Distribution of dependence of R^2 on the number of noise variables ($n_{true} \in [1, 10]$ for A and B respectively). C and D: Distribution of dependence of R^2 on the number of noise variables and amount of observations (C: mean values are represented, D: every dot is an observation).

Let f_j - model's prediction, $\overline{y}$ - average over all input y_j, then by definition (first equality) and from the (****) of part one (second equality):

$$R^2 = 1 - \frac{\sum_{j=1}^{n_{obs}}(y_j - f_j)^2}{\sum_{j=1}^{n_{obs}}(y_j - \overline{y})^2} = 1 - \frac{\sum_{j=1}^{n_{obs}}(y_j - (\sum_{i=1}^{n_{true}} \beta_i \cdot x_{j,i} + \sum_{i=1}^{n_{noise}} \beta_{n_{true}+i} \cdot z_{j,i}))^2}{\sum_{j=1}^{n_{obs}}(y_j - \overline{y})^2}$$

Denominator ($\sum_{j=1}^{n_{obs}}(y_j - \overline{y})^2$) does not depend on the model (so it is constant if dataset does not change). In numerator we subtract from y_j: $\sum_{i=1}^{n_{true}} \beta_i \cdot x_{j,i}$ and $\sum_{i=1}^{n_{noise}} \beta_{n_{true}+i} \cdot z_{j,i}$. Here is where the problem comes from!

If we will not take noise variables (z_j) into account, we will only subtract $\sum_{i=1}^{n_{true}} \beta_i \cdot x_{j,i}$; however if we will add noised variables to the model. The best case scenario, if all β_i for $i \in [n_{true}+1, n_{true}+n_{obs}]$ (coefficients before noised variables) are equal to zero: then we got the same model.

Otherwise, we will also subtract $\sum_{i=1}^{n_{noise}} \beta_{n_{true}+i} \cdot z_{j,i}$, making $\sum_{j=1}^{n_{obs}}(y_j - (\sum_{i=1}^{n_{true}} \beta_i \cdot x_{j,i} + \sum_{i=1}^{n_{noise}} \beta_{n_{true}+i} \cdot z_{j,i}))^2$ smaller, therefore making $\frac{\sum_{j=1}^{n_{obs}}(y_j - (\sum_{i=1}^{n_{true}} \beta_i \cdot x_{j,i} + \sum_{i=1}^{n_{noise}} \beta_{n_{true}+i} \cdot z_{j,i}))^2}{\sum_{j=1}^{n_{obs}}(y_j - \overline{y})^2}$ smaller, and therefore making $R^2 = 1 - \frac{\sum_{j=1}^{n_{obs}}(y_j - (\sum_{i=1}^{n_{true}} \beta_i \cdot x_{j,i} + \sum_{i=1}^{n_{noise}} \beta_{n_{true}+i} \cdot z_{j,i}))^2}{\sum_{j=1}^{n_{obs}}(y_j - \overline{y})^2}$ bigger with increasing number of noised variables.

Q.E.D.

PART 3. SOLUTION

The solution is to use adjusted R^2 (proposed by Mordecai Ezekiel, 1930).

If amount of vaiables is $m(= n_{true} + n_{noise})$, then let $df_{model} = n_{obs} - m - 1$ and $df_{total} = n_{obs} - 1$ (degrees of freedom, in model it is bigger, because we introduce variables), then:

$$R^2_{adj} = 1 - \frac{\frac{1}{df_{model}} \cdot \sum_{j=1}^{n_{obs}}(y_j - f_j)^2}{\frac{1}{df_{total}} \cdot \sum_{j=1}^{n_{obs}}(y_j - \overline{y})^2} = 1 - (1 - R^2)\frac{df_{total}}{df_{model}} = 1 - (1 - R^2)\frac{n_{obs} - 1}{n_{obs} - m - 1}$$

Since $\frac{df_{total}}{df_{model}}$ is more than one, we subtract bigger value from 1, making $R^2_{adj} \leq R^2$. The function penalises for each variable introduced.

PART 4. EXAMPLE

We will take the Kaggle dataset for Real estate price prediction (Kaggle, 2018). It has 6 predictors, one target, amount of samples is 414. We will do linear regression on the original dataset, and then start to add $n_{noise} \in [0, 1, 5, 10, 50, 100, 200]$ noise variables (from standard normal distribution), and build linear regression on this data. For each n_{noise} we will do 500 iterations, and then observe distribution of R^2 and adjusted R^2.

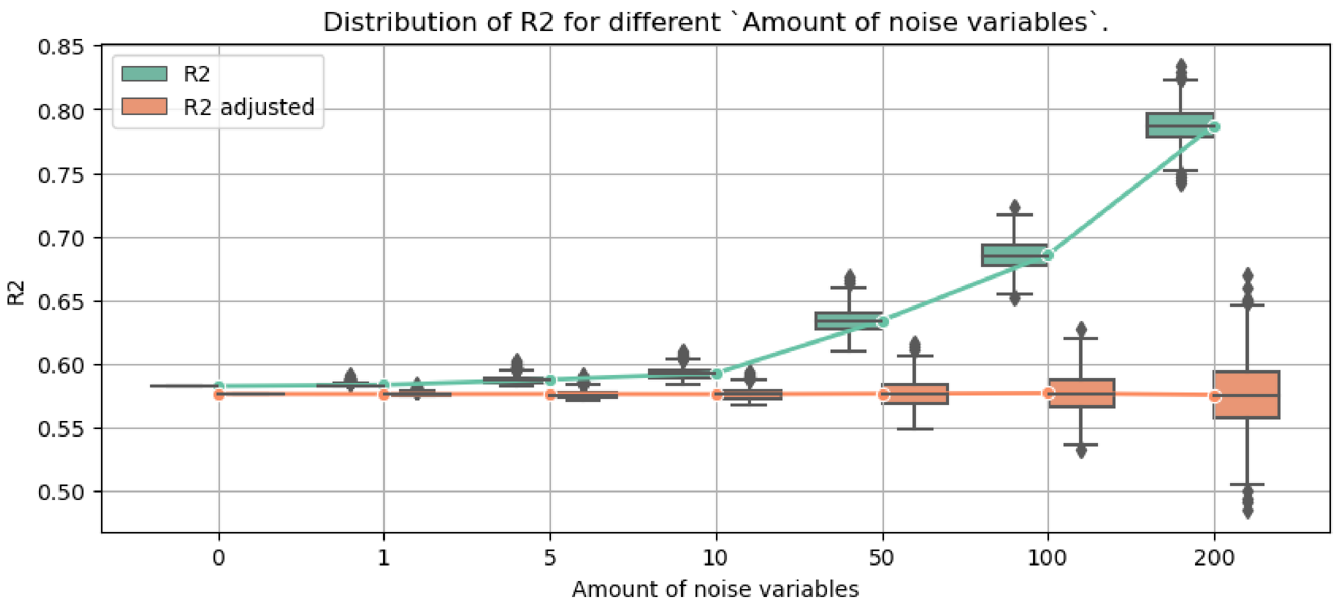

FIG. 2: Distribution of R^2 and R^2_{adj} as a function of noise variables amount.

The results are presented on the Fig.2. The increase in R^2 with increasing number of noise variables is clearly visible (from 0.57 to 0.78), with R^2_{adj} remaining almost constant (around 0.57).

That is why, in real experiments, researcher should be cautious if among many variables some variable gives an increase in R^2 (or other objective function, or it is significant in some statistical test). It is very very possible that it is just a random thing!

CODE AVAILABILITY

All code pertinent to the results presented in this work are available at:
`https://github.com/TohaRhymes/stat_um_24spring`

REFERENCES

[1] Mordecai Ezekiel (1930), Methods Of Correlation Analysis, Wiley, pp. 208-211.

[2] ALGOR_BRUCE. (2018). Real Estate Price Prediction [Dataset]. Kaggle. `https://www.kaggle.com/datasets/quantbruce/real-estate-price-prediction?resource=download`