

Chen, Qiang; Qi, Ji

Article

How much should we trust R2 and adjusted R2: Evidence from regressions in top economics journals and Monte Carlo simulations

Journal of Applied Economics

Provided in Cooperation with:

University of CEMA, Buenos Aires

Suggested Citation: Chen, Qiang; Qi, Ji (2023) : How much should we trust R2 and adjusted R2: Evidence from regressions in top economics journals and Monte Carlo simulations, Journal of Applied Economics, ISSN 1667-6726, Taylor & Francis, Abingdon, Vol. 26, Iss. 1, pp. 1-16, <https://doi.org/10.1080/15140326.2023.2207326>

This Version is available at:

<https://hdl.handle.net/10419/314225>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by/4.0/>



How much should we trust R^2 and adjusted R^2 : evidence from regressions in top economics journals and Monte Carlo simulations

Qiang Chen & Ji Qi

To cite this article: Qiang Chen & Ji Qi (2023) How much should we trust R^2 and adjusted R^2 : evidence from regressions in top economics journals and Monte Carlo simulations, Journal of Applied Economics, 26:1, 2207326, DOI: [10.1080/15140326.2023.2207326](https://doi.org/10.1080/15140326.2023.2207326)

To link to this article: <https://doi.org/10.1080/15140326.2023.2207326>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



View supplementary material [↗](#)



Published online: 02 May 2023.



Submit your article to this journal [↗](#)



Article views: 4190



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 8 View citing articles [↗](#)

How much should we trust R^2 and adjusted R^2 : evidence from regressions in top economics journals and Monte Carlo simulations

Qiang Chen and Ji Qi

School of Economics, Shandong University, Jinan, China

ABSTRACT

R^2 and adjusted R^2 may exaggerate a model's true ability to predict the dependent variable in the presence of overfitting, whereas leave-one-out R^2 (LOOR 2) is robust to overfitting. We demonstrate this by replicating 279 regressions from 100 papers in top economics journals, where the median increases of R^2 and adjusted R^2 over LOOR 2 reach 40.2% and 21.4% respectively. The inflation of test errors over training errors increases with the severity of overfitting as measured by the number of regressors and nonlinear terms, and the presence of outliers, but decreases with the sample size. These results are further validated by Monte Carlo simulations.

ARTICLE HISTORY

Received 16 November 2022

Accepted 21 April 2023

KEYWORDS

R^2 ; adjusted R^2 ; leave-one-out R^2 ; goodness of fit

1. Introduction


In empirical studies, R^2 and adjusted R^2 (denoted as \bar{R}^2) are routinely reported as measures of goodness-of-fit for linear regressions. For example, a R^2 of 0.8 is usually taken to imply that all explanatory variables jointly explain 80% of the variations in the dependent variable. But how reliable is this interpretation?

It is well known that R^2 and \bar{R}^2 only measure in-sample fit, which may not be good indicators of the model's true ability to explain or predict out of sample. In particular, it is common sense in the machine learning literature that training errors (as represented by $1 - R^2$ and $1 - \bar{R}^2$) could be poor measures of the true test errors, when the model is used to predict data that it has not yet seen. Nevertheless, as of today, most economists still happily use R^2 and \bar{R}^2 to measure goodness-of-fit, without worrying about its potential pitfalls.¹

This paper takes this issue seriously. The essential problem is that R^2 and \bar{R}^2 may exaggerate a model's true ability to explain or predict the dependent variable, especially in the presence of overfitting. Overfitting occurs when a model is excessively fit to noisy sample data (e.g., a low degree of freedom resulting from a small sample size or too many covariates, a complicated functional form with many nonlinear terms, or the presence of

CONTACT Qiang Chen  qiang2chen2@126.com  School of Economics, Shandong University, Jinan, China

¹To be sure, economics is not the only discipline in this regard. For example, Parady et al. (2021) laments the overreliance on statistical goodness-of-fit and under-reliance on model validation in the transportation literature.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/15140326.2023.2207326>.

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

outliers), which compromises the model's ability to uncover the true relationship between the dependent and explanatory variables, as well as its performance in out-of-sample prediction.

To solve this problem, we recommend using leave-one-out cross-validated R^2 ($LOOR^2$ in short) as a better measure of goodness-of-fit for linear regressions. While $LOOR^2$ has been around for some time, this paper takes it seriously and suggests that economists should routinely report $LOOR^2$ in their empirical work alongside R^2 and adjusted R^2 (if not at the expense of the latter two). There are a number of advantages associated with $LOOR^2$. First, $LOOR^2$ is robust to overfitting, as it measures the true test errors and the model's real ability to explain or predict the dependent variable. Second, while five-fold or ten-fold cross-validations are popular in machine learning to measure test errors, the results are uncertain due to random splitting of the sample into five or ten folds (parts) of roughly equal sizes. On the other hand, the results from leave-one-out cross-validation is certain, since one observation is left out at a time, and no random sampling is involved. Last but not least, for linear regressions, there is a short-cut formula for computing $LOOR^2$ such that only one regression is needed, thus the computational cost is minimal.

To support the above claims, we replicate 279 regressions from 100 empirical papers in four top economics journals during 2004–2021. In this sample, the median increases of R^2 and \bar{R}^2 over $LOOR^2$ reach 40.2% and 21.4%, respectively, implying that both R^2 and \bar{R}^2 often exaggerate the estimated model's true ability to explain or predict the variations in the dependent variable to a large extent. Moreover, we introduce “error inflation factor” (EIF) and “adjusted error inflation factor” (adjusted EIF) to measure the inflation of test errors (i.e., $1 - LOOR^2$) over training errors using R^2 and adjusted R^2 (i.e., $1 - R^2$ and $1 - \bar{R}^2$) respectively. The regression results show that both EIF and adjusted EIF increase with the severity of overfitting as measured by the number of regressors and nonlinear terms, and the presence of outliers, but decrease with the sample size. These results are further validated by Monte Carlo simulations.

Statisticians have long recognized that R^2 could be deceptively large as a measurement of a model's true predictive ability on subsequent data. In fact, this recognition motivated the development of adjusted R^2 as a way to shrink R^2 by degree-of-freedom adjustment (Larson, 1931; Wherry, 1931).² However, Mayer (1975) demonstrates empirically that even \bar{R}^2 is a poor guide to the post-sample fit, which may be caused by excessive data mining. An alternative route to the solution relies on cross-validation including leave-one-out cross-validation (Cochran, 1968; Hills, 1966; Lachenbruch & Mickey, 1968; Mosteller & Tukey, 1968), which turns out to be a more fruitful approach. Moreover, Efron and Morris (1973), Geisser (1974) and Stone (1974) propose to use cross-validation for model selection. For a modern survey on the methodology of cross-validation, see Arlot and Celisse (2010). This paper follows the tradition of cross-validation, as it measures test errors directly.

The rest of the paper is arranged as follows. Section 2 introduces leave-one-out R^2 ($LOOR^2$), error inflation factor (EIF), and adjusted error inflation factor (adjusted EIF). Section 3 studies the determinants of EIF and adjusted EIF via a meta-analysis by replicating 279 regressions from 100 prominent economic papers. Section 4 conducts

²The original formula for adjusted R^2 was first proposed in a paper by M. J. B. Ezekiel, who read it before the Mathematical Society at its annual meeting in 1928, but gave the credit to B. B. Smith.

Monte Carlo simulations for further investigation. [Section 5](#) provides conclusion and suggestions for empirical researchers.

2. Leave-one-out R^2 and error inflation factor

Consider the following linear regression model with n observations,

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \dots, n), \quad (1)$$

where y_i is the dependent variable for an individual i , and \mathbf{x}_i is a $k \times 1$ vector of explanatory variables, $\boldsymbol{\beta}$ is the corresponding $k \times 1$ vector of parameters, and ε_i is the error term. The model can be written in a matrix form,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2)$$

where $\mathbf{y} = (y_1 \dots y_n)'$, $\mathbf{X} = (\mathbf{x}_1 \dots \mathbf{x}_n)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1 \dots \varepsilon_n)'$. The well-known OLS estimator is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. With $\hat{\boldsymbol{\beta}}$ estimated and the fitted values given by $\hat{y}_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}}$, we have $R^2 = [\text{Corr}(y_i, \hat{y}_i)]^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$ in the presence of a constant term,³

and adjusted R^2 given by $\bar{R}^2 = 1 - \frac{\sum_{i=1}^n e_i^2 / (n-k)}{\sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)}$, where \bar{y} is the sample mean of y_i , and e_i is the OLS residual.

To implement leave-one-out regression omitting individual i , we simply run OLS regression with all but the i th observations. Denoting $\mathbf{X}_{(-i)}$ as the data matrix \mathbf{X} without the i th row, and $\mathbf{y}_{(-i)}$ as the outcome vector \mathbf{y} without the i th element, the OLS estimator leaving out the i th observation is simply,

$$\hat{\boldsymbol{\beta}}_{(-i)} = (\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)})^{-1}\mathbf{X}'_{(-i)}\mathbf{y}_{(-i)}. \quad (3)$$

With $\hat{\boldsymbol{\beta}}_{(-i)}$ estimated, we can make an out-of-sample prediction for the i th observation as $\hat{y}_{(-i)} = \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{(-i)}$. Repeat the procedure for all observations in the sample to obtain $\{\hat{y}_{(-i)}\}_{i=1}^n$, and the leave-one-out R^2 ($LOOR^2$) is given by

$$0 \leq LOOR^2 = [\text{Corr}(y_i, \hat{y}_{(-i)})]^2 \leq 1, \quad (4)$$

where $\text{Corr}(y_i, \hat{y}_{(-i)})$ is the correlation coefficient between y_i and $\hat{y}_{(-i)}$.

The procedure to compute $LOOR^2$ appears to be cumbersome as it entails running n regressions, which may be computationally costly if the sample size n is very large. Fortunately, for linear regressions, there is a short-cut formula for running leave-one-out regression omitting the i th observation (Hansen, 2022, Chapter 3),

³We ignore the case of linear regression without a constant term, as it is rarely encountered in practice.

$$\hat{\beta}_{(-i)} = \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\tilde{e}_i, \quad (5)$$

where $\tilde{e}_i = \frac{e_i}{1-\text{lev}_i}$ is a scaled version of the OLS residual e_i using the full sample, and $\text{lev}_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ is known as the leverage for the i th observation. Using Equation (5), the leave-one-out coefficient $\hat{\beta}_{(-i)}$ can be readily computed with existing information. Therefore, in the case of linear regressions, only one regression is needed to compute $LOOR^2$ after all. Thus, calculating $LOOR^2$ in addition to R^2 and adjusted R^2 only imposes a minimal computational cost for linear regressions.⁴

After introducing $LOOR^2$, a natural question arises about the relationship among R^2 , adjusted R^2 and $LOOR^2$. In general, R^2 and adjusted R^2 are larger than $LOOR^2$, as it is usually more difficult to make out-of-sample predictions than in-sample predictions. For example, as simulations in Section 4.1 show, when noise variables are added to the regression, R^2 keeps rising while adjusted R^2 remains stable, but $LOOR^2$ declines steadily.

To see it from a different perspective, $(1 - R^2)$ and $(1 - \text{Adjusted } R^2)$ are generally smaller than $(1 - LOOR^2)$, as training errors are usually smaller than test errors. To measure the “inflation” of test errors over training errors, we define an error inflation factor (EIF) and an adjusted error inflation factor (adjusted EIF),⁵

$$EIF = \frac{1 - LOOR^2}{1 - R^2}, \quad (6)$$

$$\text{Adjusted EIF} = \frac{1 - LOOR^2}{1 - \bar{R}^2}, \quad (7)$$

where \bar{R}^2 is adjusted R^2 .

We conjecture that both EIF and adjusted EIF increase with the severity of overfitting. Intuitively, when there is severe overfitting, training errors underestimate test errors to a great extent, resulting in large values of EIF and adjusted EIF. In the empirical study in Section 3, we consider three potential factors contributing to overfitting, i.e., the degree of freedom (sample size in excess of the number of regressors), the number of nonlinear terms (such as squared and interactive terms), and the presence of outliers. First, if the degree of freedom is small (e.g., a small sample size, or many regressors, or both), then linear regression is essentially fit to the noisy sample data, resulting in overfitting. Second, the presence of many nonlinear terms would increase the complexity of the regression function,⁶ and thus its ability to fit noisy data, which may also result in overfitting. Third, the nature of OLS estimation by minimizing the residual sum of squares implies that it is easily influenced by outliers, which again leads to overfitting.

⁴For example, the short-cut algorithm for computing $LOOR^2$ could be implemented in Stata by using the user-written command “cv_regress” (Rios-Avila, 2018) after the usual “regress” command for OLS regression.

⁵These terminologies are in the same spirit as “variance inflation factor” (VIF).

⁶In fact, the presence of many covariates also increases the complexity of regression function.

In summary, based on the fact that overfitting reduces in-sample training errors at the expense of increasing out-of-sample test errors, we hypothesize that overfitting would result in elevated EIF and adjusted EIF. The next section investigates these relationships empirically.

3. A meta-analysis

3.1. Data source and variable definitions

In this section, we empirically compare R^2 , adjusted R^2 , and $LOOR^2$, and investigate determinants of their gaps as represented by EIF and adjusted EIF. We focus on linear models where OLS is used for estimation in the recent literature. As a meta-analysis, our sample data is compiled by replicating linear regressions from 100 empirical papers selected from *American Economic Review* (23 papers), *Economic Journal* (35 papers), *European Economic Review* (18 papers) and *Review of Economic Studies* (24 papers) during 2004–2021.⁷ There are a total of 100 papers and 279 regression results in our sample with a sample size of 279, since each paper usually contains multiple OLS regressions.

For each of these 279 regressions, we calculate R^2 , adjusted R^2 , and $LOOR^2$, as well as the error inflation factor (EIF, denoted as eif) and the adjusted error inflation factor (adjusted EIF, denoted as eif_a). The explanatory variables include the sample size (n), the number of regressors including the constant term (k), the number of nonlinear terms (*nonlinear*) in each regression, and the maximum value of leverage (lev_max) as well as its variance (lev_var).

An explanation of these two measures of outliers is in order. As mentioned in Section 2, the leverage for the i th observation is given by $lev_i = \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$, which measures the influence of the i th observation on $\hat{\beta}$. Specifically, Equation (5) implies that

$$\hat{\beta} - \hat{\beta}_{(-i)} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1 - lev_i}. \quad (8)$$

It can be shown that $0 \leq lev_i \leq 1$ with a sample average of k/n (Hansen, 2022, Chapter 3). Therefore, a large lev_i implies a large discrepancy between $\hat{\beta}$ and $\hat{\beta}_{(-i)}$ according to Equation (8). The variable lev_max is simply the maximum leverage for each regression, which captures the greatest influence of a single observation in a particular regression. In the same spirit, one may consider the second largest leverage, the third largest leverage, and so on. But this approach gets tedious. Instead, we use the variance of leverage (lev_var) as a parsimonious representation. The rationale is that given that the sum of all leverages is equal to the number of regressors (i.e., $\sum_{i=1}^n lev_i = k$), when some leverages are very large (i.e., close to the largest possible value of 1), then

⁷These four journals are selected partly because their replication data and programs are more easily accessible. See the Appendix for a complete list of these 100 papers.

Table 1. Summary statistics.

| Variable | Observations | Mean | Median | S. D. | Min | Max |
|-------------------|--------------|---------|---------|----------|-------|----------|
| $R^2/LOOR^2$ | 279 | 175.156 | 1.402 | 2877.459 | 1.001 | 48065.68 |
| Adj. $R^2/LOOR^2$ | 279 | 136.268 | 1.214 | 2243.947 | 1.001 | 37483.06 |
| <i>eif</i> | 279 | 1.184 | 1.123 | 0.206 | 1.001 | 2.406 |
| <i>eif_a</i> | 279 | 1.090 | 1.060 | 0.098 | 1.000 | 1.837 |
| <i>n</i> | 279 | 621.473 | 245 | 1676.184 | 26 | 20269 |
| <i>k</i> | 279 | 29.108 | 11 | 75.084 | 2 | 694 |
| <i>nonlinear</i> | 279 | 1.487 | 0 | 2.639 | 0 | 13 |
| <i>lev_max</i> | 279 | 0.356 | 0.243 | 0.314 | 0.001 | 1 |
| <i>lev_var</i> | 279 | 0.004 | 0.00079 | 0.007 | 0.000 | 0.050 |

Table 2. Five regressions in Dower et al. (2021).

| Regressions | (1) | (2) | (3) | (4) | (5) |
|-------------------|-------|-------|-------|---------|----------|
| <i>n</i> | 58 | 57 | 57 | 57 | 56 |
| <i>k</i> | 4 | 7 | 10 | 12 | 14 |
| <i>nonlinear</i> | 1 | 1 | 4 | 5 | 6 |
| <i>lev_max</i> | 0.996 | 0.996 | 0.997 | 0.998 | 0.998 |
| <i>lev_var</i> | 0.020 | 0.024 | 0.039 | 0.041 | 0.047 |
| R^2 | 0.244 | 0.296 | 0.425 | 0.476 | 0.584 |
| Adjusted R^2 | 0.202 | 0.212 | 0.315 | 0.348 | 0.456 |
| $LOOR^2$ | 0.109 | 0.090 | 0.242 | 0.003 | 0.000012 |
| $R^2/LOOR^2$ | 2.239 | 3.287 | 1.759 | 151.725 | 48065.68 |
| Adj. $R^2/LOOR^2$ | 1.853 | 2.351 | 1.302 | 110.860 | 37483.06 |
| EIF | 1.178 | 1.293 | 1.318 | 1.902 | 2.406 |
| Adjusted EIF | 1.116 | 1.155 | 1.107 | 1.528 | 1.837 |

Data are from Table 2 of Dower et al. (2021), and by authors' calculation.

other leverages are squeezed towards their smallest possible value of 0, which results in an increase in the variance of leverage.

Summary statistics of the variables used in this study are presented in Table 1. While we focus on EIF (*eif*) and adjusted EIF (*eif_a*) in the regression analysis, it is intuitive to first look at the ratios ($R^2/LOOR^2$) and (adjusted $R^2/LOOR^2$) as reported in the first two rows of Table 1. The median of ($R^2/LOOR^2$) is 1.402, implying that the median increase of R^2 over $LOOR^2$ reaches 40.2% in the sample. Similarly, the median of (adjusted $R^2/LOOR^2$) is 1.214, implying that the median increase of adjusted R^2 over $LOOR^2$ is 21.4%. These show that R^2 and adjusted R^2 often exaggerate the estimated model's true ability to explain or predict the dependent variable to a large extent, as measured by $LOOR^2$.

The minimum values of ($R^2/LOOR^2$) and (adjusted $R^2/LOOR^2$) are both above 1 as expected. However, the maximum values of ($R^2/LOOR^2$) and (adjusted $R^2/LOOR^2$) reach alarming levels of 48,065.68 and 37,483.06, respectively. Therefore, it is instructive to take a closer look at these extreme values, which come from the fifth of five regressions in Dower et al. (2021), as shown in Table 2.

In an effort to estimate the value of a statistical life under Stalin's dictatorship, Dower et al. (2021) ran cross-sectional OLS regressions with 58 regions of the former Soviet Union as the units of observations. The dependent variable is the number of citizens repressed during the German and Polish operations of the Great Terror during 1937–1938 per 1000 capita. As typically done in empirical papers, Table 2 of Dower et al. (2021) reports results from five regressions. As more regressors and nonlinear terms are added

Table 3. Correlation matrix for major variables in the study.

| | eif | eif_a | n | k | nonlinear | lev_max | lev_var |
|------------------|----------|-----------|----------|----------|-----------|----------|---------|
| <i>eif</i> | 1 | | | | | | |
| <i>eif_a</i> | 0.926*** | 1 | | | | | |
| <i>n</i> | -0.139** | -0.154*** | 1 | | | | |
| <i>k</i> | 0.172*** | 0.0804 | 0.396*** | 1 | | | |
| <i>nonlinear</i> | 0.206*** | 0.344*** | -0.064 | -0.074 | 1 | | |
| <i>lev_max</i> | 0.575*** | 0.487*** | 0.057 | 0.223*** | 0.101* | 1 | |
| <i>lev_var</i> | 0.660*** | 0.637*** | -0.108* | 0.016 | 0.107* | 0.685*** | 1 |

* $p \leq 10\%$, ** $p \leq 5\%$, *** $p \leq 1\%$.

from regressions (1) through (5), R^2 increases steadily from 0.244 to 0.584, while adjusted R^2 increases from 0.202 to 0.456, indicating a significant boost to the goodness-of-fit at face value. However, while $LOOR^2$ improves in regression (3), it drops to alarmingly low values of 0.003 and 0.000012 in regressions (4) and (5).⁸ Consequently, $(R^2/LOOR^2)$ and (adjusted $R^2/LOOR^2$) reach outrageous levels of 48,065.68 and 37,483.06, respectively. Apparently, regressions (1) and (2) are underfit, whereas regressions (4) and (5) are severely overfit. Moreover, the maximum leverages are close to 1 in all regressions, indicating the presence of outliers.

3.2. Correlation analysis

As a preliminary exploration of determinants of EIF and adjusted EIF, Table 3 presents a correlation matrix for major variables in the study. EIF (*eif*) is negatively correlated with the sample size (*n*) at the 5% level, while positively correlated with the number of regressors (*k*), the number of nonlinear terms (*nonlinear*), the maximum leverage (*lev_max*) and the variance of leverage (*lev_var*) at the 1% level. The correlation pattern between the adjusted EIF (*eif_a*) and these determinants is qualitatively similar. The only exception is that adjusted EIF (*eif_a*) is not significantly correlated with the number of regressors (*k*), perhaps due to the degree-of-freedom adjustment already made in adjusted R^2 .

3.3. Regression analysis

For the determinants of Log(EIF), we start from the following baseline regression⁹

$$\ln eif_i = \beta_0 + \beta_1 \ln n + \beta_2 \ln k + \beta_3 nonlinear + \beta_4 lev_max_i + \beta_5 lev_var_i + \varepsilon_i. \quad (9)$$

In addition, we also interact $\ln n$ and $\ln k$ with *lev_max* and *lev_var* in Equation (9) to capture possible moderating effects of the sample size and number of regressors on the two measures of outliers. Our dataset consists of 279 observations (regressions) from 100 papers, where each paper contributes 2.79 regressions on average. Apparently, we have cluster data clustered at the paper level, where observations (regressions) from the same paper are likely correlated. Therefore, we use robust standard errors clustered at the

⁸Note that Dower et al. (2021) only report R^2 .

⁹The results of using EIF or adjusted EIF as the dependent variables are qualitatively similar, but the fit is slightly worse. To save space, we only report results using Log(EIF) and Log(Adjusted EIF) as the dependent variables.

Table 4. Determinants of log(EIF).

| | (1) | (2) | (3) | (4) |
|--------------------------------|-------------------------|-------------------------|------------------------|------------------------|
| <i>lnn</i> | -0.0645*** (0.00909) | -0.0471*** (0.00920) | -0.0106 (0.0188) | -0.0418** (0.0186) |
| <i>lnk</i> | 0.0788*** (0.0128) | 0.0607*** (0.00875) | 0.0597*** (0.0116) | 0.0528*** (0.00903) |
| <i>nonlinear</i> | 0.00578 (0.00424) | 0.00590 (0.00376) | 0.0171*** (0.00542) | 0.0119*** (0.00411) |
| <i>lev_max</i> | -0.0155 (0.0436) | 0.479*** (0.149) | -0.0989* (0.0525) | 1.300*** (0.327) |
| <i>lev_var</i> | 8.394*** (1.259) | 8.672 (6.288) | 12.84*** (3.307) | -26.05** (10.82) |
| <i>lnn*lev_max</i> | | -0.0600 (0.0431) | | -0.235*** (0.0522) |
| <i>lnn*lev_var</i> | | -12.13*** (4.244) | | -2.379 (4.590) |
| <i>lnk*lev_max</i> | | -0.0167 (0.0427) | | 0.00311 (0.0458) |
| <i>lnk*lev_var</i> | | 18.17*** (5.016) | | 14.75** (6.806) |
| <i>paper fixed effects</i> | No | No | Yes | Yes |
| <i>constant</i> | 0.284*** (0.0389) | 0.219*** (0.0461) | 0.0270 (0.104) | 0.222** (0.106) |
| <i>N</i> | 279 | 279 | 279 | 279 |
| <i>R</i> ² | 0.694 | 0.807 | 0.943 | 0.962 |
| Adjusted <i>R</i> ² | 0.689 | 0.800 | 0.909 | 0.939 |
| <i>LOOR</i> ² | 0.667 | 0.745 | 0.839 | 0.892 |

Cluster-robust standard errors in parentheses. * $p \leq 10\%$, ** $p \leq 5\%$, *** $p \leq 1\%$.

paper level throughout. In addition, we may also control for the “paper fixed effects” by giving observations from the same paper a specific intercept. However, since sample size (n) varies little within a paper,¹⁰ adding the paper fixed effects may reduce our ability to detect the effects of sample size (n). Therefore, we report regression results both with and without the paper fixed effects.

Table 4 reports results from OLS regressions with Log(EIF) as the dependent variable. Column (1) of Table 4 reports the results from the baseline regression (9) without the paper fixed effects. The coefficient of *lnn* is negatively significant at the 1% level, indicating that a large sample size decreases overfitting, thus reducing the EIF. On the other hand, the coefficient of *lnk* is positively significant at the 1% level, implying that more regressors increases the chance of overfitting, which contributes to increased EIF. The coefficient of *lev_var* (variance of leverage) is positively significant at the 1% level, as outliers may result in overfitting, whereas the coefficients of *lev_max* and *nonlinear* are insignificant.

Column (2) of Table 4 interacts *lnn* and *lnk* with *lev_max* and *lev_var*. The coefficient of *lnn*lev_var* is negatively significant at the 1% level, implying that the effect of *lev_var* on EIF may have been mitigated by increasing the sample size. On the other hand, the coefficient of *lnk*lev_var* is positively significant at the 1% level, indicating that the effect of *lev_var* on EIF may have been magnified by increasing the number of regressors. Interestingly, the coefficient of *lev_max* is now positively significant at the 1% level, whereas the coefficient of *lev_var* loses significance. Note that these two measures of

¹⁰Typically, the sample sizes of regressions within a paper change because of adding more variables, which may result in missing observations.

Table 5. Determinants of log(adjusted EIF).

| | (1) | (2) | (3) | (4) |
|-------------------------------|-------------------------|-------------------------|------------------------|------------------------|
| <i>lnn</i> | -0.0306*** (0.00530) | -0.0249*** (0.00616) | -0.00925 (0.0147) | -0.0318* (0.0187) |
| <i>lnk</i> | 0.0313*** (0.00599) | 0.0290*** (0.00529) | 0.0279*** (0.00688) | 0.0264*** (0.00608) |
| <i>nonlinear</i> | 0.00737* (0.00378) | 0.00697* (0.00376) | 0.0140*** (0.00432) | 0.0115*** (0.00373) |
| <i>lev_max</i> | -0.0193 (0.0282) | 0.173** (0.0822) | -0.0719 (0.0433) | 0.676** (0.273) |
| <i>lev_var</i> | 4.933*** (1.248) | 0.428 (4.423) | 6.521** (3.145) | -21.39** (9.379) |
| <i>lnn*lev_max</i> | | -0.0130 (0.0192) | | -0.118*** (0.0392) |
| <i>lnn*lev_var</i> | | -3.642 (2.486) | | 0.595 (4.845) |
| <i>lnk*lev_max</i> | | -0.0336 (0.0206) | | -0.0245 (0.0340) |
| <i>lnk*lev_var</i> | | 7.552*** (2.845) | | 8.173 (7.326) |
| <i>paper fixed effects</i> | No | No | Yes | Yes |
| <i>constant</i> | 0.152*** (0.0254) | 0.123*** (0.0330) | 0.0441 (0.0834) | 0.178* (0.106) |
| <i>N</i> | 279 | 279 | 279 | 279 |
| <i>R²</i> | 0.599 | 0.656 | 0.882 | 0.902 |
| <i>Adjusted R²</i> | 0.592 | 0.645 | 0.811 | 0.840 |
| <i>LOOR²</i> | 0.551 | 0.581 | 0.708 | 0.736 |

Cluster-robust standard errors in parentheses. * $p \leq 10\%$, ** $p \leq 5\%$, *** $p \leq 1\%$.

outliers are somewhat collinear, since *lev_max* and *lev_var* are positively correlated at the 1% level with a correlation coefficient of 0.685 (see Table 3).

Column (3) of Table 4 adds the paper fixed effects to the baseline regression (9). The results are qualitatively similar to column (1), but with notable differences. In particular, the coefficient of *lnn* loses significance, perhaps due to too little variation in sample size (n) within the same paper. However, the coefficient of *nonlinear* (number of nonlinear terms) is now positively significant at the 1% level, as more nonlinear terms increase the model complexity, thus contributing to overfitting.

Column (4) of Table 4 interacts *lnn* and *lnk* with *lev_max* and *lev_var* while keeping the paper fixed effects. The results in column (4) are mostly similar to column (3). However, the coefficient of *lev_var* surprisingly becomes negatively significant at the 5% level with an estimate of -26.05. Nevertheless, the coefficient of *lnk*lev_var* is positively significant at the 5% level with an estimate of 14.75. Overall, since the sample mean of *lnk* is 2.503, the marginal effect of *lev_var* evaluated at the sample mean of *lnk* is $(-26.05 + 2.503 \times 14.75) = 10.87$, which is similar in both magnitude and significance to the estimated coefficient of *lev_var* in columns (1) and (3) without interaction terms. This shows that *lev_var* increases overfitting more in high-dimensional data with a large number of covariates. Moreover, the coefficient of *lnn*lev_max* is negatively significant at the 1% level, implying that the effect of *lev_max* on overfitting could be mitigated by a large sample size.

Table 5 reports regression results for the dependent variable Log(Adjusted EIF). The results in Table 5 largely parallel those in Table 4, and the interpretations are also similar.

In summary, these empirical results show that both $\text{Log}(\text{EIF})$ and $\text{Log}(\text{Adjusted EIF})$ increase with the severity of overfitting as measured by the number of regressors ($\ln k$) and nonlinear terms (*nonlinear*), the maximum value of leverage (lev_max) and its variance (lev_var), but decreases with the sample size ($\ln n$). Moreover, the effects of outliers (lev_max and lev_var) on overfitting could be moderated by the sample size and number of regressors ($\ln n$ and $\ln k$).

4. Monte Carlo simulations

In this section, we conduct Monte Carlo simulations to study the behavior of R^2 , adjusted R^2 , LOOR^2 , EIF, and adjusted EIF as factors related to overfitting change. Overall, the results from simulations are consistent with our findings in the empirical study above.

In the baseline setting, we draw 100 random observations from a bivariate normal

distribution $\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}\right]$. With a correlation coefficient of 0.9 between Y and X , the population R^2 is 0.81. The baseline regression is simply,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (i = 1, \dots, 100). \quad (10)$$

Throughout, we repeat each simulation for 1000 times, and compute the average values of R^2 , adjusted R^2 , LOOR^2 , EIF, and adjusted EIF. We then investigate their behaviors as

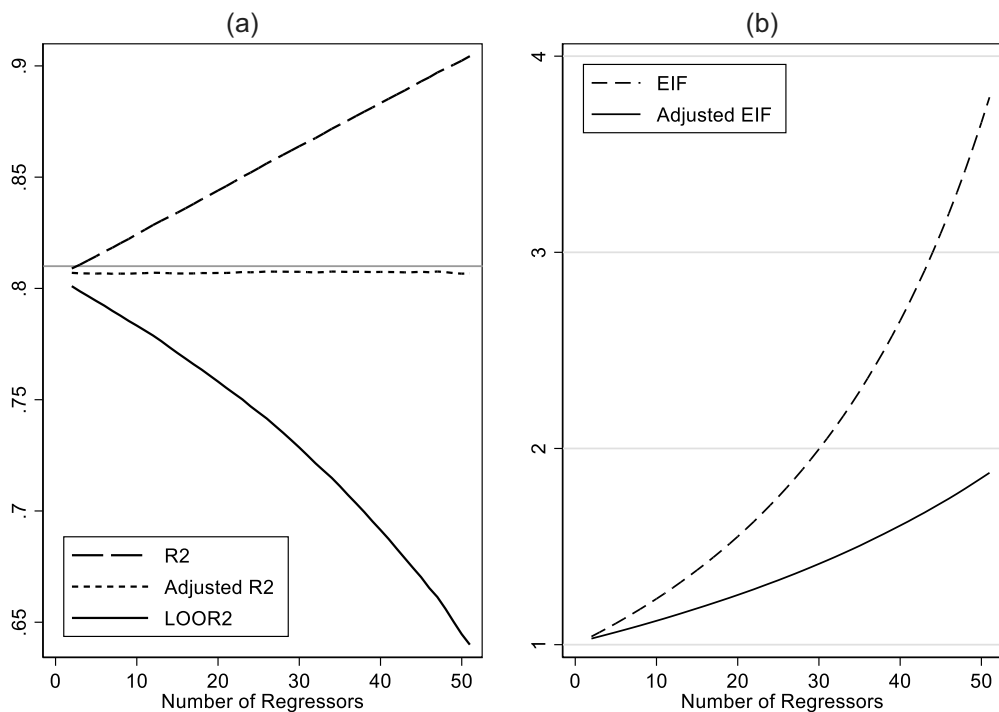


Figure 1. The effects of number of regressors.

factors related to overfitting change, including the number of regressors, the sample size, the number of nonlinear terms, and the presence of outliers.

4.1. Number of regressors

In this simulation, we increase the number of regressors simply by incrementally adding 1–50 noise variables into the baseline regression (10), where all noise variables are independently distributed as $N(0, 1)$. The sample size is kept at 100. The results are presented in Figure 1. Figure 1(a) graphs R^2 , adjusted R^2 and $LOOR^2$ against the number of regressors, where the gray horizontal line shows the population R^2 of 0.81. As the number of regressors increases from 2 to 51, R^2 increases steadily to above 0.9, clearly overestimating the ability of the model to explain the variation in y as a result of overfitting. On the other hand, adjusted R^2 hovers between 0.8 and 0.81, showing the value of degree-of-freedom adjustment. Interesting, $LOOR^2$ actually declines steadily to below 0.65, indicating that adding noise variables actually hurts the model's ability to predict out of sample. Clearly, both R^2 and adjusted R^2 exaggerate the model's true predictive ability, and the extent of exaggeration increases with the number of noise variables added. On the

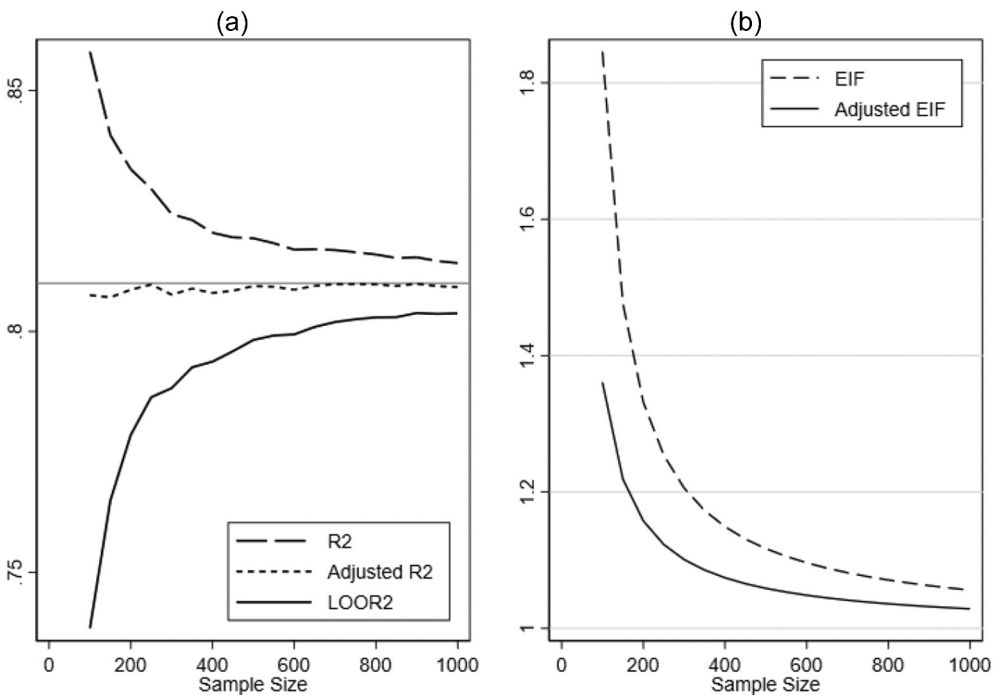


Figure 2. The effects of sample size.

other hand, $LOOR^2$ is robust to overfitting (at least as the model's real predictive ability is concerned), as overfitting resulting from adding noise variables reduces

$LOOR^2$. Figure 1(b) graphs EIF and adjusted EIF against the number of regressors. The interpretation is essentially the same as Figure 1(a).

4.2. Sample size

In this simulation, the sample size is increased from 100 to 1000 at the increment of 50. On the other hand, we keep the number of regressors at 27, including the constant term, the signal variable X , and 25 noise variables independently distributed as $N(0, 1)$. The results are presented in Figure 2. Figure 2(a) graphs R^2 , adjusted R^2 and $LOOR^2$ against the sample size, where the gray horizontal line again shows the population R^2 of 0.81. Apparently, sample size has little effect on adjusted R^2 , which hovers just below 0.81, as it has already compensated for the changing degree of freedom. On the other hand, when the sample size is relatively small (say, $n = 100$), R^2 is clearly above 0.81, indicating that the model is overfit in the presence of 25 noise variables. However, as the sample size increases towards 1000, the overfitting phenomenon diminishes, and R^2 declines towards 0.81 (but still above 0.81). On the contrary, when the sample size is relatively small, $LOOR^2$ is well below 0.81, as the model's predictive ability suffers in the presence of 25 noise variables. As the sample size is increased, $LOOR^2$ climbs up towards 0.81, as a large

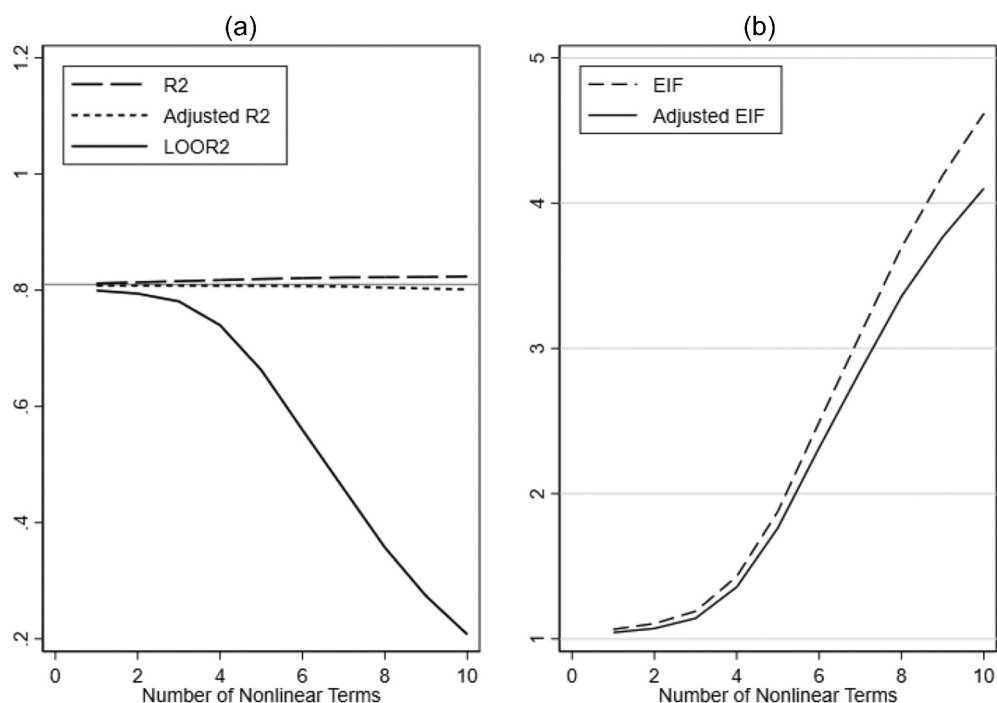


Figure 3. The effects of number of nonlinear terms.

sample size alleviates overfitting. Figure 4(b) graphs EIF and adjusted EIF against the sample size. The interpretation is similar to Figure 2(a).

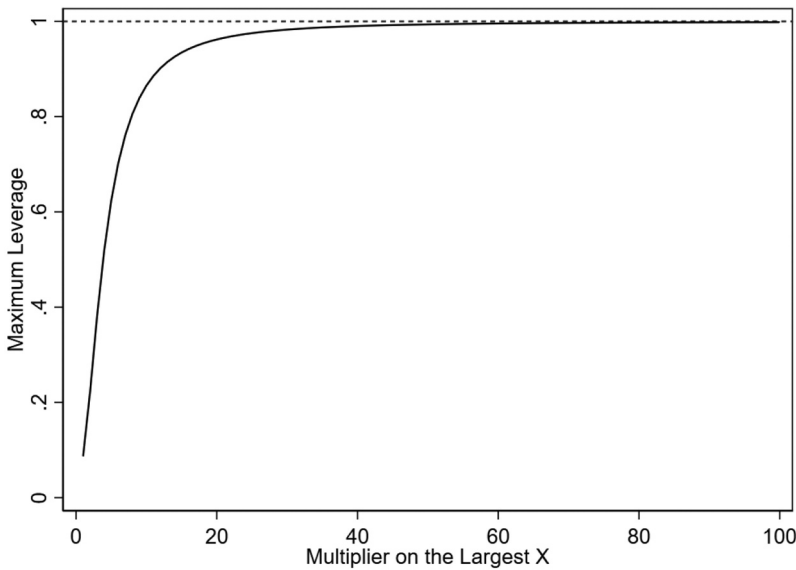


Figure 4. Maximum leverage and multiplier on the largest X.

4.3. Number of nonlinear terms

To consider the effect of nonlinear terms, we simply add second through eleventh power terms to Equation (10),¹¹

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_{11} X_i^{11} + \varepsilon_i \quad (i = 1, \dots, 100). \quad (11)$$

The sample size is still kept at 100. The results are presented in Figure 3. Figure 3(a) graphs R^2 , adjusted R^2 and $LOOR^2$ against the number of nonlinear terms. In this simple data generating process, adding more nonlinear terms does not have much effect on either R^2 or adjusted R^2 , although R^2 does climb up slightly. However, when more nonlinear terms are added, $LOOR^2$ decreases rapidly, as these nonlinear terms drive up the model's complexity, resulting in overfitting and reduced ability to predict out of sample. Figure 3(b) graphs EIF and adjusted EIF against the number of nonlinear terms, and the interpretation is similar.

4.4. Outliers

In this simulation, we generate outliers simply by multiplying the largest value of X in the sample by 2 through 100. As the multiplier on the largest X grows from 1 to 100, the maximum leverage increases rapidly, and approaches its largest possible value of 1, as shown in Figure 4.

Figure 5 presents the simulation results as the maximum leverage increases. Figure 5(a) graphs R^2 , adjusted R^2 and $LOOR^2$ against the maximum leverage. Initially,

¹¹As pointed out by an anonymous referee, adding nonlinear terms can be viewed as a particular case of including additional correlated covariates.

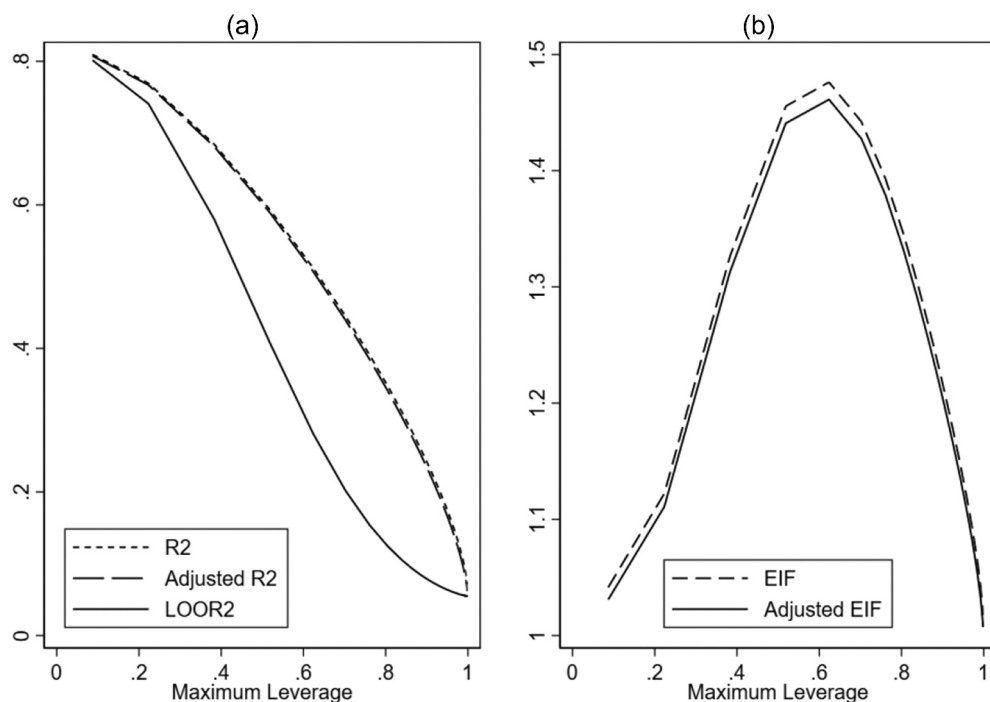


Figure 5. The effect of outliers.

as the maximum leverage grows, $LOOR^2$ drops much faster than R^2 and adjusted R^2 , as the model's true predictive ability declines, while overfitting occurs in the presence of an ever more extreme outlier. However, as $LOOR^2$ drops closer to its lower bound of 0, its speed of declining inevitably falls behind than that of R^2 and adjusted R^2 . In the end, as the multiplier on the largest X increases towards 100, the OLS fit becomes very poor, thus R^2 , adjusted R^2 , and $LOOR^2$ all decline towards their common lower bound of 0. Figure 5(b) graphs EIF and adjusted EIF against the maximum leverage, which tells a similar story. Initially, both EIF and adjusted EIF increase, but they start to decline when the maximum leverage is around 0.5 (and the multiplier on the largest X is 5), resulting in an inverted U-shape.

5. Conclusion

Goodness-of-fit measures R^2 and adjusted R^2 are routinely reported in empirical studies with the implicit presumption that they represent the percentage by which the regressors jointly explain or predict the variation of the dependent variable. This paper shows that R^2 and adjusted R^2 are inaccurate in this regard and often overly optimistic in the presence of overfitting resulting from small sample size, many regressors and nonlinear terms, and existence of outliers. As a remedy, leave-one-out R^2 ($LOOR^2$) can be readily computed, and used as a reliable measure of the model's true ability to predict out of sample.

Moreover, we introduce the concepts of “error inflation factor” (EIF) and “adjusted error inflation factor” (adjusted EIF) as the degree of inflation of test errors $(1 - LOOR^2)$ over training errors represented by $(1 - R^2)$ and $(1 - \bar{R}^2)$ respectively. We then conduct a meta-analysis about the determinants of EIF and adjusted EIF by replicating 273 regressions from 100 papers in four top economics journal during 2004–2021. The median increases of R^2 and adjusted R^2 over $LOOR^2$ reach 40.2% and 21.4%, respectively, in this sample. The regression results show that both EIF and adjusted EIF increase with the severity of overfitting, as measured by the number of regressors and nonlinear terms, and the presence of outliers, but decrease with the sample size. These results are further validated by Monte Carlo simulations.

For empirical researchers, we recommend that they report $LOOR^2$ alongside R^2 and adjusted R^2 , since $LOOR^2$ is robust to overfitting as a measure of the model’s true predictive ability out of sample. Moreover, when $LOOR^2$ diverges from either R^2 or adjusted R^2 , this is a sign of overfitting, and empirical researchers should be concerned, and look for possible causes, such as a complicated functional form (e.g., too many nonlinear terms), and the presence of outliers (e.g., the maximum leverage is close to 1). As a practical matter, while overfitting reduces bias, it usually increases variance to a greater extent, which results in increased mean squared errors of the estimator, and reduced significance of the parameter of interest. Therefore, one way to increase parameter significance is to reduce overfitting.¹²

As model validation via out-of-sample prediction becomes increasingly common in many disciplines, it is time for economists to honestly embrace $LOOR^2$ as a safeguard against overfitting, which is hard to detect by using conventional R^2 and adjusted R^2 based on in-sample fit. In this way, economists can more easily avoid the trap of overfitting, and make their empirical findings more robust. Providers of statistical software (e.g., Stata) can also help in this regard by routinely reporting $LOOR^2$ alongside traditional R^2 and adjusted R^2 in the regression output.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Qiang Chen is a professor at the School of Economics, Shandong University.

Ji Qi is a PhD student at the School of Economics, Shandong University.

References

- Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79. <https://doi.org/10.1214/09-SS054>

¹²We thank an anonymous referee for useful discussions about the relation between overfitting and parameter significance, and more studies are needed in this direction.

- Cochran, W. G. (1968). Commentary on estimation of error rates in discriminant analysis. *Technometrics*, 10(1), 204–205. <https://doi.org/10.1080/00401706.1968.10490548>
- Dower, P. C., Markevich, A., & Weber, S. (2021). The value of a statistical life in a dictatorship: Evidence from Stalin. *European Economic Review*, 133, 103663. <https://doi.org/10.1016/j.euroecorev.2021.103663>
- Efron, B., & Morris, C. (1973). Combining possibly related estimation problems (with discussion). *Journal of the Royal Statistical Society, Series B*, 35, 379–402.
- Geisser, S. (1974). A predictive approach to the random effect model. *Biometrika*, 61(1), 101–107. <https://doi.org/10.1093/biomet/61.1.101>
- Hansen, B. E. (2022). *Econometrics*. Princeton University Press.
- Hills, M. (1966). Allocation rules and their error rates. *Journal of the Royal Statistical Society Series B (Methodological)*, 28(1), 1–31. <https://doi.org/10.1111/j.2517-6161.1966.tb00614.x>
- Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10(1), 1–11. <https://doi.org/10.1080/00401706.1968.10490530>
- Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 22(1), 45–55. <https://doi.org/10.1037/h0072400>
- Mayer, T. (1975). Selecting economic hypotheses by goodness of fit. *The Economic Journal*, 85 (340), 877–883. <https://doi.org/10.2307/2230630>
- Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 2). Addison-Wesley.
- Parady, G., Ory, D., & Walker, J. (2021). The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. *Journal of Choice Modelling*, 38, 100257. <https://doi.org/10.1016/j.jocm.2020.100257>
- Rios-Avila, F. (2018). CV_REGRESS: Stata module to estimate the leave-one-out error for linear regression models. In *Statistical software components*, S458469. Boston College Department of Economics. Retrieved June 11, 2020.
- Stone, M. A. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B (Methodological)*, 36(2), 111–147. <https://doi.org/10.1111/j.2517-6161.1974.tb00994.x>
- Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics*, 2(4), 440–457. <https://doi.org/10.1214/aoms/1177732951>