Model Selection Methods: Statistical Theory and Monte Carlo Analysis

Eliyahu B. Addess

Johns Hopkins University

625.725 Theory of Statistics

Dr. Tom Woolf and Dr. Burhan Sadiq

December 2022

Abstract

The problem of model selection impacts researchers across scientific disciplines. Given a set of statistical models for a phenomena, how can data be used to determine the best model for prediction? The field of model selection seeks to answer this question.

Using statistical theory, researchers have developed numerous methods of model selection. The majority of these methods ultimately attempt to indicate a model which provides a good fit to the data but simultaneously does not suffer from over-fitting. However, the theoretical underpinnings of different model selection methods vary quite distinctly. Moreover, it can be shown that certain methods are more effective than others at indicating an accurate statistical model.

In this paper we compare several popular model selection procedures. The methods under investigation are R^2, Adjusted R^2, Akaike Information Criteria (AIC), and Bayesian Information Criteria (BIC). Following the approach of Sharma et al., 2019 in the assigned paper, we explore these criteria from both a theoretical and computational perspective. To this end, in addition to mathematical research, we performed several Monte Carlo experiments. These experiments use simulated data to compare and contrast the accuracy of the four model selection methods. Our results indicate that AIC and BIC are significantly more effective at identifying a true model than R^2 and Adjusted R^2. Moreover, BIC performs better than AIC in the event that the true model is included in the testing set. These findings support the contention of Sharma et al., 2019 that researchers should use the AIC or BIC methods for model selection, rather than the more classic R^2 and Adjusted R^2.

Model Selection Methods: Statistical Theory and Monte Carlo Analysis

Contents

Introduction

In this paper we explore the problem of statistical model selection and analyze several model selection methods in detail. The specific methods under investigation are R^2, Adjusted R^2, Akaike Information Criteria (AIC), and Bayesian Information Criteria (BIC). The paper is divided into two parts. In the first part we introduce the theoretical background of each model selection criterion. In the second part we analyze the accuracy of the criteria via Monte Carlo simulation. This project is founded on my assigned paper, Sharma et al., 2019. However, I have consulted multiple sources in order to gain a deeper understanding of the statistical theory. These books and articles will be referenced throughout the paper. For the Monte Carlo experiments, I have drawn on the ideas of Sharma et al., 2019, Raffalovich et al., 2008, and Subrahmanya, 2018.

Part 1. Theoretical Background of Model Selection Criteria

R^2 and Adjusted R^2

We begin our exploration of model selection criteria with the classical "Coefficient of Determination"(Montgomery et al., 2020), R^2. Although we will introduce R^2 in the context of simple linear regression, this definition can easily be extended to multiple linear regression. The following introduction is based on Montgomery et al., 2020.

Consider the simple linear regression model, $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, $E(Y|X = x) = \beta_0 + \beta_1 x$, and $V(Y|X = x) = \sigma^2$. Given an observed set of (x_i, y_i), we can use least squares or maximum likelihood to estimate the model parameters β_0 and β_1. We can then write the estimated $E(Y|X = x_i)$ as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{1}$$

Define the i^{th} residual as:

$$e_i = y_i - \hat{y}_i \tag{2}$$

The i^{th} residual is the difference between the observed value of Y and the fitted value.

Now we can define three sums of squares, SST, SS_{Res}, and SS_{R}. SST is the total sum of squares and is defined as

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{3}$$

where $\overline{y}$ is the sample average. SST represents the total variation in the response variable from its mean. SS_{Res} is the residual sum of squares and is defined as

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

SS_{Res} represents the variation in the observed responses from their expected value as predicted by the model. Finally, define the regression sum of squares SS_{R} as

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \tag{5}$$

The quantity $(\hat{y}_i - \overline{y})$ is the deviation of the i^{th} fitted value from the mean of the observations. This deviation can be explained by the regression with X.

It can be shown that

$$SST = SS_R + SS_{Res} \tag{6}$$

which is known as "the fundamental analysis-of-variance identity for a regression model"(Montgomery et al., 2020). Intuitively, this identity says that the total variation in the observed y_i is the sum of the variation due to the regression with the x_i (SS_R) and the random error (SS_{Res}).

The coefficient of determination[1],R^2, can be now be defined as

$$R^2 = 1 - \frac{SS_{Res}}{SST} = \frac{SS_R}{SST} \tag{7}$$

This ratio represents the proportion of the total variation in Y which is explained by the regression with X. From (6) we see that $SS_R \leq SST$ and thus $0 \leq R^2 \leq 1$. Values of R^2

[1] See Larsen and Marx, 2018 who explain that the coefficient of determination is known as R^2 because it is the square of the sample correlation coefficient R.

closer to 1 indicate that a greater proportion of the variation in the response variable is explained by the regression model.

As its title suggests, R^2 has been used as a method of model selection. Given several models, greater R^2 values indicate a better fit to the data, hence intuitively one would select the model with the highest R^2. Sharma et al., 2019 explain that this practice is problematic because it will invariably lead to overfitting the data. As terms are added to a model, it will fit the data points more precisely, leading to smaller residuals and a larger R^2. Thus from a set of potential models, R^2 will always indicate the most complex model. However, overfit models tend to provide poor inference and predictions since they often contain variables that are not truly related to the data generating process (Burnham & Anderson, 1998) and are fitted to the noise rather than the signal in the data (Sharma et al., 2019).

To address the shortcomings of R^2 an adjusted R^2 has been defined as

$$R^2_{Adj,p} = 1 - \left(\frac{n-1}{n-p}\right)\left(\frac{SS_{Res}}{SST}\right) \tag{8}$$

where p is the number of terms in the model. The coefficient $\left(\frac{n-1}{n-p}\right)$ increases with the complexity of the model which results in smaller $R^2_{Adj,p}$. Thus a penalty is incurred to models with unnecessarily many terms. This is a significant improvement over R^2 as a model selection criteria, however Sharma et al., 2019 note that adjusted R^2 "is not based on rigorous statistical theory" and hence its value is dubious. Shalizi, 2015 discusses additional issues with R^2 and states "Using adjusted R^2 instead of R^2 does absolutely nothing to fix any of this"[2]. We seek a more rigorous solution.

The Kullback-Leibler Information

In his 1973 paper (reproduced in deLeeuw, 1992), Hirotugu Akaike derived a model selection method based on the Kullback-Leibler information (also known as K-L distance or K-L divergence). Our discussion of the K-L information and foundations of AIC is based

[2] See Ford, 2015 for an interesting Monte Carlo demonstration of Shalizi's theoretical concerns.

on Burnham and Anderson, 1998 and Konishi and Kitagawa, 2008. We use the notation of Konishi and Kitagawa, 2008.

The K-L information is a measure of the discrepancy between two statistical models."The K-L distance between models is a *fundamental quantity* in science and information theory ..." (Burnham & Anderson, 1998). The K-L information can be defined as follows. Let $G(x)$ be the true data generating distribution of the random variable X and let $F(X)$ be an approximating model. Define the K-L distance from F to G as

$$I(G;F) = E_G\left[\log\left\{\frac{G(X)}{F(X)}\right\}\right] \tag{9}$$

Thus the K-L information is the expected value of $\log\left\{\frac{G(X)}{F(X)}\right\}$ with respect to the true distribution G. For continuous distributions $g(x)$ and $f(x)$ we have

$$I(g;f) = \int_{\infty}^{\infty} \log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx \tag{10}$$

The definition of the K-L information has some elegant properties, namely $I(g;f) \geq 0$ and $I(g;f) = 0$ if and only if $g(x) = f(x)$. Given approximating distributions f_1 and f_2 with $I(g;f_1) < I(g;f_2)$, these properties allow us to view f_1 as closer to the true distribution than f_2 is.

It is important to note that in general $I(g;f) \neq I(f;g)$. Intuitively we should not expect these two quantities to be equal. $I(g;f)$ assumes that g is the true model and represents the "information lost" (Burnham & Anderson, 1998) when f approximates g. Hence we take the expectation with respect to g. Conversely, $I(f;g)$ assumes that f is the true model and g the approximation. Consequently, $I(g;f)$ has nothing to do with $I(f;g)$. As Burnham and Anderson, 1998 put it, "*nor should they be equal, because the roles of truth and model are not interchangeable.*"

It would seem that the K-L information is an ideal criteria for model selection. The model with the smallest $I(g;f)$ is closest to the true data generating model g and as such is the best model. The obvious problem, however, is that in real world situations it is impossible to compute $I(g;f)$. This is because the K-L information is an expected value

with respect to the true model g, *which we do not know*. The question becomes, given several competing models, can we use data to *estimate* the K-L distances?

Foundations of AIC

When developing an estimator of $I(g; f)$, a key observation is that the K-L information can be split into two terms.

$$I(g; f) = E_g\left[\log\left\{\frac{g(X)}{f(X)}\right\}\right] = E_g[\log g(X) - \log f(X)] \tag{11}$$

$$= E_g[\log g(X)] - E_g[\log f(X)] \tag{12}$$

Notice that the first term in (12) depends only on the true distribution g. Since g is unknown, we have no way of estimating this term. However, since the first term of (12) depends only on g, it will be constant across all possible approximating models. Hence when comparing the K-L distances of different approximating models, it suffices to consider just the second term in (12), which is

$$E_g[\log f(X)] \tag{13}$$

For a model f, (13) is known as "the expected log-likelihood"(Konishi & Kitagawa, 2008). Greater values of (13) will yield smaller K-L distances, implying a better model.

Although we cannot compute the true value of $E_g[\log f(X)]$ (because it is an expectation with respect to the unknown distribution g), we *can* obtain an estimate using the empirical distribution. For IID data points $(x_1, \ldots, x_n)$, the empirical distribution, $\hat{g}(x)$ is a discrete probability distribution that assigns probability $\frac{1}{n}$ to each data point[3]. That is $\hat{g}(x_i) = \frac{1}{n}$. If we replace g with $\hat{g}$ in (13), we have

$$E_{\hat{g}}[\log f(X)] = \sum_{i=1}^{n} \hat{g}(x_i) \log f(x_i) \tag{14}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \log f(x_i) \tag{15}$$

[3] See the course text, Wasserman, 2004, section 7.1 for a formal definition of the empirical cumulative distribution function.

By the law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n}\log f(x_i) \xrightarrow{p} E_g[\log f(X)]$$

Hence (15) is a consistent estimator for $E_g[\log f(X)]$.

Note that if we multiply the estimator $\left[\frac{1}{n}\sum_{i=1}^{n}\log f(x_i)\right]$ by n, we obtain $\sum_{i=1}^{n}\log f(x_i)$. For a parametric model $f(x|\boldsymbol{\theta})$ this gives us the well known log-likelihood[4],

$$\ell(\boldsymbol{\theta}) = \log\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) = \sum_{i=1}^{n}\log f(x_i|\boldsymbol{\theta}) \tag{16}$$

Here Akaike showed a remarkable connection between an estimator for the relative K-L information and the classical log-likelihood of a parametric model. Specifically, for a model whose parameters are estimated using maximum likelihood, the maximum log-likelihood $\ell(\hat{\boldsymbol{\theta}})$ is an estimator for $nE_g[\log f(X|\hat{\boldsymbol{\theta}}]$ (Konishi & Kitagawa, 2008).

Although the maximized log-likelihood $\ell(\hat{\boldsymbol{\theta}})$ is proportional to an estimator of the expected log-likelihood, it is a biased estimator. This is because the data are first used to estimate the model parameters and then the same data are used to estimate the expected log-likelihood. This reuse of data causes the expected value of the estimator to actually be greater than the expected log-likelihood. Under conditions given in Akaike, 1974, it can be shown that this bias is approximated by K, the number of estimated parameters in the model f. Hence

$$\ell(\hat{\boldsymbol{\theta}}) - K \tag{17}$$

is proportional to an accurate estimator of the expected log-likelihood, which is the relative K-L information of the approximating model f.

Multiplying (17) through by -2, we obtain the official definition of the Akaike Information Criteria (AIC),

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}) + 2K \tag{18}$$

---

[4] See Wasserman, 2004, section 9.3.

where K is the number of parameters in $f(x|\boldsymbol{\theta})$. Smaller values of AIC indicate greater expected log-likelihoods, hence smaller K-L distances. Therefore, for a set of approximating models, one should select the model with the smallest AIC.

BIC

The Bayesian Information Criteria (BIC) was introduced by Schwarz, 1978. Our introduction is based on Konishi and Kitagawa, 2008. Although the computational form of BIC resembles that of AIC, its theoretical basis is entirely different. As discussed earlier, AIC is based on estimating the K-L divergence of an approximating model from the true model. BIC, however, is based on the posterior probability of the approximating model. This follows from the Bayesian philosophy that a prior and posterior probability distribution can be established with regard to model parameters.[5] We present a brief derivation of BIC.

Let $Z_1, Z_2, \ldots, Z_s$ be competing parametric models, with distributions $f_i(x|\boldsymbol{\theta}_i)$ and priors $\pi_i(\boldsymbol{\theta}_i)$. Consider a sample $\mathbf{x}_n = \{x_1, \ldots, x_n\}$. For model Z_i, the marginal distribution of x_n is

$$p_i(\mathbf{x}_n) = \int g_i(\mathbf{x}_n, \boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \tag{19}$$

$$= \int f_i(\mathbf{x}_n|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \tag{20}$$

The integral in (20) is known as the "*marginal likelihood* of the data" (Konishi & Kitagawa, 2008) and represents the likelihood of the data given the i^{th} model.

Bayes theorem states that for partitioning events $\{A_1, \ldots, A_s\}$ and event B,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{s} P(B|A_j)P(A_j)} \tag{21}$$

Let the prior probability of model Z_i be $P(Z_i)$. Then applying Bayes theorem gives us

$$P(Z_i|\mathbf{x}_n) = \frac{p_i(\mathbf{x}_n)P(Z_i)}{\sum_{j=1}^{s} p_j(\mathbf{x}_n)P(Z_j)} \tag{22}$$

[5] See Wasserman, 2004. Ch. 11 for a concise introduction to Bayesian Inference.

This is the posterior probability of model Z_i given the data. In terms of model selection, it is reasonable to accept the model with the highest posterior probability. Hence we are looking for the model that maximizes the numerator of (22), for the denominator is the same across all models.

If the situation is such that we can assume that $P(Z_i)$ are all equal[6] then the numerator of (22) is maximized by the model that maximizes $p_i(\mathbf{x}_n)$. Hence we should select a model with the greatest marginal likelihood $p_i(\mathbf{x}_n)$. BIC is officially defined by multiplying the $\log p_i(\mathbf{x}_n)$ by -2. As shown before, the marginal likelihood is equal to the integral in (20). Thus we have

$$\text{BIC} = -2 \log p_i(\mathbf{x}_n) = -2 \log \left\{ \int f_i(\mathbf{x}_n|\boldsymbol{\theta}_i) \pi_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \right\} \tag{23}$$

The integral in (20) can be approximated using Laplace's method, giving us the computationally friendly form,

$$\text{BIC} \approx -2 \log f_i(\mathbf{x}_n|\hat{\boldsymbol{\theta}}_i) + k_i \log n \tag{24}$$

where $\hat{\boldsymbol{\theta}}_i$ is the MLE of $\boldsymbol{\theta}_i$ and k_i is the number of parameters in model Z_i. Since we multiplied $\log p_i(\mathbf{x}_n)$ by a negative, the method will indicate the model with the smallest BIC value.

Recall that $\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}) + 2K$. For a given model, $\log f_i(\mathbf{x}_n)$ is the maximized log-likelihood, $\ell(\hat{\boldsymbol{\theta}})$. Hence for a K parameter model, BIC can be written in a similar form to AIC, that is

$$\text{BIC} \approx -2\ell(\hat{\boldsymbol{\theta}}) + K \log n \tag{25}$$

In most cases, $K \log n$ will be greater than $2K$, thus BIC has a greater penalty for model complexity than does AIC. This distinction will become apparent in the upcoming Monte Carlo experiments[7].

[6] In a Bayesian context this simply means that in the absence of data we do not have reason to believe that one model is more accurate than another.

[7] The observations in this last paragraph are based on Wasserman, 2004, the assigned paper Sharma et al.,

Part 2. Monte Carlo Analysis

General Approach

The general framework for a Monte Carlo assessment of model selection criteria can be described as follows. We first design a model and use random number generation to simulate data from this model. This model is the "true" model in the sense that it is the distribution from which the data were generated. Knowledge of this true model is the key distinction between real world data and simulated data.

We then proceed to analyze the simulated data as if they were real data. We fit several models to the data and calculate the model selection criteria for each model. If a criterion indicates the true data generating model or a parsimonious approximation of the true model, this is taken to be evidence of the accuracy of the given criterion. Conversely, if a criterion indicates incorrect models this is evidence that the criterion is flawed. We repeat this process numerous times and report the proportion of success for each criterion.

In order to conduct the Monte Carlo experiment, we need a model from which to generate data. Since the project instructions called for the use of an existing data set, I thought that instead of coming up with a model from scratch, we could use real world data to help us develop a model. We can then use this model to generate new data and test our criteria. I later saw that Raffalovich et al., 2008 also used this approach in their research.

Our general approach follows that of Sharma et al., 2019 in the assigned paper, who used Monte Carlo Simulation to analyze model selection methods in the context of Information Systems Partial Least Squares modeling. We chose to conduct a similar analysis using Multiple Linear Regression models. This aligns more closely with material covered in 625.725. We now describe our experimental design in detail.

2019, and the R documentation on AIC/BIC.

Data Set

The data set used in this project was obtained from Kaggle at this link and pertains to housing prices in King County, USA. I used Excel to make some modifications to the original data set before conducting our analysis. Several columns were removed and the remaining column names were changed. The original column names pertained to different housing features, with price as the response variable. In this project we plan on using this data merely as a springboard to generate new data sets, hence we are not overly concerned with analyzing and understanding the factors affecting housing prices in King County. Therefore in order to simplify our experiment, I changed the predictor names to generic $x_1, x_2, \ldots, y$. This also reflects the more theoretical nature of this project. There are 15 independent variables labeled $x_1, \ldots x_{15}$ and a response variable y. The data set contains $21,613$ observations for which y varies from $75,000$ to $7,700,000$. This new data set is saved as "data_generic.csv" and can be obtained here at the project GitHub repository. Henceforth this data set will be known as DataSet1.

Experimental Design

Data Generation

The first step in our Monte Carlo Analysis is to design a "true" model from which to simulate data. For this experiment I chose to generate data from a multiple linear regression model with 9 covariates. Thus the simulated random variable Y_{sim} will be of the form

$$Y_{sim} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_9 X_9 + \varepsilon \tag{26}$$

where ε is normally distributed with mean 0 and standard deviation σ. I chose a model with 9 covariates because we are going to need to fit models with significantly different number of predictor variables than the data generating model. Since the data set contains 15 predictors, choosing a data generator with 9 predictors gives us the necessary flexibility.

The next step is to use our data set to obtain coefficients for the "true" model. With the help of the R lm function(R Core Team, 2022), I fit a linear regression model from the

response variable y to the first 9 predictor variables, $x_1, \ldots, x_9$. The coefficients from this model were then used as the coefficients for the data generating model, that is $\beta_0, \ldots, \beta_9$. For the random error component ε, I generated a vector of random deviates from a normal distribution with mean 0 and standard deviation 10000. This standard deviation is reasonable relative to the range of y values in DataSet1. The number of elements in the vector ε was set to be equal to the number of observations in DataSet1, which is 21613.

The actual data generation process can be described simply using vector notation. Consider a data set similar to the ones described before. Let n be the number of observations in the data set, p be the number of predictor variables in the data set, $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$ be the p columns of covariate values, $\boldsymbol{\beta_0}$ be a vector of length n with each component equal to β_0, and $\boldsymbol{\varepsilon}$ be a vector of length n where each component is a random deviate from a normal distribution with mean 0 and standard deviation σ. Now we can generate n observations from a response variable y_{sim} with the following formula

$$y_{sim} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\mathbf{x_1} + \boldsymbol{\beta_2}\mathbf{x_2} + \cdots + \boldsymbol{\beta_9}\mathbf{x_9} + \boldsymbol{\varepsilon} \tag{27}$$

The R programming language natively supports this vector arithmetic. Note the similarity between the generating formula in (27) and the random variable form of Y_{sim} in (26).

After using this process to generate y_{sim}, we combined the first 15 columns of DataSet1 with y_{sim} to form a new data frame called sim_data. This is the data set upon which we will perform our model selection analysis.

Model Fitting and Selection

After generating the simulated data set, we proceed to fit several models to the data. For this experiment I fit five models to sim_data. They are:

1. Model 1: The true data generating model. This model includes predictors $x_1, x_2 \ldots x_9$.

2. Model 2: A parsimonious approximation of the true model. This model includes $x_1, x_2 \ldots x_6$.

3. Model 3: An under-fit model. This model includes x_1, x_2, and x_3.

4. Model 4: An entirely incorrect model that does not include all of the true predictors and also includes several spurious predictors. This model includes $x_1, x_3, x_5, x_7, x_{10}, x_{11}, x_{12}$, and x_{14}.

5. Model 5: An over-fit model that includes all the predictors in the data set. This model includes $x_1, x_2, \ldots x_{15}$.

After fitting the five models, we compute R^2, Adjusted R^2, AIC, and BIC for each model and record the model selections. That is, for R^2 and Adjusted R^2 the model with the greatest value is recorded, while for AIC and BIC the model with the smallest value is recorded. Analyzing these model selections will give us insight into the functionality and accuracy of the various criteria.

Monte Carlo Replications

Our process so far can be summarized with the following algorithm:

1. Take an existing data set and fit a linear model from the response variable to the first 9 predictor variables.

2. Generate an normal random error vector.

3. Simulate a response variable using the coefficients obtained in step 1 and the error generated in step 2.

4. Fit the five models to the simulated data.

5. Compute the four model selection criteria for each model.

6. Record model selections for each criterion.

These steps are necessary for one iteration of the experiment and form a base case for the Monte Carlo analysis. However, we are interested in assessing the performance of model

selection criteria over numerous iterations (i.e. many different data sets). The key difference between iterations is the normal random error vector, $\boldsymbol{\varepsilon}$. Since $\boldsymbol{\varepsilon}$ is randomly generated, it takes on different values on each iteration, leading to a new data set. Therefore we need to repeat steps 2 - 6 on each iteration. To practically accomplish this, I wrapped these steps into a single function called experiment1.

For this experiment, I ran 1000 iterations of the experiment1 function and recorded the model selection proportions for each criterion. For example, I recorded the percent of times AIC indicated Model 1, Model 2, etc. The same was done for each criterion. Since 1000 iterations takes a significant amount of processing power, I used the R package future.apply(Bengtsson, 2021) for multiprocessing and specifically the function future_replicate to implement the Monte Carlo replications.

Results

The results are given in the following table. Each row corresponds to one of the 5 models discussed before. For each model, the table records the proportion of times that each criterion indicated that particular model.

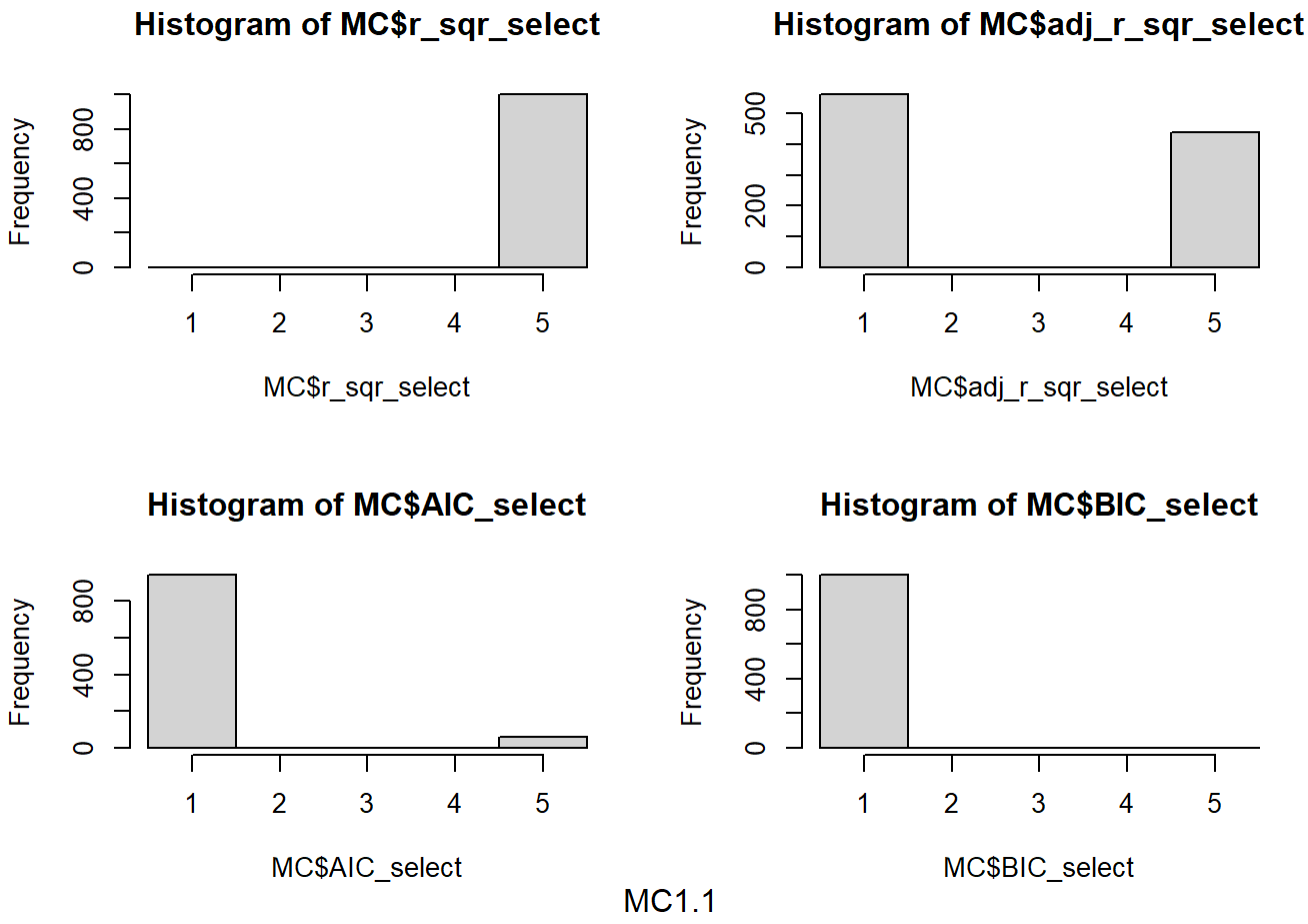
Experiment Results:

model	r_sqr_prop	adj_r_sqr_prop	AIC_prop	BIC_prop
1	0	0.594	0.942	1
2	0	0.000	0.000	0
3	0	0.000	0.000	0
4	0	0.000	0.000	0
5	1	0.406	0.058	0

What we see is that R^2 always indicated Model 5, the over-fit model containing all possible predictors. This behavior is consistent with the theory discussed in Part 1. Adjusted R^2 did significantly better than R^2 and indicated the true model (Model 1) about 60% of the time. However, Adjusted R^2 also suffered from over-fitting and indicated Model 5 on 40% of experimental runs. AIC performed far better than the classical methods,

indicating the true model approximately 94% of the time. BIC had the best performance, indicating the true model on every run.

These results can also be visualized with the following histograms.



MC1.1

Our results resemble those of Sharma et al., 2019 in the assigned paper and support their claim that the classical R^2 and Adjusted R^2 should not be used for model selection. AIC and BIC detected the true model with far greater accuracy and hence are recommended methods.

It is interesting to note that AIC did indicate the over-fit model on about 6% of the trials, while BIC never indicated this model. This can be explained by our earlier observation that BIC has a stronger penalty for model complexity than AIC does. Thus BIC is less likely to indicate over-fit models.

For researchers who continue to use R^2 or Adjusted R^2, it is comforting to observe that neither method ever indicated the completely incorrect or severely under-fit models. Burnham and Anderson, 1998 state that "Shibata, 1989 argues convincingly that under-fitted models are a more serious issue in data analysis and inference than over-fitted models."

Conclusions

We have used Monte Carlo simulation to demonstrate that AIC and BIC have superior ability to select the true model in the event that the *true data generating model is including in the set of competing models.* This is rarely the case in real world data analysis; "exploratory research typically involves situations where researchers are unlikely to have access to or awareness of the complete set of variables and linkages that formed the observed reality." (Sharma et al., 2019) The goal of practical model selection then is not to select the objectively true model, rather it is to select a parsimonious approximating model that is consistent with the true data generating process, if not capturing complete reality.[8] If so, we have yet to compare the performance of different model selection criteria in the more realistic case that the true data generating model is *not* included in the set of competing models.

Sharma et al., 2019 conducted a Monte Carlo simulation under these conditions and

[8] Based on Sharma et al., 2019

obtained results comparable to the those obtained when the true model was included in the set of competing models. That is R^2 and Adjusted R^2 tended to select the saturated/over-fit model, while AIC and BIC had a much stronger tendency to select the models that were parsimonious approximations of reality.

The author also conducted several Monte Carlo experiments under similar conditions. This was accomplished by a modification of our experimental design. In our original experiment, after generating the simulated data set, we fit five models to the data, including the true model. Now we followed the same process but only fit the latter four models, which were the parsimonious approximation, the under-fit model, the completely false model, and the over-fit model. The entire process was replicated 1000 times.

To my surprise, R^2, Adjusted R^2, AIC, and BIC all consistently indicated the over-fit model. These results are in significant contrast to those of Sharma et al., 2019. I have not yet obtained a complete explanation for this discrepancy. As discussed earlier, our true model included $x_1, \ldots x_9$, while the original parsimonious approximation model included $x_1, \ldots x_6$. I suspected that vis-a-vis our data set, $x_1, \ldots x_6$ really constituted an under-fit model and hence was not selected. I ran the experiment again with the parsimonious approximation model including $x_1, \ldots, x_8$, however again all the criteria indicated the over-fit model. This behavior remains an open question and requires further analysis.

Appendices

Appendix A: Data

The data sets used in this project are available at https://github.com/eb-addess/625.725-Project/tree/main/data. The original data set is called kc_house_data.csv. The modified version is called housing_data_adjusted.csv. The final generic version that was used for our analysis is called data_generic.csv. Our modifications were explained in the Data Set section of the paper.

Appendix B: Code

All of the programming in this project was done using the R language(R Core Team, 2022). The code is available at https://github.com/eb-addess/625.725-Project. This GitHub repository contains several folders. The folder entitled Experiments contains the scripts for the three experiments described in the paper. Experiment1.1 is the main Monte Carlo experiment, which contained the true model among the competing models. Experiment 2.1 is the first experiment referenced in the Conclusions section, which did not include the true model among competing models. Experiment 3.1 is the second experiment mentioned in Conclusions, which also did not include the true model, but had a parsimonious model including predictors $x_1, \ldots x_8$.

The reader will notice that the experiment scripts are rather short. This is because they use the project functions. These functions do all the heavy lifting and can be found in the script ProjectFunctions.R located in the folder entitled Functions. These include specific functions for each experiment and the functions used to produce tables and histograms of results.

Finally the folder Base_Cases contains the base case code for each experiment. The base cases include all the steps for one iteration of the Monte Carlo simulations. These scripts represent an important step in the development of the final project code. The base case was written first and was then wrapped into the experiment function for replication. This allowed me to think more clearly about the steps necessary for the basic simulation without the complications of repetition. I think that this approach might be helpful for anyone looking to conduct a similar analysis.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Bengtsson, H. (2021). A unifying framework for parallel and distributed processing in r using futures. *The R Journal*, *13*(2), 208–227. https://doi.org/10.32614/RJ-2021-048

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference: A practical information-theoretic approach*. Springer.

deLeeuw, J. (1992). Introduction to akaike (1973) information theory and an extension of the maximum likelihood principle. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Foundations and basic theory* (pp. 599–609). Springer. https://doi.org/10.1007/978-1-4612-0919-5_37

Ford, C. (2015). Is r-squared useless? | university of virginia library research data. https://data.library.virginia.edu/is-r-squared-useless/

Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer.

Larsen, R. J., & Marx, M. L. (2018). *An introduction to mathematical statistics and its applications* (Sixth edition). Pearson.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2020). *Introduction to linear regression analysis* (Fifth edition). Wiley.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raffalovich, L. E., Deane, G. D., Armstrong, D., & Tsao, H.-S. (2008). Model selection procedures in social research: Monte-Carlo simulation results [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03081070802203959]. *Journal of Applied Statistics*, *35*(10), 1093–1114. https://doi.org/10.1080/03081070802203959

Schwarz, G. (1978). Estimating the Dimension of a Model [Publisher: Institute of Mathematical Statistics]. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shalizi, C. (2015). Lecture 10: F-tests, r2, and other distractions |carnegie mellon university statistics & data science. https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/10/lecture-10.pdf

Sharma, P., Sarstedt, M., Shmueli, G., Kim, K., & Thiele, K. (2019). PLS-Based Model Selection: The Role of Alternative Explanations in Information Systems Research. *Journal of the Association for Information Systems*, *20*(4). https://doi.org/10.17005/1.jais.00538

Shibata, R. (1989). Statistical aspects of model selection. In *From data to model* (pp. 215–240). Springer.

Subrahmanya, R. (2018). *Analysis of boston housing data using linear regression ,trees and gam*. https://rpubs.com/Rashmi_Subrahmanya/371719

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. Springer.