Global Scientific JOURNALS

GSJ: Volume 8, Issue 12, December 2020, Online: ISSN 2320-9186
www.globalscientificjournal.com

A New Method for Regression Model Selection

NIYONZIMA Felix

Teacher of Physics at G.S. Kabuga (A Senior High School; Kigali, Rwanda)

E-mail: nhashimfelix2014@gmail.com

Tel: +250788841115.

Abstract

In this paper we intend to present a new method for regression model selection. The new model selection method that we developed uses clear criteria for variables selection as well as best model selection and its strength lies in its ability to select a best model even before accomplishing all required steps. We have used the built-in R data sets and the provided R packages to develop the procedures through which our new model selection method selects the best model from all possible models. Our new method can be performed via two approaches which always lead to the same result, this will avoid a confusion for statisticians and data scientists about the choice of the approach to be used.

Our new regression model selection method also confirms the weaknesses of some existing model selection methods presented in many papers by various researchers. For the same data, the new method we wish to present in this paper results either in the best model better than that results if the stepwise regression have been used or in the same model as the stepwise regression. This paper presents the detail of the method, the variable selection criteria and processes as well as the best model selection criterion will be presented. The new method's distinctive features relative to the most popular and most used methods like the stepwise regression and the best subset regression will be stated.

Keywords: Regression model selection, Stepwise regression, Best subsets regression, Adjusted R-squared, Akaike information criterion (AIC) and Model sigma (Residual standard error, RSE).

1. Introduction

Regression model selection has been proved to be an important part of the regression analysis since not all available predictor variables are appropriate for explaining the variation in the response variable as well as predicting the new response outcome. Regression model selection aims at selecting the important predictor variables which accurately explain and predict the response variable. Researchers have spent much effort to cope with such task, as the result, the methods like the stepwise regression, the best subsets regression, the LASSO regression, the ridge regression, etc, are available to be applied for selecting the best model from all possible models (Joseph B. Kadane & Nicole A. Lazar 2004) . The quantities such as adjusted R-squared, p-value for F-test/t-test, mean square error (MSE), Akaike information criterion (AIC), Bayesian information criterion (BIC) and Mallows Cp are the most used for variable selection as well as model selection criteria (Joseph B. Kadane & Nicole A. Lazar 2004, Frank Emmert-Streib & Matthias Dehmer 2019, Obubu Maxwell et al. 2019, Heinze G., Wallisch C., & Dunkler D. 2018).

The best subsets regression is a regression model selection method which works by building all possible models according to available predictor variables and compare them with aid of mean square error (MSE) or the coefficient of determination (R-squared) which leads to “p” models, “p” is the number of the predictor variables. A best model is selected from such “p” models after comparing them with aid of either AIC, BIC or Mallows Cp (Frank Emmert-Streib & Matthias Dehmer 2019, Maya Lozinski 2018, Zhang Z. 2016).

The stepwise regression is another popular regression model selection method, it operates through two approaches: forward selection and backward elimination. Forward selection approach starts by a model with only intercept and proceeds to adding one by one all significant variables according to a predetermined significance level until none of the predictor variables fulfills such criterion. Backward elimination approach starts from the full model and proceeds to removing one by one all insignificant variables according to a predetermined significance level until no variable is still insignificant (Loann David Denis

Desboulets 2018, Joseph B. Kadane & Nicole A. Lazar 2004, Obubu Maxwell et al. 2019, Zhang Z. 2016).

While the best subsets regression and the stepwise regression are the most popular and the most available in many statistical software, their applicability is feasible to some cases and not to other cases or the result is not reliable. Thus we still need a method which is applicable to a wide range of problems and whose the result is reliable (Joseph B. Kadane & Nicole A. Lazar 2004).

The new method for regression model selection that will be presented in this paper uses the p-value for t-test and F-test for the variables selection criteria and the adjusted R-squared or AIC for the best model selection criterion, and it can reveal a best model which performs even better than the model built upon all available predictor variables. We shall start by explaining the reason for the need for the model selection and a short review of some existing model selection methods. Next we shall explain the processes of selecting the best model using our new method, thereafter, we shall introduce the strength of our method over the most popular model selection methods like the stepwise regression and the best subsets regression and finally draw a conclusion.

2. Need for Regression Model Selection and a Brief Review of Some Existing Methods

Suppose that a researcher collects data about p variables, $x_1, x_2, x_3, \ldots, x_{1-p}, x_p$, that he/she thinks to be the major factors which affect the variable, y, he/she will need to write a mathematical equation which describes the relationship between such variables and y. This equation is in this field called the regression model and has the following form:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \cdots + \beta_{1-p} x_{i,(1-p)} + \beta_p x_{i,p} + \varepsilon_i$$

However, researchers showed that, it is not always important to include all available predictor variables, $x_1, x_2, x_3, \ldots, x_{1-p}, x_p$, in the model since the issues of overfitting may occurs, on the other hand, ignoring many variables may result in underfitting issue, all such issues have to be avoided (Frank Emmert-Streib & Matthias Dehmer 2019).

Underfitting occurs in the model when important variables affecting the response variable are not included in the model. On the other hand, a model which include important

predictor variables together with the variables which have less or no relationship with the response variable is subjected to overffiting problem. The model with underfitting problem is poor in explaining the variation in the response variable and the researcher needs to collect more data so as to solve for such problem (Frank Emmert-Streib & Matthias Dehmer 2019). Overfitting leads to a complex model which is sometimes hard to interpret, it causes the confusion in identifying the major factors which affect the response variable. In the context of prediction, a model with overfitting problem can perform better on the training data but performs worse on the test data (Frank Emmert-Streib & Matthias Dehmer 2019, Obubu Maxwell et al. 2019). Therefore, a best model have to keep a trade-off between overfitting and underfitting; that is, it must be simple, easy to interpret, strongly explain the variation in the response variable and perform better when used to predict new unseen data (Frank Emmert-Streib & Matthias Dehmer 2019).

A move to finding the technique of selecting a best model from all possible models resulted in the development of the methods such as the best subsets regression, the stepwise regression, the ridge regression and the LASSO regression. However, not only such methods are available for the model selection purpose, we choose to mention only those since they have been proved to be the most performer among others and they are also the most discussed in many literature, among them also the stepwise regression is the most popular and the most applied in many statistical analyses particularly in big data analysis (Frank Emmert-Streib & Matthias Dehmer 2019, Heinze G., Wallisch C., & Dunkler D. 2018).

The ridge regression is first a technique for building the regression model as does the ordinary least squares technique. The ordinary least squares (OLS) minimises the residual sum of squares and it includes all the predictor variables in the model which can lead to overfitting issue. In addition, the OLS does not care about the relationship between the predictor variables which can lead to multicollinearity issue (RANJIT KUMAR PAUL 2008, Ali BAGER et al. 2017). The ridge regression works by adding a ridge parameter to ordinary least squares information matrix, in other words, it puts constraint to the regression coefficients which make them to be shrunk towards the population parameters (RANJIT KUMAR PAUL 2008, Ali BAGER et al. 2017, Olga Morozova et al. 2015). Such property allows the ridge regression technique to differentiate the most important predictor variables from the least important variables, which leads to solving the overfitting issue. In addition, the technique reduces the inter-relations between the predictor variables which

leads to solving the multicollinearity issue. These qualities make the ridge regression to be used beyond the regression model building, but also for regression model selection (RANJIT KUMAR PAUL 2008, Ali BAGER et al. 2017). The weakness of the ridge regression is that, the regression model built using the ridge regression is hard to interpret since the technique does not eliminate any predictor variable, instead it includes all of them, and thus the additional model selection method is needed (Joseph B. Kadane and Nicole A. Lazar 2004, Olga Morozova et al. 2015).

The LASSO regression also applies shrinkage to the regression coefficients as does the ridge regression. Shrinkage in regression analysis implies that the model coefficients are made closer to the population parameters. The difference between the LASSO regression and the ridge regression is that, if multicollinearity is present between for example four variables, the LASSO chooses only one variable and set the remaining three coefficients to zero (Trevor Hastie, Robert Tibshirani & Ryan J. Tibshirani 2018; Maya Lozinski 2018; Olga Morozova et al. 2015; Heinze G, Wallisch C. & Dunkler D. 2018), which results in a reduced model relative to the obtained model if the ridge regression has been used. However, if more coefficient are set to zero, we may lose much information which leads to the model with low accuracy (Trevor Hastie, Robert Tibshirani & Ryan J. Tibshirani 2018; Maya Lozinski 2018).

The best subsets regression selects a best model from all possible models that can be built from “p” predictor variables. The method builds all possible models considering “k” predictor variables (k= 1, 2, 3, …, p), for each “k” many models are built and only one model with the largest MSE or highest R-squared is chosen, resulting in “k” models. Since such “k” models differ in their complexity, the best model is selected from them with aid of either AIC or BIC or Mallows Cp, it is the model with the lowest AIC or lowest BIC or lowest Mallows Cp. (Frank Emmert-Streib & Matthias Dehmer 2019, Zhang Z. 2016). However, for “p” predictor variables, the model will build “2^p” models, for example for 20 variables, 1048576 models will be built, this causes a computational problem, such models are so hard to compute when the predictor variables are too many, thus the method remains feasible when the predictor variables are moderate (Loann David Denis Desboulets 2018).

The stepwise regression is the most popular and the most available model selection method in many statistical software (Loann David Denis Desboulets 2018; Joseph B. Kadane & Nicole A. Lazar 2004; Gary Smith 2018). Instead of building all possible models

as does the best subsets regression, the stepwise regression builds a subset of all possible models (Loann David Denis Desboulets 2018) which makes it suitable and the only feasible method for big data analysis (Joseph B. Kadane & Nicole A. Lazar 2004, Zhang Z. 2016). It comprises two approaches, backward elimination and forward selection. Forward selection approach starts from a model with only intercept and proceeds to including one by one those predictor variables whose the p-value is less than the chosen alpha-to-enter value given that the whole model is significant, the algorithm stops when no other predictor variable fulfills such criteria (Joseph B. Kadane & Nicole A. Lazar 2004, Obubu Maxwell et al. 2019, Gary Smith 2018). Backward elimination approach starts from a full model and removes one by one those predictor variables which have a p-value greater than the chosen alpha-to-remove threshold until no variable fulfills the condition for removal. An improved version of the stepwise regression is the combination of both backward elimination and the forward selection approaches (Joseph B. Kadane & Nicole A. Lazar 2004, Gary Smith 2018).

However, for the same data, backward elimination and forward selection approaches may result in the different best models. Furthermore, the stepwise regression do not search all possible models and thus it is not guaranteed to find the best model (Frank Emmert-Streib & Matthias Dehmer 2019, Joseph B. Kadane & Nicole A. Lazar). Gary Smith (2018) mentioned various criticisms about the stepwise regression, the major one is that, it may include the predictor variables which explain less the response variable while it leaves the most important predictors. Therefore, a new regression model selection method is still awaited by statisticians and data scientists based on the fact that the existing methods are bordered by various limitations. The new regression model selection method that we wish to present in this paper tries to provide solutions to such limitations as it will be discussed in detail in the sections ahead.

3. Processes and Criteria for our New Regression Model Selection Method

As it is the purpose of this paper, we would like to illustrate a new method of selecting the best regression model. The task of selecting such model involves the work of selecting important predictor variables that are pretty best to explain and predict the response variable. Our new regression model selection method operates through two approaches: **the any-predictors start-up approach** and **the single-predictor start-up approach**. The

processes through which the important predictor variables are selected in such approaches are based on the criteria constituted by the p-value for t-test and the p-value for F-test, the best model selection is performed with aid of the adjusted R-squared or AIC.

Here below are the summary of the steps of the procedures for variables selection as well as best model selection in our new model selection method, the detail that will help the deep understanding of such procedures will be presented with aid of examples in the next section.

- The method (any-predictors start-up approach and the single-predictor start-up approach) starts by setting the initial best model (called again first best model). The first best model is composed by some of the available predictor variables and they are all significant according to a chosen significance level. The significance level is either 0.05 or 0.1

Next thing to do after setting the first best model, is to enter all available predictor variables into the first best model in the following order:

- Enter each variable, one by one into the first best model until all of them are tested for inclusion.
- Enter two predictor variables simultaneously into the best model until all possible combinations of two variables are tested for inclusion.
- Enter three predictor variables simultaneously, four predictor variables simultaneously, five predictor variables simultaneously, etc, until all possible combinations are tested for inclusion.
- Suppose there are “p” predictor variables, the last step is to enter “p” variables simultaneously into the best model.

The any-predictors start-up approach and the single-predictor start-up approach differ only by the structure of the first best model and the way of setting it, other processes of selecting the predictor variables as well as the best model are the same. At each step of entering the predictor variables into the best model, important variable(s) is (are) selected, it/they is/are the predictor variable(s) which has/have a p-value below the chosen significance level, this leads to the new best model. This best model is compared to the best

model selected at the preceding step with aid of either adjusted R-squared or AIC so as to select the better one and continue the process.

4. Demonstration of the New Regression Model Selection Method

In this section we present some examples to facilitate the full understanding of the new model selection processes above. In all examples the significance level is 0.05, the data are the built-in R data sets.

4.1. First approach: The any-predictors start-up

Example1:

In this example, we use the "biopsy" data of the "MASS" package to examine the relationship between the variable "V5" and the remaining variables except the variable "ID". After removing "ID", we have stored the data into the object "biopsy1".

The first few observations of the data are shown below.

V1	V2	V3	V4	V5	V6	V7	V8	V9
5	1	1	1	2	1	3	1	1
5	4	4	5	7	10	3	2	1
3	1	1	1	2	2	3	1	1
6	8	8	1	3	4	3	7	1

The any-predictors start-up approach starts by setting the initial best model (the first best model) which is determined with aid of full model. Let's build a full model, only the predictor variables are shown, the response variable is "V5."

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	1.16245	0.11218	10.36245	0.00000
V1	0.00507	0.02621	0.19343	0.84668
V2	0.29868	0.04543	6.57398	0.00000
V3	0.08282	0.04575	1.81011	0.07072
V4	0.02731	0.02934	0.93067	0.35236
V6	0.03418	0.02365	1.44533	0.14883
V7	0.03499	0.03707	0.94375	0.34563
V8	0.07738	0.02719	2.84533	0.00457
V9	0.18768	0.03581	5.24169	0.00000

After building the full model, we classify the predictor variables into two classes: **the first class predictors** and **the second class predictors**. The second class predictors include initially all insignificant predictor variables found in the full model. The first class predictors are determined by building another model containing only the predictor variables which are found to be significant in the full model, if they all remain significant in such model, we proceed to performing what we mane in this paper "filtering", if some of them become insignificant we move them into the second class predictors, thereafter we perform filtering to the remaining variables. Looking at the table above, the variables "V2", "V8" and "V9" are significant (their p-values are less than 0.05). Let's build a model with such variables, the response variable remains the same, here below the predictor variables are shown.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.29546	0.08277	15.65220	0e+00
V2	0.41906	0.02607	16.07717	0e+00
V8	0.10792	0.02578	4.18641	3e-05
V9	0.19256	0.03555	5.41685	0e+00

We see from above table that, the predictor variables remain significant in this model (the model with only the significant predictor variables found in the full model), then we proceed to performing "filtering". This model may include some variables which add nothing to it, and which reduce nothing when removed from the model. Filtering consists of checking and removing such predictor variables from the model, if found we move them into the second class predictors and the remaining variables constitute the first class predictors, if the significant predictor variables from the full model remain significant in the model containing only them, we classify all of them into the first class predictors. Filtering is done by removing one by one some predictor variables starting from the least significant variable without affecting the model's adjusted R-squared.

The model with only the significant variables found in the full model has the following formula.

```
fit = lm(V5~V2+ V8+ V9, data = biopsy1)
```

The model's adjusted R-squared is 0.5990. During filtering for this example, we first remove "V8" since it has the least p-value among others, then we will have a model with only "V2" and "V9". The formula is shown below:

```
fit_t= lm(V5~V2+ V9, data = biopsy1)
```

This model is associated with the adjusted R-squared equals 0.5892. We compare then this value to the adjusted R-squared (0.5990) of the model containing also "V8" (named "fit" above), we see that the adjusted R-squared reduces after removing "V8", we conclude that the variable "V8" as well as "V2" and "V9" are all important in the model "fit" and thus we classify them into the first class predictors, the remaining variables are all classified into the second class predictors, they are: "V1", V3","V4","V6" and "V7". Now, we form our first best model, it is the model whose the predictor variables are into the first class predictors, the formula is shown below:

```
best1 = lm(V5~V2+ V8+ V9, data = biopsy1)
```

We keep this model and we proceed to entering into it the second class predictors, this action will results in many best models that we will compare to select the final best model. But in order to enter the variables of the second class predictors into the first best model, we have to mix them with the variables of the first class predictors and enter all of them into the first best model following a specific order. In other words, we will enter all available predictor variables into our first best model so as to select the final best one. We start by entering one by one all available predictor variables in any order.

Step 1: Enter "V1" into the first best model, "best1", only the predictor variables are shown below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.22678	0.10304	11.90627	0.00000
V2	0.40558	0.02871	14.12616	0.00000
V8	0.10433	0.02597	4.01650	0.00007
V9	0.19018	0.03561	5.34137	0.00000
V1	0.02821	0.02521	1.11877	0.26363

It is seen that "V1" is insignificant (the significance level is 0.05), we remove it from this model and check the significance of the remaining predictor variables as shown below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.29546	0.08277	15.65220	0e+00
V2	0.41906	0.02607	16.07717	0e+00
V8	0.10792	0.02578	4.18641	3e-05
V9	0.19256	0.03555	5.41685	0e+00

We see that after removing "V1", the remaining variables remain significant, thus we get our second best model, its formula is shown below:

```
best2 = lm(V5~V2+ V8+ V9, data = biopsy1)
```

We compare the second best model to the first best model with aid of either adjusted R-squared or Akaike information criterion (AIC). Both models, "best1" and "best2" are the same and should have the same adjusted R-squared or the same AIC, we discard any one of them. Let's choose to discard "best1" and keep the second best model, "best2". From "best2" we build the third best model by entering into it the next predictor variable, "V2".

Step 2: Enter "V2" into the second best model, "best2".

At this step, "V2" is already in the "best2", thus the third best model will the same as the second best model, its formula is:

```
best3 = lm(V5~V2+ V8+ V9, data = biopsy1)
```

We compare this new best model to the previous one. Since both models, "best3" and "best2" are the same, they have the same adjusted R-squared, we then discard "best2" and keep "best3" as our new best model. We enter then into it another variable and find another new best model.

Step 3: Enter "V3" into the previous best model, "best3".

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.25776	0.08369	15.02812	0.00000
V2	0.32978	0.04313	7.64629	0.00000
V8	0.09315	0.02630	3.54229	0.00042
V9	0.19016	0.03541	5.36991	0.00000
V3	0.11360	0.04383	2.59190	0.00975

We see from the above table that "V3" is significant as well as other predictor variables, we then get our new best model, we name it "best4".

```
best4 = lm(V5~V2+ V8+ V9+ V3, data = biopsy1)
```

We compare this model, "best4" to the previous best model, "best3" with aid of either adjusted R-squared or AIC. We have found that "best4" has a larger value of adjusted R-squared than "best3", 0.602 over 0.599, thus we discard "best3" and consider "best4" as our new best model. We keep it and from it, we build another best model by entering into it the variable "V4".

Step 4: Enter "V4" into the previous best model, "best4", only the predictor variables are shown below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.23534	0.08453	14.61490	0.00000
V2	0.31229	0.04419	7.06698	0.00000
V8	0.08631	0.02654	3.25203	0.00120
V9	0.18274	0.03561	5.13218	0.00000
V3	0.10512	0.04402	2.38778	0.01722
V4	0.04816	0.02732	1.76252	0.07843

We see that "V4" is insignificant, we remove it and then check the significance of the remaining predictor variables as shown below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.25776	0.08369	15.02812	0.00000
V2	0.32978	0.04313	7.64629	0.00000
V8	0.09315	0.02630	3.54229	0.00042
V9	0.19016	0.03541	5.36991	0.00000
V3	0.11360	0.04383	2.59190	0.00975

The remaining variables are all significant, we then get a new best model, "best5". We compare this best model to the previous one, "best4". Both models are the same, thus they

have the same adjusted R-squared, we discard "best4" and keep the new best model, "best5".

```
best5 = lm(V5~V2+ V8+ V9 + V3, data = biopsy1)
```

Into "best5" we enter the next variable to form a new best model.

Step 5: Enter "V6" into the previous best model, "best5", only the predictor variables are shown.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.22713	0.08455	14.51319	0.00000
V2	0.31816	0.04331	7.34589	0.00000
V8	0.08628	0.02640	3.26834	0.00114
V9	0.19043	0.03531	5.39350	0.00000
V3	0.08764	0.04520	1.93884	0.05294
V6	0.04796	0.02136	2.24504	0.02509

We see that "V3" is insignificant, we remove it and then check the significance of the remaining predictor variables as shown below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.24755	0.08406	14.84040	0.00000
V2	0.37997	0.02938	12.93179	0.00000
V8	0.09541	0.02603	3.66580	0.00027
V9	0.19222	0.03537	5.43506	0.00000
V6	0.05855	0.02069	2.82958	0.00480

We see from above table that, after removing "V3" the remaining variables remain all significant, we then get our new best model, its formula is:

```
best6 = lm(V5~V2+ V8+ V9+ V6, data = biopsy1)
```

We compare this model to the previous one, we have seen that "best6" has the adjusted R-squared of 0.603 while "best5" has the adjusted R-squared of 0.602, thus we discard

"best5" and consider "best6" as our new best model, we keep it and enter into it "V7" to find the new best model.

Step 6: Enter "V7" into the previous best model, "best6", only the predictor variables are shown below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.18855	0.09659	12.30523	0.00000
V2	0.36432	0.03197	11.39469	0.00000
V8	0.08771	0.02675	3.27885	0.00110
V9	0.19460	0.03541	5.49644	0.00000
V6	0.05038	0.02171	2.32085	0.02059
V7	0.04515	0.03645	1.23885	0.21583

"V7" is insignificant, we remove it and check the significance of the remaining variables. In the previous step, we have found "V2", "V8", "V9" and "V6" to be significant into a single model, then we get the new best model, its formula is:

```
best7 = lm(V5~V2+ V8+ V9+ V6, data = biopsy1)
```

We compare this best model to the previous best model, "best6". "best7" and "best6" are the same and have the same adjusted R-squared, we discard "best6" and keep "best7", thereafter we enter "V8" into it to form the new best model. However, "V8" is already into "best7", therefore, the new best model named "best8" will be the same as "best7", we then keep "best8" and discard "best7".

```
best8 = lm(V5~V2+ V8+ V9+ V6, data = biopsy1)
```

We enter "V9" into "best8" to find the new best model, "V9" is already into "best8", thus our new best model, named "best9", will be the same as "best8", we will keep it and discard the previous one. At this step, our best model has the following formula:

```
best9 = lm(V5~V2+ V8+ V9+ V6, data = biopsy1)
```

We finish to enter one by one the second class predictor variables into the best model, the next thing to do is to enter two predictor variables simultaneously into the best model. That is, we will enter in any order (V1, V2), (V1, V3), (V1, V4), (V1, V6), (V1, V7), (V1, V8), (V1, V9), (V2, V3), (V2, V4), (V2, V6) …, (V7, V8), (V7, V9) and finally (V8, V9).

Step 9: Enter "V1" and "V2" simultaneously into the previous best model, "best9". Note that "V2" is already in "best9".

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.22254	0.10260	11.91526	0.00000
V2	0.37622	0.03069	12.25896	0.00000
V8	0.09449	0.02613	3.61594	0.00032
V9	0.19130	0.03545	5.39583	0.00000
V6	0.05626	0.02139	2.62975	0.00874
V1	0.01104	0.02594	0.42579	0.67040

"V1" is insignificant, we remove it and check the significance of the remaining variables. The remaining predictor variables have been found to be significant in the previous model, then, we get a new best model, named "best10"

```
best10 = lm(V5~V2+ V8+ V9+ V6, data = biopsy1)
```

"best10" and "best9" are the same, we discard "best9" and keep "best10".

Step 10: Enter "V1" and "V3" simultaneously into "best10". The predictor variables are shown below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.21795	0.10244	11.89000	0.00000
V2	0.31746	0.04357	7.28709	0.00000
V8	0.08603	0.02646	3.25156	0.00120
V9	0.19011	0.03539	5.37145	0.00000
V6	0.04722	0.02188	2.15813	0.03127
V1	0.00416	0.02614	0.15899	0.87372
V3	0.08663	0.04568	1.89657	0.05831

We see that some predictor variables are insignificant, we will remove them starting from the most insignificant, here it is "V1", we remove it and check the significance of the remaining variables as shown below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.22713	0.08455	14.51319	0.00000
V2	0.31816	0.04331	7.34589	0.00000
V8	0.08628	0.02640	3.26834	0.00114
V9	0.19043	0.03531	5.39350	0.00000
V6	0.04796	0.02136	2.24504	0.02509
V3	0.08764	0.04520	1.93884	0.05294

After removing "V1", "V3" remains insignificant, we also remove it and check the significance of the remaining variables as shown below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.24755	0.08406	14.84040	0.00000
V2	0.37997	0.02938	12.93179	0.00000
V8	0.09541	0.02603	3.66580	0.00027
V9	0.19222	0.03537	5.43506	0.00000
V6	0.05855	0.02069	2.82958	0.00480

The remaining predictor variables are all significant, we get then a new best model, its formula is shown below.

```
best11 = lm(V5~V2+ V8+ V9+ V6, data = biopsy1)
```

We compare this model, "best11" to the previous model "best10". Both model are the same, we then discard "best10" and keep "best11".

Step 11: Enter "V1" and "V4" simultaneously into "best11".

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.20768	0.10335	11.68588	0.00000
V2	0.36432	0.03230	11.28030	0.00000
V8	0.09058	0.02633	3.43971	0.00062
V9	0.18583	0.03575	5.19835	0.00000
V6	0.04717	0.02273	2.07510	0.03836
V1	0.01287	0.02597	0.49536	0.62051
V4	0.03408	0.02891	1.17884	0.23888

"V1" is the most insignificant, we remove it and check the significance of the remaining variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.23709	0.08453	14.63433	0.00000
V2	0.36897	0.03089	11.94600	0.00000
V8	0.09174	0.02621	3.49982	0.00050
V9	0.18703	0.03564	5.24704	0.00000
V6	0.05006	0.02196	2.27964	0.02294
V4	0.03322	0.02884	1.15202	0.24972

"V4" is still insignificant, we remove it and check the significance of the remaining predictor variables. It is seen that after removing "V4", the remaining variables will be the same as the variables in the previous best model, therefore the new best model is the same as the previous one, we discard the previous best model and keep the new one, it is named "best12" and has the following formula:

```
best12 = lm(V5~V2+ V8+ V9+ V6, data = biopsy1)
```


Step 12: Enter "V1" and "V6" simultaneously into "best12".
Note that "V6" is already in "best12".

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.22254	0.10260	11.91526	0.00000
V2	0.37622	0.03069	12.25896	0.00000
V8	0.09449	0.02613	3.61594	0.00032
V9	0.19130	0.03545	5.39583	0.00000
V6	0.05626	0.02139	2.62975	0.00874
V1	0.01104	0.02594	0.42579	0.67040

"V1" is insignificant and should be removed from this model, thus we get the new best model which is the same as the previous best model, we discard the previous model and keep the new best model, it is named "best13" and its formula is shown below.

```
best13 = lm(V5~V2+ V8+ V9+ V6, data = biopsy1)
```

This example contains eight predictor variables, performing all the remaining steps by hand is so long, let's summarise how they must be done.

After entering all couples of predictor variables into the best model, we must enter three predictor variables simultaneously, four predictor variables simultaneously, five predictor variables simultaneously, six predictor variables simultaneously, seven predictor variables simultaneously and finally enter eight predictor variables simultaneously.

The best model obtained after entering simultaneously eight predictor variables will be compared to the best model obtained at the preceding step so as to find the final best model.

Example2

Consider the "auction" data of the "yarrr" package, and consider all its variables except "color" for only the purpose of our new method illustration, we have stored the data into the object "auction1". The first few rows of our data are shown below.

cannons	rooms	age	condition	style	jbb	price
18	20	140	5	classic	3976	3502
21	21	93	5	modern	3463	2955
20	18	48	2	classic	3175	3281
24	20	81	5	classic	4463	4400

The any-predictors start-up approach starts by setting the initial best model (first best model) which is determined with aid of the full model. Let's build a full model, only the predictor variables are shown, the dependent variable is "price."

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.85223	82.19512	-0.01037	0.99173
cannons	8.66970	5.65498	1.53311	0.12557
rooms	0.13007	3.64963	0.03564	0.97158
age	0.21668	0.33013	0.65634	0.51176
condition	17.53950	8.23951	2.12871	0.03352
stylemodern	-72.65554	28.62611	-2.53809	0.01130
jbb	0.91383	0.05052	18.08999	0.00000

We will classify these predictor variables into two classes: the first class predictors and the second class predictors. The second class predictors include initially all the predictor variables which are insignificant in the full model. We build another model whose the predictor variables are the variables which are significant in the full model. From above table, such variables are “jbb”, “style” and “condition” considering the significance level of 0.05. The result of the model associated to them is shown below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.08670	70.82437	0.18478	0.85344
jbb	0.97310	0.02023	48.09314	0.00000
stylemodern	-49.15486	21.80775	-2.25401	0.02441
condition	11.41337	6.59490	1.73064	0.08383

We see from above table that the variable “condition” becomes insignificant in this model, we remove it from the model and then move it into the second class predictors, and then we check the significance of the remaining variables as shown below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.01830	70.36719	0.39817	0.69059
jbb	0.98483	0.01908	51.60393	0.00000
stylemodern	-44.97805	21.69547	-2.07315	0.03841

It is seen that the remaining predictor variables remain significant, we proceed to performing “filtering”. The above model whose the predictor variables are only “jbb” and “style” has the adjusted R-squared of 0.7554, the variable “style” is the least significant than “jbb”, then we will remove “style” and check the change in the adjusted R-squared. We have seen that after removing “style”, the adjusted R-squared drops to 0.7546. Since there is a change in the adjusted R-squared, we conclude that both “jbb” and “style” are all important in the above model and we classify them into the first class predictors which are the variables for our first best model, “best1”, it has the following formula:

```
best1 = lm(price~jbb+ style, data = auction1)
```

We will mix the variables of the second class predictors together with the variables of the first class predictors and enter them into the first best model; first, one by one; second, in

couples; third, three predictor variables simultaneously;…; and finally all the predictor variables simultaneously so as to find the final best model. The predictor variables we have are "cannons", "rooms", "age", "condition", "style" and "jbb", we may enter them in any order.

Step 1: Enter the variable "cannons" into the first best model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.33198	70.76135	0.49931	0.61767
jbb	0.96113	0.03077	31.23935	0.00000
stylemodern	-54.54354	23.78160	-2.29352	0.02203
cannons	3.91109	3.98217	0.98215	0.32626

"cannons" is insignificant, we remove it and check the significance of the remaining predictor variables as shown below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.01830	70.36719	0.39817	0.69059
jbb	0.98483	0.01908	51.60393	0.00000
stylemodern	-44.97805	21.69547	-2.07315	0.03841

The remaining predictor variables are all significant, we get then our second best model, its formula is as follows.

```
best2 = lm(price~jbb+ style, data = auction1)
```

We compare this best model to the previous best model with aid of either adjusted R-squared or AIC to select the better one. "best2" and "best1" are the same and thus have the same AIC or the same adjusted R-squared, we discard "best1" and keep "best2". Into "best2" we enter another variable to find a new best model.

Step 2: Enter the variable "rooms" into the previous best model, "best2".

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	53.93320	72.43702	0.74455	0.45672
jbb	0.99611	0.02052	48.55199	0.00000
stylemodern	-40.19962	21.91738	-1.83414	0.06693
rooms	-4.16999	2.79477	-1.49207	0.13600

"rooms" is the most insignificant and should be removed first. After removing it, the new best model will be the same as the previous best model, we discard the previous and keep the new best model, let's name it "best3".

```
best3 = lm(price~jbb+ style, data = auction1)
```

Step 3: Enter the variable "age" into the previous best model, "best3".

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.53035	77.29328	0.14918	0.88144
jbb	0.98476	0.01909	51.58044	0.00000
stylemodern	-44.90532	21.70391	-2.06900	0.03880
age	0.16750	0.32431	0.51649	0.60562

"age" is insignificant, after removing it the new best model will be similar to the previous model, we discard the previous and keep the new best model, let's name it "best4", its formula is as follows.

```
best4 = lm(price~jbb+ style, data = auction1)
```

Step 4: Enter the variable "condition" into the previous best model, "best4".

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.08670	70.82437	0.18478	0.85344
jbb	0.97310	0.02023	48.09314	0.00000
stylemodern	-49.15486	21.80775	-2.25401	0.02441
condition	11.41337	6.59490	1.73064	0.08383

"condition" is insignificant, after removing it the new best model will be the same as the previous best model, then we discard the previous one and keep the new best model, we call it "best5".

```
best5 = lm(price~jbb+ style, data = auction1)
```

The remaining two steps for entering a single predictor variables will cover entering "style" and "jbb" one by one. However, these variables are already in the previous best model, thus after entering them the new best model will be the same as the previous model. So, let's

name “best8”, the new best model that we shall keep after such steps, its formula is the following:

```
best8 = lm(price~jbb+ style, data = auction1)
```

We proceed our model selection method by entering into the previous best model two predictor variables simultaneously, we shall have the couples of variables such as (cannons, rooms), (cannons, age), …, (cannons, jbb), (rooms, age), (rooms, condition), …., (condition, jbb) and finally (style, jbb).

Step 7 Enter simultaneously the variable “cannons” and “rooms” into the previous best model, “best8”.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	53.60814	72.47832	0.73964	0.45969
jbb	0.98663	0.03779	26.10484	0.00000
stylemodern	-44.05107	25.43630	-1.73182	0.08362
cannons	1.35870	4.54787	0.29876	0.76519
rooms	-3.70885	3.19381	-1.16126	0.24581

Three predictor variables are insignificant, we remove them one by one starting from the most insignificant and check the significance of the remaining variables every time we remove one variable. Let’s remove “cannons”.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	53.93320	72.43702	0.74455	0.45672
jbb	0.99611	0.02052	48.55199	0.00000
stylemodern	-40.19962	21.91738	-1.83414	0.06693
rooms	-4.16999	2.79477	-1.49207	0.13600

It is seen that “rooms” is the most insignificant, we remove it also. After removing it, the new best model will be the same as the previous best model, thus we discard the previous and keep the new one, we call it “best9” and its formula is the following.

```
best9 = lm(price~jbb+ style, data = auction1)
```

The remaining steps are so long to be performed by hand in this paper, let’s summarise how they must be done.

Since in this example the predictor variables are seven, after entering the couples of variables, we have to proceed to entering three predictor variables simultaneously, four predictor variables simultaneously, five predictor variables simultaneously, six predictor variables simultaneously and finally seven predictor variables simultaneously.

We did a hard work to perform all such steps and find that the final best model is (the formula):

```
best_final = lm(price~jbb+ style, data = auction1)
```

Example 3

In this example, we use the data, "UScrime", of the package ,"MASS". Note that, the data we shall use are modified by removing outliers, thus if you use that you have in your computer as whole, you may get different answers to ours. The variables "Time", "So" and "M" are also not used in this example. The first few observations of the data are shown below.

	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	GDP	Ineq	Prob	y
5	121	109	101	591	985	18	30	91	20	578	174	0.041399	1234
6	110	118	115	547	964	25	44	84	29	689	126	0.034201	682
10	118	71	68	632	1029	7	15	100	24	526	174	0.044498	705
12	108	75	71	595	972	47	59	83	31	580	172	0.031201	849

We start by building the full model so as to classify the predictor variables and set the first best model. The predictor variables in the full model are shown below, the dependent variable is "Ineq".

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	235.30257	354.27636	0.66418	0.53597
Ed	-1.07173	2.07169	-0.51732	0.62699
Po1	1.01823	3.09671	0.32881	0.75562
Po2	-1.04719	3.06403	-0.34177	0.74642
LF	0.14311	0.34194	0.41852	0.69294
M.F	0.17429	0.49044	0.35537	0.73681
Pop	0.12706	0.80912	0.15704	0.88136
NW	-0.05283	0.61622	-0.08574	0.93500
U1	-0.69683	1.17903	-0.59102	0.58021
U2	1.43941	1.91447	0.75186	0.48600

GDP	-0.37458	0.15325	-2.44432	0.05834
Prob	-81.65746	527.61625	-0.15477	0.88306
y	0.04731	0.07061	0.66998	0.53256

In this example, none of predictor variable is significant in the full model. In order to classify the variables into first and second class predictors, we consider the least insignificant variable and build a model with only such variable, if it is found to be significant, we classify it into the first class predictors, otherwise we consider the next variables one by one starting from the least to the most insignificant until we get one variable which is significant. For the case like this example, the first class predictors always includes one predictor variable.
In this example "GDP" is the least insignificant, let's build a model with only it.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	355.10386	31.58047	11.24441	0e+00
GDP	-0.32942	0.05652	-5.82787	3e-05

"GDP" is significant, we then classify it into the first class predictors, and the remaining variables are classified into the second class predictors. Note that, in this case we do not perform filtering since filtering is done only if the variables are more than one. Our first best model is the model with the variable "GDP", we keep it and we proceed to entering the second class predictors into it. Note also that, the second class predictors must be mixed with the first class predictors so as to enter them into the initial best model. The first best model has the following formula:

```
best1 = lm(Ineq~GDP, data = UScrime1)
```

We now start entering one by one all the predictor variables into "best1". During these steps, we enter them in any order. The variables are the following: Prob, NW, LF, Pop, Ed, y, M.F, U1, U2, GDP, Po1 and Po2.

According to how we list them, the first variable to enter into the first best model is "Prob".

Step 1: Enter the variable "Prob" Into "best1"

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	383.30158	49.48491	7.74583	0.00000
GDP	-0.36228	0.07224	-5.01471	0.00015
Prob	-234.96477	314.34957	-0.74746	0.46634

"Prob" is insignificant and we have to remove it. After removing it, the new best model named, "best2", will be the same as "best1", we discard "best1" and keep "best2". Its formula is the following:

```
best2 = lm(Ineq~ GDP, data = UScrime1)
```

Step 2: Enter the variable "NW" Into "best2"

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	356.15346	29.54160	12.05600	0.00000
GDP	-0.35150	0.05425	-6.47949	0.00001
NW	0.44041	0.24274	1.81432	0.08968

"NW" is insignificant, the new best model at this step is the same as "best2", we discard "best2" and keep the new model, "best3".

```
best3 = lm(Ineq~ GDP, data = UScrime1)
```

Step 3: Enter the variable "LF" Into "best3"

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	217.17935	68.49282	3.17083	0.00633
GDP	-0.30307	0.05208	-5.81986	0.00003
LF	0.21111	0.09546	2.21167	0.04293

All the predictor variables in this model are significant, thus the new best model has the following formula:

```
best4 = lm(Ineq~ GDP+ LF, data = UScrime1)
```

We compare "best4" to "best3"; "best4" has a greater adjusted R-squared, thus we discard "best3" and keep "best4".

Step 4: Enter the variable "Pop" Into "best4"

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	241.10910	61.50340	3.92026	0.00154
GDP	-0.35910	0.05225	-6.87235	0.00001
LF	0.20498	0.08449	2.42603	0.02937
Pop	0.51744	0.22768	2.27261	0.03934

All the predictor variables in this model are significant, thus the new best model has the following formula:

```
best5 = lm(Ineq~ GDP+ LF+ Pop, data = UScrime1)
```

"best5" has a greater adjusted R-squared than "best4", we discard "best4" and keep "best5".

Step 5: Enter the variable "Ed" Into "best5"

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	246.08614	63.31890	3.88646	0.00187
GDP	-0.33104	0.06906	-4.79336	0.00035
LF	0.23874	0.10114	2.36052	0.03454
Pop	0.45669	0.25121	1.81792	0.09219
Ed	-0.34607	0.54014	-0.64072	0.53284

"Ed" and "Pop" are insignificant, but "Ed" is the more insignificant than "Pop", thus "Ed" will be removed first. It is seen that after removing it the new best model named, "best6", will be the same as the previous best model. We discard the previous model and keep the new model.

```
best6 = lm(Ineq~ GDP+ LF + Pop, data = UScrime1)
```

If someone continues running these steps, he/she will find that, all steps involving entering a single predictor will end with the above best model, "best6". So let's skip showing these steps and go to the steps which involve entering two predictor variables simultaneously. We will have many combinations of two variables including (Prob, NW), (LF, Pop), (Ed, y),

etc. Let's show one step by entering the couple (Ed, y) as shown below (only the predictor variables are shown)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	311.61199	61.08508	5.10128	0.00026
GDP	-0.35026	0.05993	-5.84490	0.00008
LF	0.25060	0.08709	2.87740	0.01390
Pop	0.36092	0.21973	1.64253	0.12641
Ed	-1.12332	0.56895	-1.97436	0.07181
y	0.03952	0.01672	2.36417	0.03578

From the table above "Pop" and "Ed" are insignificant, but "Pop" is more insignificant, we remove it and check the significance of the remaining variables. The result is shown below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	310.24594	64.94579	4.77700	0.00036
GDP	-0.29740	0.05375	-5.53293	0.00010
LF	0.28327	0.09016	3.14184	0.00779
Ed	-1.50580	0.55198	-2.72800	0.01725
y	0.04459	0.01747	2.55193	0.02410

we see that the remaining predictors are all significant, thus we get a new best model named "best_x".

```
best_x= lm(Ineq~ GDP+ LF + Ed+ y, data = UScrime1)
```

Comparison between "best_x" and the previous best model "best6" proves "best_x" to be better than "best6", we discard "best6" and keep "best_x".

There are still many steps to perform so as to select the final best model, we choose to stop by here since it is so hard to perform all of them by hand.

Example 4

In this example we use the "movies" data of the "yarrr" package. The first few observations of the data are shown below:

sequel	budget	revenue.all	revenue.dom	revenue.int	revenue.inf
0	425	2783.919	760.5076	2023.411	826.1981
0	200	2207.616	658.6723	1548.943	1139.1828
1	215	1665.444	651.4436	1014.000	651.4436
0	225	1519.480	623.2795	896.200	655.3831

The dependent variable is “revenue.all”. The predictor variables for the full model look as follows:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0	0	2.845955e+01	0.00000
sequel	0	0	1.905745e+01	0.00000
budget	0	0	-4.584897e+01	0.00000
revenue.dom	1	0	5.864871e+14	0.00000
revenue.int	1	0	1.048770e+15	0.00000
revenue.inf	0	0	2.206130e+00	0.02742

We build another model whose the predictor variables are the significant variables found in the full model, such model looks as follows (the dependent variable is still the same):

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0	0	2.885717e+01	0
sequel	0	0	1.891686e+01	0
budget	0	0	-4.681627e+01	0
revenue.dom	1	0	6.666313e+14	0
revenue.int	1	0	1.054866e+15	0

The model formula is:

```
fit = lm(revenue.all~sequel+ budget+ revenue.dom+ revenue.int,data = movies1)
```

This model has the adjusted R-squared equals 1. We proceed to performing filtering by removing some predictor variables starting from the least significant. From the table above, it seems like all the variables have the same p-values, however, the variable “budget” has the least p-value since it has the least test statistic (t-value), thus we remove it first and check the change in the adjusted R-squared. The new model after removing “budget” will have the following formula:

```
fit1 = lm(revenue.all~sequel+ revenue.dom+ revenue.int, data = movies1)
```

This model has the adjusted R-squared equals 1. This means that the variable "budget" is useless when it is added to the variables "sequel", "revenue.dom" and "revenue.int" to explain the variation in the variable "revenue.all" and thus we move it into the second class predictors. We stop filtering when the adjusted R-squared reduces after removing a variable, then we have to continue since it is still 1. In the model "fit1" above, the variable "sequel" is the least significant since it has the least test statistic as shown below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0	0	-6.350512e+01	0
sequel	0	0	-8.190030e+00	0
revenue.dom	1	0	1.677345e+15	0
revenue.int	1	0	2.852259e+15	0

After removing it, we shall have a model which has the following formula:

```
fit2 = lm(revenue.all~revenue.dom+ revenue.int, data = movies1)
```

This model has also the adjusted R-squared equals 1, we continue filtering by removing the variable "revenue.dom". The new model after removing "revenue.dom" has the adjusted R-squared equals 0.722. Since the adjusted R-squared reduces, we stop filtering and take back the removed variable into the model, then the first best model is the model, "fit2", above. This first best model has the R-squared equals one, then it is meaningless to add other predictor variables (the second class predictors) to it since the R-squared cannot exceed one, thus the final best model for our data is the model whose the equation is "fit2" above.

4.2. Second Approach: The Single-predictor start-up

The single-predictor start-up approach differs from the any-predictors start-up approach by the number of the predictor variables that the first best model includes. In the single-predictor start-up approach, the first best model includes only one predictor variable while in the any-predictors start-up approach the first best model includes any number of the predictor variables. The way we set the first best model for these two approaches also differ, other procedures such as entering the second class predictors into the best model and choosing the best model at each step are the same. In the single-predictor start-up

approach, the first best model takes only one predictor variable, this predictor variable is the variable which results in the largest R-squared when the response variable is regressed on each predictor variable in the separated models.

Consider the example 4 above, to apply this approach to that data, we have to regress the response variable, "revenue.all", to each of the predictor variables. The predictor variables are five, then we shall build five models as shown below:

```
fit1 = lm(revenue.all~ sequel, data = movies1)

fit2 = lm(revenue.all~ budget, data = movies1)

fit3 = lm(revenue.all~ revenue.dom, data = movies1)

fit4 = lm(revenue.all~ revenue.int, data = movies1)

fit5 = lm(revenue.all~ revenue.inf, data = movies1)
```

After building them, we compare their R-squared, we have seen that, the model "fit4" has the largest R-squared, therefore the variable "revenue.int" will be the variable for the first best model which we can name for example "best1", its formula is:

```
best1 = lm(revenue.all~ revenue.int, data = movies1)
```

Thereafter, we start entering all predictor variables (including also "revenue.int") into the first best model in the same way as we do for the any-predictors start-up approach. Filtering is not done for the single-predictor start-up, since the first best model has only one predictor variable.

5. Foundation of our New Regression Model Selection Method

The technique for developing this regression model selection method arose after observing a complex relationship between the predictor variables in the multiple regression model. Multicollinearity is the most known cause of the effects between the predictor variables, however, the predictor variables can affect each other even when no multicollinearity is present in the model.

Consider for example this regression model formula and the corresponding variance inflation factor values for the predictor variables ("swiss" data of the "datasets" package):

```
fit = lm(Fertility~., data = swiss)
```

Variance inflation factor results

	x
Agriculture	2.2841
Examination	3.6754
Education	2.7749
Catholic	1.9372
Infant.Mortality	1.1075

The values for the variance inflation factor indicate no multicollinearity between the variables. Let' s also explore the significance of the predictor variables.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	66.91518	10.70604	6.25023	0.00000
Agriculture	-0.17211	0.07030	-2.44814	0.01873
Examination	-0.25801	0.25388	-1.01627	0.31546
Education	-0.87094	0.18303	-4.75849	0.00002
Catholic	0.10412	0.03526	2.95297	0.00519
Infant.Mortality	1.07705	0.38172	2.82157	0.00734

In this model, only the variable "Examination" is insignificant (significance level equals 0.05), and we have noticed that, if it is removed, the remaining variables remain significant. But, let's see what happens when the variable "Education" is removed, the remaining variables behave as follows:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	59.60267	13.04246	4.56989	0.00004
Agriculture	-0.04759	0.08032	-0.59251	0.55669
Examination	-0.96805	0.25284	-3.82865	0.00042
Catholic	0.02611	0.03843	0.67950	0.50055
Infant.Mortality	1.39597	0.46259	3.01771	0.00431

We see from this table that, after removing the variable "Education", the variable "Examination" becomes significant and the variables "Agriculture" and "Catholic" become insignificant, thus we can conclude that "Education" affects "Examination", "Agriculture"

and “Catholic”; in other words, there is a relationship between the variable “Education” and the three variables mentioned above.

Another feature we observed between the predictor variables is that, multicollinearity between two variables can be influenced by another third variable. Consider the following model formula:

```
fit = lm(PerimCh1~ FiberLengthCh1+  FiberWidthCh1, data =
segmentationOriginal130)
```

Variance inflation factor results

	x
FiberLengthCh1	1.81015
FiberWidthCh1	1.81015

The variance inflation factor values for these two variables indicate no multicollinearity, but let’s see what happens when another variable is added to them.

```
fit2 = lm(PerimCh1~ FiberLengthCh1+  FiberWidthCh1+ EqSphereAreaCh1,
          data = segmentationOriginal130)
```

Variance inflation factor results

	x
FiberLengthCh1	14.97006
FiberWidthCh1	11.88061
EqSphereAreaCh1	8.82733

We observe now multicollinearity between all variables, however, the multicollinearity between “FiberLengthCh1” and “FiberWidthCh1” is due to “EqSphereAreaCh1”, in other words, the first two variables (FiberLengthCh1 and FiberWidthCh1) affect each other through the third variable (EqSphereAreaCh1).

After observing these behaviors of predictor variables, we tried to link them to the real life situations. In real life, there exists companies, associations, or other particular groups whose the members work for common interest. However, the members may contribute equally to the development of the group, or some members can contribute more than others. If a disagreement occurs for example between the most contributors and the less contributors, they may separate and thus two groups arise, after separation there will be

always a strong group and a weak group. However, some members of the weak group may choose by their will to come back to join the strong group, if they come and obey the order of the group, they stay in it, if they come and do not obey the order or cause a disorder, they will be sent out again. Not only by the will of the members of the weak group they may join again the strong group, the members of the strong group may also try to convince them to come back to join them again. Our new regression model selection method has been developed referring to these real life relationships.

The multiple regression model behaves as the company, association or other particular group. The predictor variables are analogous to the members of the group. In the significant full model, predictor variables can all be significant, or some variables are significant and others are insignificant. This is similar to the inequality between the contributions of the group members towards the group development. The classification of the predictor variables into the first and the second class predictors is analogous to the separation between the members of the group when a disagreement occurs. In some groups, there are intermediate members, these are the members who do not make the group to move back or to move forward (to advance), if the separation between the group members occurs, the intermediate members stay bound to the strong group, but losing them also cannot weakens the group, this is the reason behind filtering in our new model selection method. After the predictor variables classification, we try many times to enter the second class predictors into the best model, this is compared to the events during which the members of the weak group are coming again to join the strong group either by their will or convinced by the strong group members.

6. Discussion

In this section we present the various features of our new regression model selection method as well as its similarity and discrepancy to some existing model selection methods especially the stepwise regression and the best subsets regression.

- This new regression model selection method operates through two approaches which lead to the same final best model. However, the any-predictors start-up approach is better than the single-predictor start-up approach since the best model can be obtained even in earlier steps.

- This model selection method consists of selecting the important predictor variables, the selection can be done using the usual significance levels of 0.05 or 0.1

- The final best model resulted from our model selection method will never includes insignificant variables relative to the usual significance levels.

- For the same data and the same significance level, the best model resulted from our new model selection method is either the same or better than the best model resulted if the stepwise regression was used.

- This model selection method allows the predictor variables to be tested for inclusion in the best model many times and in different conditions. After setting the first best model, our new model selection method proceeds to entering the predictors into best model in different formats, one variable alone, couple of variables, three variables simultaneously, four variables simultaneously, etc; thus, a variable that has been removed at the preceding steps will always come back to be tested again for inclusion in the next steps.

- This method works by selecting the important variables as does the stepwise, forward selection approach. However, it possesses a strong feature that the stepwise regression fails to have, the new model selection method keeps the best model until a new better one is found to replace it. In the stepwise regression, forward selection approach, every time a predictor variable is selected, a new best model results, but this new model is considered without comparing it to the best model obtained at the preceding step. Furthermore, the process of finding a variable to consider for inclusion can result in more than one best model according to the selection criteria, but one of them is considered without effective comparison to others.

- This new regression model selection method builds in sequence all possible models as does the best subsets regression. However; many steps are similar that some many steps can be skipped. In addition, in the any-predictors start-up approach, the best model can be obtained even in the earlier steps; models comparison is also easy: at each step we compare only two models. Furthermore, according to the chosen significance level; this new method cares always about the significance of the predictor

variables individually, while the best subsets regression cares only about the significance of the whole model.

- The fact that the new method builds all possible models limits it to be used when the number of the predictor variables is large as it is the case for the best subsets regression. But, we have observed that; the best model can be obtained in the earlier steps especially for the any-predictors start-up approach. This has arisen a new idea of determining a specific order the predictor variables can be entered into the first best model so as to pull the final best model without accomplishing all method's steps (all possible models); thus, the future research will cover such matter.

- In section 4 above, we have used four examples to demonstrate the procedures for our new regression model selection method. Let's compare the results we obtain to the results we should obtain if the stepwise regression was applied.

In example 1, we have obtained the best model whose the formula is:

```
best13 = lm(V5~V2+ V8+ V9+ V6, data = biopsy1)
```

Significance of the predictor variables (new method)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.24755	0.08406	14.84040	0.00000
V2	0.37997	0.02938	12.93179	0.00000
V8	0.09541	0.02603	3.66580	0.00027
V9	0.19222	0.03537	5.43506	0.00000
V6	0.05855	0.02069	2.82958	0.00480

If we apply the stepwise regression to the same data (with aid of "caret" package), the best model has the formula:

```
# Using backward elimination approach

best_backward = lm(V5~V2+ V3+ V4+ V6+  V7+ V8+ V9, data = biopsy1)
```

Significance of the predictor variables (backward approach)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.17342	0.09674	12.12951	0.00000
V2	0.29957	0.04517	6.63239	0.00000
V3	0.08404	0.04528	1.85617	0.06386
V4	0.02691	0.02925	0.92010	0.35785
V6	0.03514	0.02311	1.52026	0.12891
V7	0.03526	0.03702	0.95257	0.34115
V8	0.07767	0.02713	2.86240	0.00433
V9	0.18816	0.03570	5.27093	0.00000

```
# Using forward selection approach

best_forward = lm(V5~V2+ V3+ V4+ V6+ V7+ V8+ V9, data = biopsy1) # same as
                                                                "best_backward"

# Using both backwawd and forward approaches

best_both = lm(V5~ V2+ V3+  V6+  V7+  V8+  V9, data = biopsy1)
```

Significance of the predictors (both approaches)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.17392	0.09673	12.13625	0.00000
V2	0.30593	0.04463	6.85494	0.00000
V3	0.08471	0.04527	1.87125	0.06174
V6	0.04085	0.02226	1.83516	0.06692
V7	0.04125	0.03644	1.13195	0.25806
V8	0.07954	0.02705	2.94029	0.00339
V9	0.19267	0.03536	5.44943	0.00000

The comparison between these three models is given in the table below.

Prediction Performance Comparison

X	adj.r.squared	sigma	AIC
best13 (new method)	0.60305	1.40063	2405.484
best_backward	0.60474	1.39765	2405.553
best_both	0.60483	1.39750	2404.409

With aid of the table above; consider the models "best_both" and "best13 (new method)", the best model resulted from the stepwise regression has a lower AIC. However, the adjusted R-squared difference is 0.00178, the model sigma (Residual Standard Error, RSE) difference is 0.00313 and the AIC difference is 1.075, such difference values are all small that the three models are almost the same. In addition, the best model we have obtained using our new method has addition good feature: it does not contain insignificant variables while they are present in the best models resulted from the stepwise regression considering the significance level of 0.05. We also observed that, if insignificant predictor variables are removed from the model "best_both", the remaining significant predictors (significance level equals 0.05) form a model which is the same as the model obtained using the new method.
Recall that, the best model according to our new method has been obtained after 13 steps, thus we hope to get a better model than that one above after performing all steps, otherwise, the best model will be what we obtain.

In example 2, we have obtained the best model which has the following formula:

```
best_final = lm(price~jbb+ style, data = auction1)
```

Significance of the predictor variables (new method)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.01830	70.36719	0.39817	0.69059
jbb	0.98483	0.01908	51.60393	0.00000
stylemodern	-44.97805	21.69547	-2.07315	0.03841

Applying the stepwise regression to the same data (with aid of "caret" package), the best model has the following formula:

```
#Using both backward and forward approaches

best_both = lm(price~ jbb, data = auction1)
```

```
#Using forward approach

best_forward = lm(price~ jbb, data = auction1) # same as "best_both"

# Using backward elimination approach

best_backward = lm(price~ cannons+ condition+ style+ jbb, data = auction1)
```

Significance of the predictor variables (backward approach)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	21.00176	70.85512	0.29640	0.76698
cannons	8.31988	4.38580	1.89700	0.05812
condition	17.25548	7.27075	2.37327	0.01782
stylemodern	-71.64105	24.79610	-2.88921	0.00395
jbb	0.91668	0.03596	25.49342	0.00000

Let' summarise the comparison between these models through the following table.

Prediction Performance Comparison

X	adj.r.squared	sigma	AIC
best_final (new method)	0.75536	323.1456	14399.08
best_backward	0.75649	322.4024	14396.47
best_both	0.75456	323.6791	14401.38

Consider the models "best_backward" and "best_final (new method)", the best model resulted from the stepwise regresion has a lower AIC. However, the adjusted R-squared difference is 0.00113, the model sigma (Residual Standard Error, RSE) difference is 0.7432 and the AIC difference is 2.61, such difference values are all small that the three models are almost the same. Note that, in the model "best_backward", the variable "cannons" is insignificant, the significance level = 0.05; if removed, the variable "condition" also becomes insignificant, if "condition" is removed, we get the best model which is similar to the best model obtained using our new method.

In example 3, to the step we were, the best model have had the following formula:

```
best_x= lm(Ineq~ GDP+ LF + Ed+ y, data = UScrime1)
```

Significance of the predictor variables (new method)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	310.24594	64.94579	4.77700	0.00036
GDP	-0.29740	0.05375	-5.53293	0.00010
LF	0.28327	0.09016	3.14184	0.00779
Ed	-1.50580	0.55198	-2.72800	0.01725
y	0.04459	0.01747	2.55193	0.02410

while backward elimination approach, forward selection approach and both approaches (with aid of "caret" package) resulted in the same best model which has the following formula:

```
best_stepwise = lm(Ineq~ GDP, data = UScrime1)
```

The comparison between such models is summarised in the table below:

Prediction Performance

X	adj.r.squared	sigma	AIC
best_x (new method)	0.8131	11.4175	144.8895
best_stepwise	0.6598	15.4066	153.4143

The best model resulted from our new method has a lower AIC and the AIC difference between such two models is 8.5248. Recall that the model "best_x (new method)" has been obtained without accomplishing all method's steps, if all steps were accomplished we should have another model better than it, otherwise it is still the same.

Note:

using the stepAIC() function for the stepwise regression, backward elimination approach and both approaches result in the same best model whose the formula is:

```
best_both = lm(formula = Ineq ~ GDP+ LF + Ed+ y+ U1 + U2, data = UScrime1)
```

Significance of the Predictor Variables

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	358.42309	88.70662	4.04054	0.00195
GDP	-0.35523	0.06143	-5.78224	0.00012
LF	0.27597	0.09257	2.98128	0.01249

Ed	-1.57822	0.54214	-2.91110	0.01416
y	0.05540	0.01886	2.93661	0.01353
U1	-0.48029	0.27687	-1.73472	0.11068
U2	1.14869	0.64252	1.78777	0.10136

From above table it is seen that “U1” and “U2” are insignificant. If we remove “U1”, “U2” remains insignificant; if “U2” is removed, the remaining predictors remain all significant which form the same model as the best model obtained using our new regression model selection method. In addition, the best model according to the new method has been obtained during the steps of entering one variable into the first best model, if all steps were accomplished we should have another model better than it, otherwise it is still the same.

In example 4, the best model obtained using our new method has the following formula:

```
fit2 = lm(revenue.all~revenue.dom+ revenue.int, data = movies1)
```

while backward elimination approach, forward selection approach and both approaches applied to the same data result in the same best model whose the formula is:

```
best_stepwise = lm(revenue.all~ revenue.dom+ revenue.int, data = movies1)
```

The models are the same.

7. Conclusion

The new regression model selection presented in this paper selects the best model from all possible models and the best model can be obtained even in earlier steps of the variable selection processes, therefore it is not always necessary to accomplish all the required steps especially when the number of the predictor variables is too large. Two alternative approaches can be used for this new model selection method, they are the any-predictors start-up approach and the single-predictor start-up approach, and both approaches lead to the same result when applied to the same data. The variables selection processes in both approaches is based on the statistical significance of the predictor variables, the significance level can be the usual values of 0.05 or 0.1 and the best model can never include the insignificant variables according to the usual significance levels. The any-predictors start-up approach is better than the single-predictor start-up approach since the best model can be obtained in earlier steps, thus the next paper will focus on the effective

order through which the predictor variables can be entered into the first best model so as to get the final best model without accomplishing all the method's steps and with no error.

Acknowledgments: I thank Dr. Evariste MINANI and Prof. L. L. Yadav, lecturers at University of Rwanda College of Education; Antoine MAHORO and Ndihokubwayo Kizito, PhD students; for their guidance and support. I thank also Prof. Brian Caffo, Johns Hopkins Univeristy for his lectures available on Coursera.

References

1. Joseph B. Kadane and Nicole A. Lazar (2004). Methods and Criteria for Model Selection.
   Corresponding Link

2. Loann David Denis Desboulets (2018). A Review on Variable Selection in Regression Analysis.
   Corresponding Link

3. T. Hastie, R. Tibshirani and R. Tibshirani (2018). Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons.
   Corresponding Link

4. Frank Emmert-Streib and Matthias Dehmer (2019). Evaluation of Regression Models: Model Assessment, Model Selection and Generalization Error.
   Corresponding Link

5. Obubu Maxwell (2019). Comparison of Some Variable Selection Techniques in Regression Analysis. Am J Biomed Sci & Res. 2019 - 6(4). AJBSR.MS.ID.001044. DOI: 10.34297/AJBSR.2019.06.001044.
   Corresponding Link

6. RANJIT KUMAR PAUL (2008). MULTICOLLINEARITY: CAUSES, EFFECTS AND REMEDIES. Corresponding Link

7. Ali BAGER et al. (2017). ADDRESSING MULTICOLLINEARITY IN REGRESSION MODELS: A RIDGE REGRESSION APPLICATION.
Corresponding Link

8. Riccardo Parviero (2017). Improving ridge regression via model selection and focussed fine-tuning.
Corresponding Link

9. Maya Lozinski (2018). Best Subset Selection: Some Recommendations for Practitioners.
Corresponding Link

10. Zhang Z. Variable selection with stepwise and best subset approaches. Ann Transl Med 2016;4(7):136. doi: 10.21037/atm.2016.03.35
Corresponding Link

11. Gary Smith (2018). Step away from stepwise.
Corresponding Link

12. Olga Morozova et al. (2015). Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia
Corresponding Link

13. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biometrical Journal. 2018; 60: 431-449.
https://doi.org/10.1002/bimj.201700067
other Link