Vol. 13. Issue.2. 2025 (April-June)

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

http://www.bomsr.com

Email:editorbomsr@gmail.com

INTERNATIONAL STANDARD SERIAL NUMBER 2348-0580

RESEARCH ARTICLE

Behaviour of Statistical Measures of Goodness of Fit Subject to Variation in Sample Size and Number of Regressors

Nirmali Borah
Assistant Professor, Dept. of Statistics
Paschim Guwahati Mahavidyalaya, Guwahati, Assam

DOI:10.33329/bomsr.13.2.31

Article Info
Article Received: 12/04/2025
Article Accepted: 29/05/2025
Published online: 26/06/2025

Abstract

R^2 (Adj. R^2) is the most widely pursued statistical measure in judging the goodness of the fit of the model as well as in explaining the total variation of the explained variable via explanatory variables, especially in applied scientific fields. As regards the application of R^2 (Adj. R^2) there is a variation of its interpretation across linear models with and without Intercept along with equally spaced and unequally spaced sample observations. In this paper, an attempt has been made to see the behaviour of R^2 (Adj. R^2) on varying sample sizes from small to moderate. What has been observed in this paper is that R^2 is relatively large for a small sample compared to a moderate (and large) sample. The results obtained are found to corroborate those of the prominent work by Gerald J. Hahn and Others. In addition, this paper aims to study the behaviour of R^2 (Adj. R^2) on varying numbers of regressors. The inferences are that R^2 (Adj.R^2) is an increasing function of regressors.
Keywords: Coefficient of determination, adjusted regression coefficient of determination, Regression analysis.

1.Introduction

In regression analysis, goodness-of-fit measures assess the extent to which the explanatory variable explains the variation in the explained variable. These measures are critical for model selection, comparison, and validation. However, if both the sample size and the number of regressors change, these measures can behave differently, which leads to varying interpretations of model quality. Data analysts often rely on the coefficient of determination (R^2) to evaluate how well a model fits a given dataset. Working with non-linear models with an intercept using the wrong R^2 statistic can lead to misleading conclusions

The root of the problem lies in two key challenges. First, there are various expressions of R^2 in statistical literature. While they produce similar results for linear regression with an intercept and they can vary significantly for nonlinear models or models without an intercept. Second, many statisticians are unaware of how widespread this issue is.

The selection of R^2 requires to be based on the types of models being formulated, the model fitting technique used, the purpose for which R^2 is used and the desirable properties of R^2. It is significantly dependent on the number of observations in a sample as well as the number of Regressors. R^2 with a large sample size is different from a small sample size dataset. Understanding these dynamics is crucial for accurate model evaluation. There appear a lot of discrepancies and debate in statistical literature as regards the use of R^2. Samples are as under.

The literature on R^2 and its limitations sheds light on several important perceptions of how this statistic can mislead researchers, especially in complex models with varying sample sizes and predictors. Gao (2023) highlights that R^2 tends to overestimate the proportion of variance explained (PVE) as sample size increases, which could lead to incorrect conclusions about a model's effectiveness. Li et al. (2020) discuss how Adjusted R^2 (adj. R^2) attempts to address this by accounting for sample size and the number of regressors. Onyutha (2020) points out that the inclusion of multiple predictors can artificially inflate R^2 by giving the false impression that a model fits the data well, even if it does not. This can result in models that seem effective but are actually inadequate.

Rights and Sterba (2022) further complicate the picture, suggesting that in multilevel models, the choice of centering strategy can impact R^2 calculations, making it difficult to accurately interpret the explained variance at different levels. This highlights the challenges in interpreting R^2, particularly in more complex models. Despite R^2's widespread use, its limitations are clearly visible. Onyutha (2020) suggests that alternative metrics like the coefficient of model accuracy (CMA) could offer more reliable evaluations of model performance.

Sangeeta Dev (2016) adds another layer of complexity by pointing out that R^2 behaves unpredictably when data is contaminated, further complicating its utility as a universal measure for assessing the performance of linear regression models. Jeremy, N., V., and Miles (2014) confirm that while R^2 measures the proportion of variance explained by predictors, it is highly sensitive to both sample size (n) and the number of regressors (k).

In conclusion, though R^2 is a commonly used statistic, it is clear that it has several limitations that can lead to misinterpretation.

2. Methodology:

The methodology and methods used here are simple linear regression and multiple linear regression with or without an intercept. They go as:

$$\left.\begin{aligned} E((Y/X)) &= \alpha + \beta X \\ E((Y/X)) &= \beta X \end{aligned}\right\} \quad (2.1)$$

And

$$\left.\begin{aligned} E(Y/X) &= \beta_0 + \sum_{i=1}^{k} \beta_i X_i \\ E(Y/X) &= \sum_{i=1}^{k} \beta_i X_i \end{aligned}\right\} \quad (2.2)$$

Where Y stands for the Explained variable and X's stand for explanatory variables

The least squares estimators are

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad \rightarrow (2.3)$$

$$D(\hat{\beta}) = (X'X)^{-1}\sigma^2$$

Where X = (x_{ij}) ; i=1, 2, …….., n

= regression matrix j= 0, 1, 2, ……., k;

$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots\ldots, \hat{\beta}_k)^T$

$\underset{\sim}{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots\ldots, \varepsilon_n)^T$

And X is a **n x k+1** matrix of rank k+1

2.1. R-squared (R^2):

R-squared is a statistical measure that is used to explain the variability of the dependent variables or regressand in a regression model with respect to the variation in the independent variable or regressor. It is the proportion of variance in the dependent variable that is predictable from the independent variables.

The value of $\mathbf{R^2}$ ranges from 0 to 1. $\mathbf{R^2}$ tends to 1 implies that the model explains a large portion of the variability, while a value close to 0 means the model explains very little variation.

Expressions of R^2:

$$\boldsymbol{R^2 = 1 - \frac{SSR}{SST}} \quad \textbf{or}$$

$$\boldsymbol{R^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where SSR is the sum of squared residuals (errors) and SST is the total sum of squares, which measures the total variance of the dependent variable.

Adjusted R-squared (Adj. R^2): Adjusted R-squared adjusts the $\mathbf{R^2}$ value depending on the number of regressors in the model. It is useful because $\mathbf{R^2}$ will always increase when more variables are added, even if they don't significantly improve the model. Adjusted R-squared penalizes the model for including unnecessary predictors and prevents overfitting.

$$AdjR^2 = 1 - \left(\frac{1-R^2}{n-p-1}\right) \cdot (n-1)$$

Where, n is the number of observations,

and p is the number of regressors.

A higher adjusted R-squared value implies a better model fit without considering the unnecessary regressors.

Mean Squared Error (MSE): MSE measures the average squared difference between the observed actual outcomes and the outcomes predicted by the model. It gives us an idea of how well the model predicts the dependent variable.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where y_i is the observed value and $\hat{y}_i$is the predicted value.

Lower MSE indicates that the model's predictions are closer to the actual values. A higher MSE indicates poor model performance.

Akaike Information Criterion (AIC): AIC is a measure of the quality of a statistical model, balancing between goodness-of-fit and model complexity. It helps in model selection, comparing multiple models, and choosing the one that best explains the data without overfitting

$$AIC = 2k - 2\,ln(L)$$

Where:

k is the number of model parameters (predictors/regressors),

L is the likelihood of the model (how likely the observed data is given the model).

A lower AIC value indicates a better model, as it suggests that the model fits the data well without including unnecessary parameters.

Bayesian Information Criterion (BIC): BIC is similar to AIC but adds a stronger penalty for having more parameters. It is also used for model selection, comparing models to find the one that best balances fit and complexity.

$$BIC = ln(n) \cdot k - 2\,ln(L)$$

Where,

n is the number of data points,

k is the number of parameters,

L is the likelihood.

A lower BIC value indicates a better model. However, BIC tends to favour simpler models compared to AIC, especially when the sample size is large.

3. Results and Analysis:

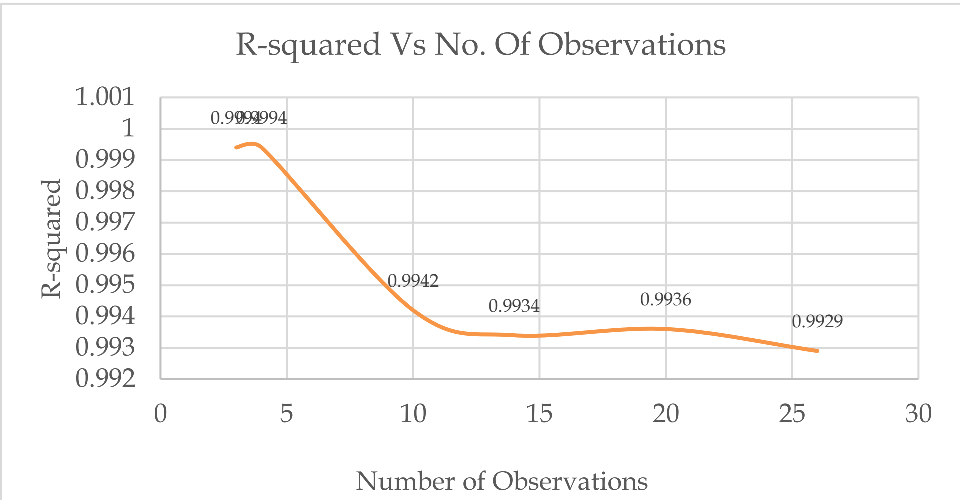
Models (2.1) and (2.2) have been run for small to moderate sample sizes. The various statistical measures such as MSE, AIC, BIC, R^2, Adj. R^2 has been computed. The results are presented below.

3.1. Model: Simple linear regression (with intercept)

3.1.1: The following Table shows the values of MSE, AIC, BIC, R^2, Adj. R^2 calculated by using Python Language as per increasing number of observations for simple linear regression with intercept.

No. of Obsns.	**3**	**4**	**10**	**14**	**20**	**26**
R-squared	0.9994	0.9994	0.9942	0.9934	0.9936	0.9929
Adjusted R-squared	0.9991	**0.9992**	0.9936	0.9928	0.9932	0.9926
MSE	3541.8013	3261.8399	5470.1543	4451.9707	4435.6024	4815.2091
AIC	17.9294	22.7479	77.6942	110.1411	156.3745	206.5087
BIC	17.0280	22.1343	77.9968	110.7802	157.3703	207.7668
Significance at 1% level	Significant	Significant	Significant	Significant	Significant	Significant
Significance at 5% level	Significant	Significant	Significant	Significant	Significant	Significant

Data courtesy: Example taken from of the book' Introduction to Linear Regression Analysis' by Montgomery et.al, Pp. 61

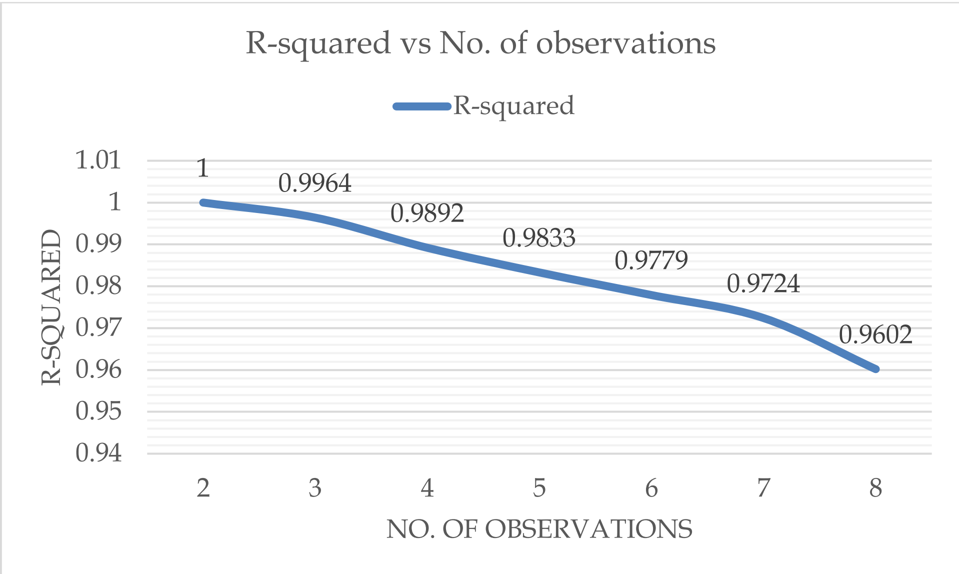


Graph 3.1.1: R-squared vs no. of observations

3.1.2: The following Table shows the values of MSE, AIC, BIC, R^2, Adj. R^2 calculated by using Python Language as per increasing number of observations.

No. of Obsns.	02	03	04	05	06	07	08
R-squared	1.0000	0.9964	0.9892	0.9833	0.9779	0.9724	0.9602
Adjusted R-squared	0.9891	0.9929	0.9838	0.9777	0.9724	0.9669	0.9535
MSE	0.0000	0.0000	0.0002	0.0004	0.0007	0.0011	0.0017
AIC	-134.498	-17.3999	-18.4007	-20.4008	-22.367	-24.0875	-24.3732
BIC	-137.112	-19.2027	-19.6281	-21.1819	-22.7835	-24.1957	-24.2143
Significance at 1% level	Not Significant	Not Significant	Significant	Significant	Significant	Significant	Significant
Significance at 5% level	Not Significant	Significant	Significant	Significant	Significant	Significant	Significant

Data Courtesy: Example 2.15 of the book' Introduction to Linear Regression Analysis' by Montgomery et.al, Pp 63



Graph 3.1.2: R-squared vs no. of observations

What has been revealed from the table and the graph is that R-Squared values decrease as sample size increases, including high R2 for smaller sample size(n). In other words, R2 decreases as 'n' increases. It is also revealed that the statistical measures MSE, AIC, and BIC are much smaller for a small sample size as compared to a moderately large sample.

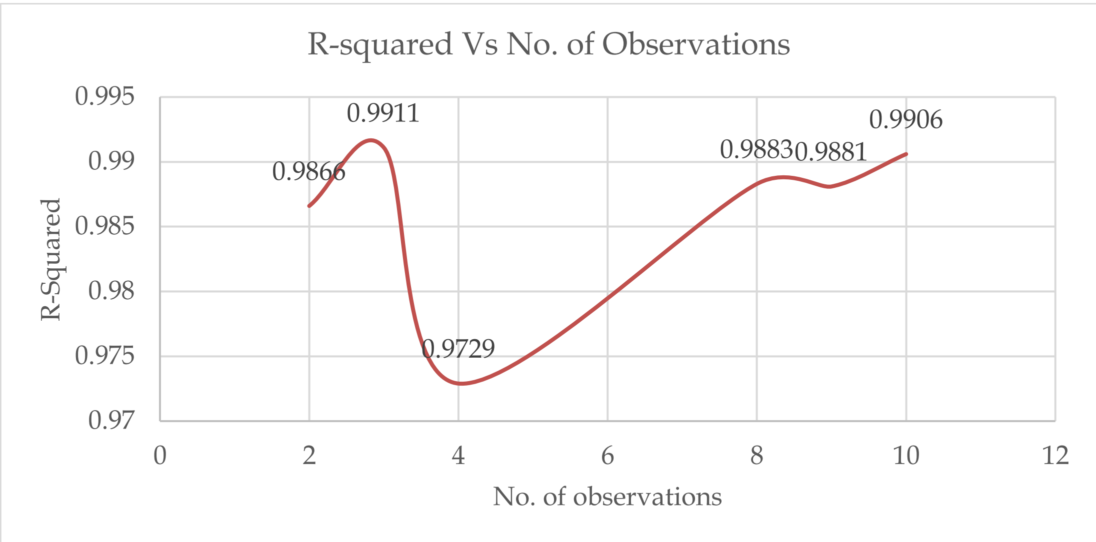
3.2: Simple linear regression (equally spaced observations)

Regression results for equally spaced data are as follows:

3.2.1: The following Table shows the values of MSE, AIC, BIC, R^2, Adj. R^2 calculated by using Python Language as per increasing number of observations for equally spaced observations.

No. of Obsn	**02**	**03**	**04**	**08**	**09**	**10**
R-squared	0.9866	0.9911	0.9729	0.9883	0.9881	0.9906
Adjusted R-squared	0.9867	0.9867	0.9639	0.9866	0.9866	0.9895
MSE	1063.8432	1618.1256	3389.3666	7071.4964	8860.3264	9996.9310
AIC	16.2375	22.5843	33.9233	63.9818	72.7073	79.5355
BIC	14.9307	33.3096	33.3096	64.0612	72.9045	79.8381
Significance at 1% level	Not Significant	Significant	Significant	Significant	Significant	Significant
Significance at 5% level	Not Significant	Significant	Significant	Significant	Significant	Significant

Data Courtesy: Table 2.5 of 'Basic Econometrics' by Damodor and Sangeeta, Pp49

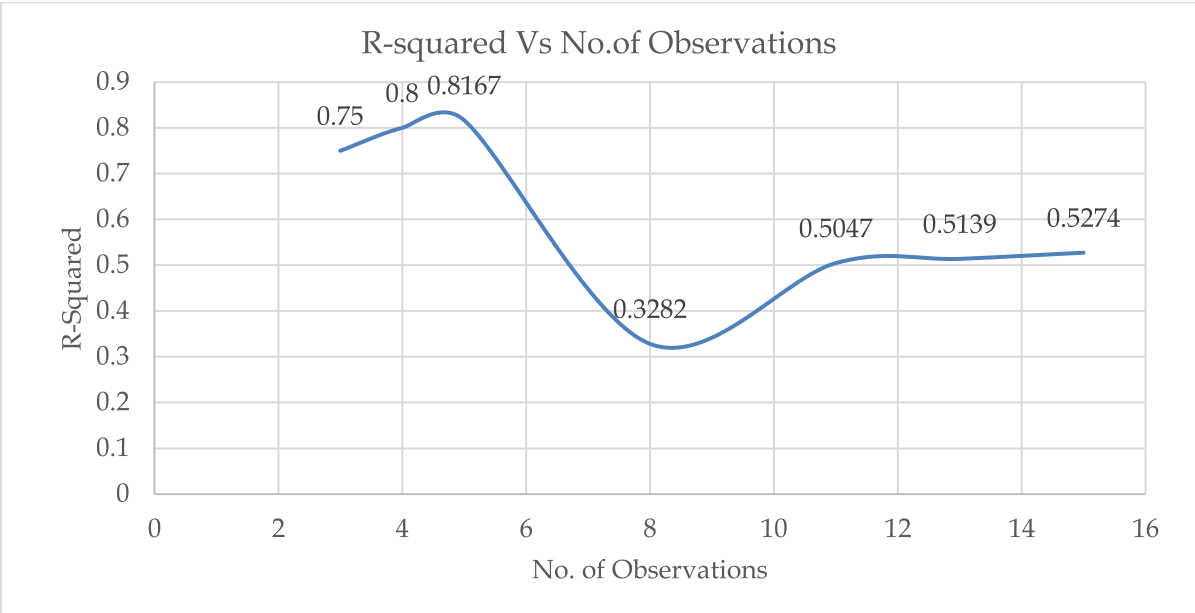


Graph 3.2.1: R-squared Vs no. of observations

3.2.2: The following Table shows the values of MSE, AIC, BIC, R^2, Adj. R^2 calculated by using Python Language as per increasing number of observations for equally spaced observations.

No. of Obsns.	**03**	**04**	**05**	**8**	**11**	**13**	**15**
R-squared	**0.7500**	0.8000	**0.8167**	0.3282	**0.5047**	**0.5139**	**0.5274**
Adjusted R-squared	**0.5000**	0.7000	**0.7556**	0.2162	**0.4496**	**0.4697**	**0.4910**
MSE	**0.0006**	0.0005	**0.0022**	0.0084	**0.0081**	**0.0074**	**0.0066**
AIC	**-9.9730**	-15.0521	**-12.4071**	-11.5356	**-17.7506**	**-22.9527**	**-28.7515**
BIC	**-11.7758**	-16.2795	**-13.1882**	-11.3767	**-16.9548**	**-21.8228**	**-27.3354**
Significance at 1% level	Not Significant	Not Significant	Not Significant	Not Significant	Not Significant	Significant	Significant
Significance at 5% level	Not Significant	Not Significant	Significant	Not Significant	Significant	Significant	Significant

Data Courtesy: Example J of the book'Applied Regression Analysis' by Draper & Smith, Pp 101



Graph 3.2.2: R-squared Vs no. of observations

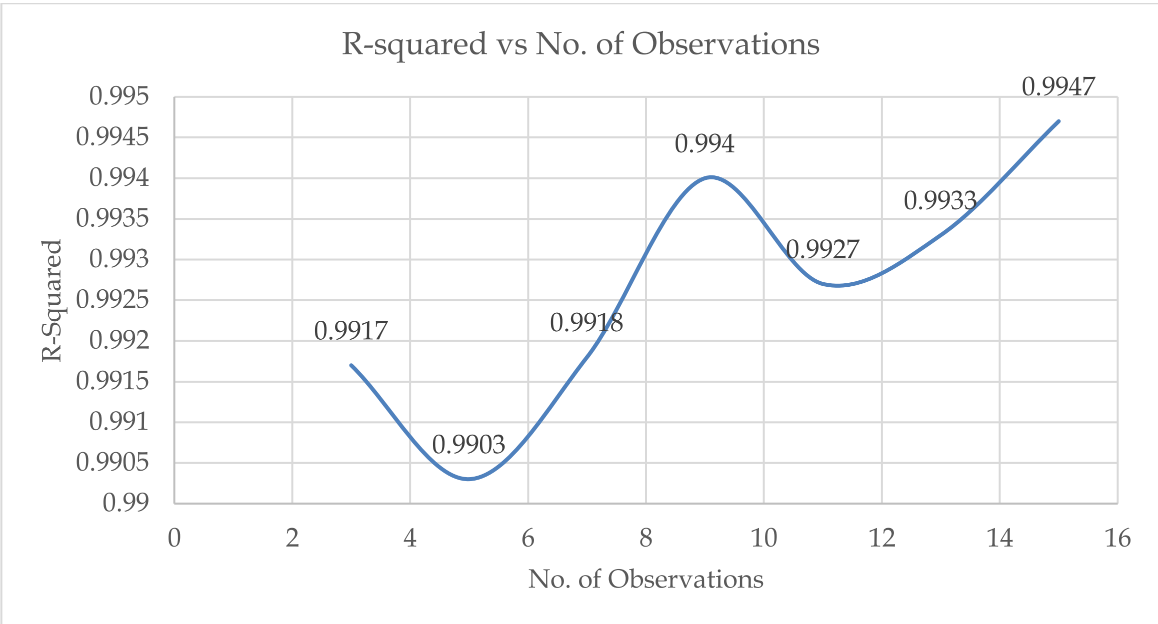
The table above indicates that R2 values oscillate as sample sizes increase from small to moderate. Its behaviour appears to contradict what is revealed for R2 with the intercept model for unequal observations. Under this model, MSE, AIC, BIC have been observed to be lowest for small size as compared to moderate sample size.

3.3: Simple linear regression (without intercept)

3.3.1. The following Table shows the values of MSE, AIC, BIC, R2, Adj. R2 calculated by using Python Language as per the increasing number of observations for a simple linear regression model without an intercept.

No. of Obsns.	**03**	**05**	**07**	**09**	**11**	**13**	**15**
R-squared	0.9917	0.9903	0.9918	0.9940	0.9927	0.9933	0.9947
Adjusted R-squared	0.9834	0.9871	0.9901	0.9931	0.9919	0.9927	0.9943
MSE	0.0945	0.1158	0.0902	0.0939	0.0940	0.0892	0.0807
AIC	5.4377	7.4097	7.0260	8.2502	9.2060	9.4759	8.8120
BIC	3.6349	6.6286	6.9178	8.6446	10.0018	10.6058	10.2281
Significance at 1% level	Not Significant	Significant	Significant	Significant	Significant	Significant	Significant
Significance at 5% level	Not Significant	Significant	Significant	Significant	Significant	Significant	Significant

Data Courtesy: Example 2.8 of the book' Introduction to Linear Regression Analysis' by Montgomery et.al, Pp 48



Graph 3.3.1: R-squared vs no. of observations

What has been laid bare is that R2 increases as sample size increases which is completely in opposition to what has been for R2 for the intercept model. While the MSE, AIC and BIC values are observed to be lower for small sample size as compared to moderate sample size. The MSE, AIC, and BIC are found to be marginally large for a moderate sample.

3.4: Linear Regressions (upon increasing the number of regressors)

Linear regressions have been run for various sizes of regressors. The results are underlined as follows:

3.4.1: The following Table shows the values of MSE, AIC, BIC, R^2, Adj. R^2 calculated by using the Python Language as per the increasing number of regressors for linear regression model.

No. of Regressor	**1**	**2**	**3**	**4**	**5**
R-squared	0.0442	0.4153	0.6013	0.6722	0.8228
Adjusted R-squared	-0.0240	0.3254	0.5016	0.5530	0.7341
MSE	1327450.7482	812045.1944	553804.8250	455255.4269	246176.4737
AIC	274.9864	269.1230	264.9991	263.8639	256.0269
BIC	276.5315	271.4408	268.0895	267.7268	260.6624
Significance at 1% level	Not Significant	Not Significant	Significant	Not Significant	Significant
Significance at 5% level	Not Significant	Significant	Significant	Significant	Significant

Data Courtesy: Table 8.10 of 'Basic Econometrics' by Damodar and Sangeeta, Pp 297

Graph 3.4.1: R-squared vs no. of Regressors

3.4.2: The following Table shows the values of MSE, AIC, BIC, R^2, Adj. R^2 calculated by using Python Language as per the increasing number of regressors for linear regression model.

No. of Regressor	**1**	**2**	**3**	**4**
R-squared	0.5336	0.7005	0.7048	0.7923
Adjusted R-squared	0.4978	0.6506	0.6243	0.7092
MSE	1457421.1575	935834.5016	922443.8041	649081.9642
AIC	259.4508	254.8061	256.5899	253.3179
BIC	260.8669	256.9302	259.4221	256.8581
Significance at 1% level	Significant	Significant	Significant	Significant
Significance at 5% level	Significant	Significant	Significant	Significant

Data Courtesy: Table 7.6 of 'Basic Econometrics' by Damodor and Sangeeta, Pp241

Graph 3.4.2: R-squared vs no. of Regressors

The table and the graph have brought about higher R^2 for the higher number of regressors, which, in turn, establishes the theoretical justification of non-decreasing R^2 as a result of non-increasing Residual Sum of Squares. The model fitting using a higher number of regressors is warranted by lesser MSE, AIC, and BIC as compared to regressions with a lesser number of regressors. The same picture is revealed in both examples.

4. Conclusion

The present study undertook an inclusive investigation into the behaviour of the coefficient of determination (R^2) and its adjusted form (Adjusted R^2) under varying sample size and number of regressors in linear regression models, both with and without intercepts.

The empirical results reveal that R^2 tends to show relatively higher values when applied to small sample sizes, which may create an overestimation of model performance. As the sample size increases, R^2 generally shows a declining trend, thereby presenting a more conservative and possibly more realistic measure of model fit. These findings are consistent with those of Hahn et al. and other prior studies.

Furthermore, it is observed that both R^2 and Adjusted R^2 increase with the number of regressors, which is theoretically consistent with the non-increasing nature of the Residual Sum of Squares (RSS). In this context, Adjusted R^2 serves as a more reliable problem-solving tool as it accounts for model complexity by penalizing the inclusion of additional predictors.

The study also observed corresponding statistical measures such as Mean Squared Error (MSE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). Notably, models with smaller sample sizes yielded lower values of MSE, AIC, and BIC.

In light of these findings, it is evident that while R^2 and Adjusted R^2 are indispensable tools in regression analysis, they are not free from limitations. Their application must be sensitive in models involving small sample sizes or a large number of regressors.

In conclusion, researchers are advised to critically assess the suitability of R^2 in their specific modelling contexts and to adopt additional statistics and model validation techniques. Only through such a comprehensive approach can one ensure robust inference and prevent the misinterpretation of statistical results.

References

[1] Gao, J. (2023). R-Squared (R^2) – How much variation is explained? *Journal of Statistical Analysis*, 10(2), 1-12. https://doi.org/10.1177/26320843231186398

[2] Rights, J. D., & Sterba, S. K. (2022). R-squared measures for multilevel models with three or more levels. *Multivariate Behavioral Research*, 58(2), 340-367. https://doi.org/10.1080/00273171.2021.1985948

[3] Onyutha, C. (2020). From R-squared to coefficient of model accuracy for assessing "goodness-of-fits". *Geoscientific Model Development Discussions*, 13(2), 1-25. https://doi.org/10.5194/GMD-2020-51

[4] Zaarour, N., & Melachrinoudis, E. (2019). What's in a coefficient? The "not so simple" interpretation of R^2 for relatively small sample sizes. *Journal of Education and Training Studies*, 7(12), 27. https://doi.org/10.11114/jets.v7i12.4492Mendeley

[5] Deb, S. (2016). A novel robust R-squared measure and its applications in linear regression. In *Advances in Statistical Research* (pp. 131-142). Springer. https://doi.org/10.1007/978-3-319-48517-1_12

[6] Miles, J. N. V., & V., Jeremy. (2014). R Squared, Adjusted R Squared. In *Encyclopedia of Statistics in Behavioral Science* (pp. 189-192). Wiley. https://doi.org/10.1002/9781118445112.STAT06627

[7] Li, J., Smith, A., & Lee, S. (2000). More accurate estimators of multiple correlation coefficients? *Statistics in Practice*, 12(3), 45-60. https://doi.org/10.35566/isdsa2019c11

[8] Seber, G.A.F.& Lee, A.J (2003). Linear Regression Analysis. *Wiley Series in Probability and Statistics, Section Edition*

[9] Montgomery, D.C., Peck, E.A., & Vining, G.G. (2003). Introduction to Linear Regression Analysis. *3rd ed., John Wiley & Sons.*

[10] Draper, N.R. & Smith, H. (2019). Applied Regression Analysis. *Wiley Series in Probability and Statistics, 3rd Edition*

[11] Gujarati, D.N., Porter, D.C., Gunasekar, S. (2009). Basic Econometrics. *Fifth Edition, The McGraw Hill Companies*