Exploring Common Variable Selection Approaches

Table Of Contents

Introduction:

The present report has been prepared by performing various tasks on the 'Parkinsons' dataset using the R platform. The dataset is imported onto R and named 'park'. The purpose of this report is to understand the various selection methods, the effects of having many predictors and multicollinearity.

Before performing the tasks, the 'park' dataset has been cleaned and named 'parkclean'. The dimension of the 'park' and the 'parkclean' datasets is as follows:

```
> dim(park)
[1] 5875   25
> dim(parkclean)
[1] 5869   25
```

1. **Variable selection method I:**

   a. Backward Elimination:

      Backward stepwise regression is a step-by-step iterative construction of a regression model where a **reduced final model is built from a full model**. The method at each step removes a variable from the full model that does not have a significant impact on the dependent variable. The significance of a variable here is determined based on the **exit criteria** set before the start of the process. Some of the parameters used to define the criteria are p-value, AIC, BIC, F-value, etc. This process is repeated until all the variables which are not significant are removed leaving behind the final model.

   b. Backward Model:

      A final model is built on R using the backward elimination method, where the dependent variable is 'total_UPDRS' and all other variables are independent. All the variables in the parkinsons dataset are included except for sex, test_time_hr and test_time_min. The variables **'test_time_hr' and 'test_time_min' are excluded** as they are mathematical derivations of test_time and including them **leads to singularities** in the model. The 'olsrr' package has been used to implement the backward elimination.

The **exit criteria used is p-value which is set to 0.05**. So, at any step, the variable is checked for significance against this critical value.

On implementing the r code, the **following result is generated**:

- Elimination Steps:

```
Backward Elimination: Step 1

 Variable Jitter.DDP Removed

                        Model Summary
--------------------------------------------------------------
R                       0.953       RMSE                3.255
R-Squared               0.908       Coef. Var          11.218
Adj. R-Squared          0.908       MSE                10.596
Pred R-Squared          0.907       MAE                 2.431
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
--------------------------------------------------------------------------
                 Sum of
                Squares        DF    Mean Square       F         Sig.
--------------------------------------------------------------------------
Regression   610082.234        19      32109.591    3030.461    0.0000
Residual      61920.752      5844         10.596
Total        672002.986      5863
--------------------------------------------------------------------------
```

```
Backward Elimination: Step 2

 Variable Shimmer.dB Removed

                        Model Summary
--------------------------------------------------------------
R                       0.953       RMSE                3.255
R-Squared               0.908       Coef. Var          11.218
Adj. R-Squared          0.908       MSE                10.594
Pred R-Squared          0.907       MAE                 2.431
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
--------------------------------------------------------------------------
                 Sum of
                Squares        DF    Mean Square       F         Sig.
--------------------------------------------------------------------------
Regression   610080.294        18      33893.350    3199.257    0.0000
Residual      61922.692      5845         10.594
Total        672002.986      5863
--------------------------------------------------------------------------
```

```
Backward Elimination: Step 3

 Variable Shimmer.DDA Removed

                        Model Summary
--------------------------------------------------------------
R                       0.953       RMSE                3.255
R-Squared               0.908       Coef. Var          11.217
Adj. R-Squared          0.908       MSE                10.594
Pred R-Squared          0.907       MAE                 2.431
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
--------------------------------------------------------------------------
                 Sum of
                Squares        DF    Mean Square       F         Sig.
--------------------------------------------------------------------------
Regression   610071.701        17      35886.571    3387.511    0.0000
Residual      61931.285      5846         10.594
Total        672002.986      5863
--------------------------------------------------------------------------
```

```
Backward Elimination: Step 4

 Variable Shimmer.APQ3 Removed

                        Model Summary
--------------------------------------------------------------
R                       0.953       RMSE                3.255
R-Squared               0.908       Coef. Var          11.217
Adj. R-Squared          0.908       MSE                10.593
Pred R-Squared          0.907       MAE                 2.431
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
--------------------------------------------------------------------------
                 Sum of
                Squares        DF    Mean Square       F         Sig.
--------------------------------------------------------------------------
Regression   610065.707        16      38129.107    3599.462    0.0000
Residual      61937.279      5847         10.593
Total        672002.986      5863
--------------------------------------------------------------------------
```

```
Backward Elimination: Step 5

 Variable Jitter.PPQ5 Removed

                        Model Summary
--------------------------------------------------------------
R                       0.953       RMSE                3.255
R-Squared               0.908       Coef. Var          11.217
Adj. R-Squared          0.908       MSE                10.593
Pred R-Squared          0.907       MAE                 2.430
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
--------------------------------------------------------------------------
                 Sum of
                Squares        DF    Mean Square       F         Sig.
--------------------------------------------------------------------------
Regression   610052.439        15      40670.163    3839.177    0.0000
Residual      61950.547      5848         10.593
Total        672002.986      5863
--------------------------------------------------------------------------
```

```
Backward Elimination: Step 6

 Variable NHR Removed

                        Model Summary
--------------------------------------------------------------
R                       0.953       RMSE                3.255
R-Squared               0.908       Coef. Var          11.217
Adj. R-Squared          0.908       MSE                10.593
Pred R-Squared          0.907       MAE                 2.431
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
--------------------------------------------------------------------------
                 Sum of
                Squares        DF    Mean Square       F         Sig.
--------------------------------------------------------------------------
Regression   610043.669        14      43574.548    4113.466    0.0000
Residual      61959.317      5849         10.593
Total        672002.986      5863
--------------------------------------------------------------------------
```

- Elimination summary:

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Jitter.DDP	0.9079	0.9076	19.0993	30504.8964	3.2551
2	Shimmer.dB	0.9079	0.9076	17.2823	30503.0801	3.2549
3	Shimmer.DDA	0.9078	0.9076	16.0932	30501.8938	3.2548
4	Shimmer.APQ3	0.9078	0.9076	14.6589	30500.4613	3.2547
5	Jitter.PPQ5	0.9078	0.9076	13.9108	30499.7173	3.2548
6	NHR	0.9078	0.9076	12.7385	30498.5475	3.2547

- Final model summary:

R	0.953	RMSE	3.255
R-Squared	0.908	Coef. Var	11.217
Adj. R-Squared	0.908	MSE	10.593
Pred R-Squared	0.907	MAE	2.431

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	610043.669	14	43574.548	4113.466	0.0000
Residual	61959.317	5849	10.593		
Total	672002.986	5863			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	1.911	1.046		1.828	0.068	-0.139	3.961
age	0.068	0.005	0.056	13.330	0.000	0.058	0.078
sex	-1.403	0.102	-0.061	-13.692	0.000	-1.604	-1.202
test_time	0.003	0.001	0.013	3.180	0.001	0.001	0.004
motor_UPDRS	1.226	0.006	0.931	215.774	0.000	1.215	1.238
Jitter	-271.965	52.637	-0.143	-5.167	0.000	-375.152	-168.778
Jitter.Abs	14032.325	3087.240	0.047	4.545	0.000	7980.192	20084.457
Jitter.RAP	385.396	87.157	0.113	4.422	0.000	214.537	556.255
Shimmer	-47.224	10.465	-0.114	-4.512	0.000	-67.740	-26.708
Shimmer.APQ5	90.382	16.251	0.141	5.562	0.000	58.523	122.241
Shimmer.APQ11	-33.614	6.833	-0.063	-4.919	0.000	-47.010	-20.218
HNR	-0.098	0.023	-0.039	-4.318	0.000	-0.143	-0.054
RPDE	3.219	0.603	0.030	5.337	0.000	2.036	4.401
DFA	-2.123	0.709	-0.014	-2.994	0.003	-3.514	-0.733
PPE	-3.812	0.954	-0.033	-3.995	0.000	-5.682	-1.941

c. Model equation:

The **final model after 6 steps of backward elimination consists of 14 independent variables**. A total of 6 variables are eliminated from the full model and they are 'Jitter.DDP', 'Shimmer.dB', 'Shimmer.DDA', 'Shimmer.APQ3', 'Jitter.PPQ5' and 'NHR'.

The **equation for the model** is written below based on the beta values generated from the model summary:

total_UPDRS = 1.911 + 0.068***age** – 1.403***sex** + 0.003***test_time** + 1.226***motor_UPDRS** – 271.965***Jitter** + 14032.325* **Jitter.Abs** + 385.396* **Jitter.RAP** - 47.224***Shimmer** + 90.382***Shimmer.APQ5** - 33.614***Shimmer.APQ11** - 0.098* **HNR** + 3.219***RPDE** - 2.123***DFA** – 3.812***PPE**

2. Variable selection method II:

a. Forward Selection:

Forward variable selection is a step-by-step iterative construction of a regression model where the **final model is built from a null model** (model with no variables) with only an intercept. The forward selection method at each step adds the most significant variable that gives the best improvement to the model. The significance of a variable here is determined based on the **entry criteria** set before the start of the process. Some of the parameters used to define the criteria are p-value, AIC, BIC, Fvalue, etc.

This process of adding variables and testing at each step continues as long as the model improves. The process stops once the model is no longer improving on adding more variables thus leading us to the final model.

b. Forward Model:

A final model is built on R using the Forward selection method, where the dependent variable is 'total_UPDRS' and all other variables are independent. All the variables in the parkinsons dataset are included except for test_time_hr and test_time_min. The reason is the same as stated in the previous section for not excluding the above two variables. The 'olsrr' package has been used to implement the Forward selection. The **entry criteria used is p-value which is set to 0.05**. So, at any step, the variable is checked for significance against this critical value.

On implementing the r code, the **following result is generated**:

Selection Steps:

```
Forward Selection: Step 1

+ motor_UPDRS

                    Model Summary
------------------------------------------------------
R                  0.947    RMSE          3.432
R-Squared          0.897    Coef. Var    11.827
Adj. R-Squared     0.897    MSE          11.778
Pred R-Squared     0.897    MAE           2.545
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

```
Forward Selection: Step 2

+ sex

                    Model Summary
------------------------------------------------------
R                  0.950    RMSE          3.353
R-Squared          0.902    Coef. Var    11.557
Adj. R-Squared     0.902    MSE          11.246
Pred R-Squared     0.902    MAE           2.423
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

```
Forward Selection: Step 3

+ age

                    Model Summary
------------------------------------------------------
R                  0.951    RMSE          3.309
R-Squared          0.905    Coef. Var    11.405
Adj. R-Squared     0.904    MSE          10.951
Pred R-Squared     0.904    MAE           2.440
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

```
Forward Selection: Step 4

+ RPDE

                    Model Summary
------------------------------------------------------
R                  0.951    RMSE          3.301
R-Squared          0.905    Coef. Var    11.375
Adj. R-Squared     0.905    MSE          10.894
Pred R-Squared     0.905    MAE           2.436
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

```
Forward Selection: Step 5

+ Shimmer.APQ11

                    Model Summary
------------------------------------------------------
R                  0.952    RMSE          3.290
R-Squared          0.906    Coef. Var    11.337
Adj. R-Squared     0.906    MSE          10.822
Pred R-Squared     0.905    MAE           2.434
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

```
Forward Selection: Step 6

+ Shimmer.APQ5

                    Model Summary
------------------------------------------------------
R                  0.952    RMSE          3.280
R-Squared          0.906    Coef. Var    11.303
Adj. R-Squared     0.906    MSE          10.757
Pred R-Squared     0.906    MAE           2.433
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

```
Forward Selection: Step 7

+ PPE

                    Model Summary
------------------------------------------------------
R                  0.952    RMSE          3.276
R-Squared          0.906    Coef. Var    11.292
Adj. R-Squared     0.906    MSE          10.735
Pred R-Squared     0.906    MAE           2.429
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

```
Forward Selection: Step 8

+ HNR

                    Model Summary
------------------------------------------------------
R                  0.952    RMSE          3.271
R-Squared          0.907    Coef. Var    11.272
Adj. R-Squared     0.907    MSE          10.697
Pred R-Squared     0.906    MAE           2.428
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

```
Forward Selection: Step 9

+ Jitter.Abs

                    Model Summary
------------------------------------------------------
R                  0.952    RMSE          3.268
R-Squared          0.907    Coef. Var    11.264
Adj. R-Squared     0.907    MSE          10.681
Pred R-Squared     0.907    MAE           2.423
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

```
Forward Selection: Step 10

+ Shimmer

                    Model Summary
------------------------------------------------------
R                  0.952    RMSE          3.265
R-Squared          0.907    Coef. Var    11.254
Adj. R-Squared     0.907    MSE          10.663
Pred R-Squared     0.907    MAE           2.422
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

```
Forward Selection: Step 11

+ test_time

                    Model Summary
------------------------------------------------------
R                  0.953    RMSE          3.263
R-Squared          0.907    Coef. Var    11.247
Adj. R-Squared     0.907    MSE          10.649
Pred R-Squared     0.907    MAE           2.425
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

```
Forward Selection: Step 12

+ Jitter

                    Model Summary
------------------------------------------------------
R                  0.953    RMSE          3.262
R-Squared          0.907    Coef. Var    11.241
Adj. R-Squared     0.907    MSE          10.638
Pred R-Squared     0.907    MAE           2.422
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

```
Forward Selection: Step 13

+ Jitter.RAP

                    Model Summary
------------------------------------------------------
R                  0.953    RMSE          3.257
R-Squared          0.908    Coef. Var    11.225
Adj. R-Squared     0.907    MSE          10.608
Pred R-Squared     0.907    MAE           2.430
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

```
Forward Selection: Step 14

+ DFA

                    Model Summary
------------------------------------------------------
R                  0.953    RMSE          3.255
R-Squared          0.908    Coef. Var    11.217
Adj. R-Squared     0.908    MSE          10.593
Pred R-Squared     0.907    MAE           2.431
------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

Selection summary:

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	motor_UPDRS	0.8972	0.8972	655.5125	31133.6489	3.4319
2	sex	0.9019	0.9019	361.7038	30837.1672	3.3535
3	age	0.9045	0.9045	199.5796	30682.3511	3.3092
4	RPDE	0.9050	0.9050	169.1757	30652.8881	3.3006
5	Shimmer.APQ11	0.9057	0.9056	130.1451	30614.8051	3.2897
6	Shimmer.APQ5	0.9062	0.9061	95.2202	30580.4973	3.2798
7	PPE	0.9065	0.9063	83.9762	30569.4197	3.2764
8	HNR	0.9068	0.9067	64.3753	30550.0274	3.2707
9	Jitter.Abs	0.9070	0.9068	56.3739	30542.0960	3.2682
10	Shimmer	0.9071	0.9070	47.5117	30533.2908	3.2655
11	test_time	0.9073	0.9071	40.7418	30526.5523	3.2633
12	Jitter	0.9074	0.9072	35.5771	30521.4038	3.2616
13	Jitter.RAP	0.9077	0.9075	19.6963	30505.5249	3.2569
14	DFA	0.9078	0.9076	12.7385	30498.5475	3.2547

- Model summary:

R	0.953	RMSE	3.255
R-Squared	0.908	Coef. Var	11.217
Adj. R-Squared	0.908	MSE	10.593
Pred R-Squared	0.907	MAE	2.431

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	610043.669	14	43574.548	4113.466	0.0000
Residual	61959.317	5849	10.593		
Total	672002.986	5863			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	1.911	1.046		1.828	0.068	-0.139	3.961
motor_UPDRS	1.226	0.006	0.931	215.774	0.000	1.215	1.238
sex	-1.403	0.102	-0.061	-13.692	0.000	-1.604	-1.202
age	0.068	0.005	0.056	13.330	0.000	0.058	0.078
RPDE	3.219	0.603	0.030	5.337	0.000	2.036	4.401
Shimmer.APQ11	-33.614	6.833	-0.063	-4.919	0.000	-47.010	-20.218
Shimmer.APQ5	90.382	16.251	0.141	5.562	0.000	58.523	122.241
PPE	-3.812	0.954	-0.033	-3.995	0.000	-5.682	-1.941
HNR	-0.098	0.023	-0.039	-4.318	0.000	-0.143	-0.054
Jitter.Abs	14032.325	3087.240	0.047	4.545	0.000	7980.192	20084.457
Shimmer	-47.224	10.465	-0.114	-4.512	0.000	-67.740	-26.708
test_time	0.003	0.001	0.013	3.180	0.001	0.001	0.004
Jitter	-271.965	52.637	-0.143	-5.167	0.000	-375.152	-168.778
Jitter.RAP	385.396	87.157	0.113	4.422	0.000	214.537	556.255
DFA	-2.123	0.709	-0.014	-2.994	0.003	-3.514	-0.733

c. Model equation:

The **final model after 14 steps of forward selection consists of 14 independent variables**. A total of 6 variables are not included and they are ‘Jitter.DDP’, ‘Shimmer.dB’, ‘Shimmer.DDA’, ‘Shimmer.APQ3’, ‘Jitter.PPQ5’ and ‘NHR’.

The **equation for the model** is written below based on the beta values generated from the model summary:

total_UPDRS = 1.911 + 0.068***age** – 1.403***sex** + 0.003***test_time** + 1.226***motor_UPDRS** – 271.965***Jitter** + 14032.325* **Jitter.Abs** + 385.396* **Jitter.RAP** - 47.224***Shimmer** + 90.382***Shimmer.APQ5** - 33.614***Shimmer.APQ11** - 0.098* **HNR** + 3.219***RPDE** - 2.123***DFA** – 3.812***PPE**

3. Variable selection method III:

a. Stepwise Selection:

Stepwise regression is a **combination of 'forward selection' and 'backward elimination'**. This method **is more flexible** compared to the above two methods. Like the above two methods, this method too is a step-by-step iterative construction. At each step first a variable is added based on pre-defined entry criteria and then all the variables already added in the model are checked for their significance against predefined exit criteria. If any variable in the added list is found to be nonsignificant, it is removed. Hence there are **two conditions set, one for entry and one for the exit**. These steps are continued until no more variables can be added or removed thus leading us to a final model.

b. Stepwise Model:

A final model is built on R using the stepwise regression method, where the dependent variable is 'total_UPDRS' and all other variables are independent. All the variables in the parkinsons dataset are included except for test_time_hr and test_time_min. The 'olsrr' package has been used to implement the Forward selection. The **entry and exit criteria used is p-value which is set to 0.05**. So, at any step, the variable is checked for significance against this critical value upon adding and for the significance of the variables already present in the model. If, at any step, a variable is found to be insignificant after adding, it is removed from the model.

On implementing the r code, the **following result is generated**:

- <u>Stepwise selection/elimination steps:</u>

```
Stepwise Selection: Step 1

+ motor_UPDRS

                        Model Summary
--------------------------------------------------------------
R                       0.947       RMSE                3.432
R-Squared               0.897       Coef. Var          11.827
Adj. R-Squared          0.897       MSE                11.778
Pred R-Squared          0.897       MAE                 2.545
```

```
Stepwise Selection: Step 2

+ sex

                        Model Summary
--------------------------------------------------------------
R                       0.950       RMSE                3.353
R-Squared               0.902       Coef. Var          11.557
Adj. R-Squared          0.902       MSE                11.246
Pred R-Squared          0.902       MAE                 2.423
```

```
Stepwise Selection: Step 3

+ age

                        Model Summary
--------------------------------------------------------------
R                       0.951       RMSE                3.309
R-Squared               0.905       Coef. Var          11.405
Adj. R-Squared          0.904       MSE                10.951
Pred R-Squared          0.904       MAE                 2.440
```

```
Stepwise Selection: Step 4

+ RPDE

                        Model Summary
--------------------------------------------------------------
R                       0.951       RMSE                3.301
R-Squared               0.905       Coef. Var          11.375
Adj. R-Squared          0.905       MSE                10.894
Pred R-Squared          0.905       MAE                 2.436
```

```
Stepwise Selection: Step 5

+ Shimmer.APQ11

                        Model Summary
--------------------------------------------------------------
R                       0.952       RMSE                3.290
R-Squared               0.906       Coef. Var          11.337
Adj. R-Squared          0.906       MSE                10.822
Pred R-Squared          0.905       MAE                 2.434
```

```
Stepwise Selection: Step 6

+ Shimmer.APQ5

                        Model Summary
--------------------------------------------------------------
R                       0.952       RMSE                3.280
R-Squared               0.906       Coef. Var          11.303
Adj. R-Squared          0.906       MSE                10.757
Pred R-Squared          0.906       MAE                 2.433
```

```
Stepwise Selection: Step 7

+ PPE

                        Model Summary
--------------------------------------------------------------
R                       0.952       RMSE                3.276
R-Squared               0.906       Coef. Var          11.292
Adj. R-Squared          0.906       MSE                10.735
Pred R-Squared          0.906       MAE                 2.429
```

```
Stepwise Selection: Step 8

+ HNR

                        Model Summary
--------------------------------------------------------------
R                       0.952       RMSE                3.271
R-Squared               0.907       Coef. Var          11.272
Adj. R-Squared          0.907       MSE                10.697
Pred R-Squared          0.906       MAE                 2.428
```

```
Stepwise Selection: Step 9

+ Jitter.Abs

                        Model Summary
--------------------------------------------------------------
R                       0.952       RMSE                3.268
R-Squared               0.907       Coef. Var          11.264
Adj. R-Squared          0.907       MSE                10.681
Pred R-Squared          0.907       MAE                 2.423
```

```
Stepwise Selection: Step 10

+ Shimmer

                        Model Summary
--------------------------------------------------------------
R                       0.952       RMSE                3.265
R-Squared               0.907       Coef. Var          11.254
Adj. R-Squared          0.907       MSE                10.663
Pred R-Squared          0.907       MAE                 2.422
```

```
Stepwise Selection: Step 11

+ test_time

                        Model Summary
--------------------------------------------------------------
R                       0.953       RMSE                3.263
R-Squared               0.907       Coef. Var          11.247
Adj. R-Squared          0.907       MSE                10.649
Pred R-Squared          0.907       MAE                 2.425
```

```
Stepwise Selection: Step 12

+ Jitter

                        Model Summary
--------------------------------------------------------------
R                       0.953       RMSE                3.262
R-Squared               0.907       Coef. Var          11.241
Adj. R-Squared          0.907       MSE                10.638
Pred R-Squared          0.907       MAE                 2.422
```

```
Stepwise Selection: Step 13

+ Jitter.RAP

                        Model Summary
--------------------------------------------------------------
R                       0.953       RMSE                3.257
R-Squared               0.908       Coef. Var          11.225
Adj. R-Squared          0.907       MSE                10.608
Pred R-Squared          0.907       MAE                 2.430
--------------------------------------------------------------
```

```
Stepwise Selection: Step 14

+ DFA

                        Model Summary
--------------------------------------------------------------
R                       0.953       RMSE                3.255
R-Squared               0.908       Coef. Var          11.217
Adj. R-Squared          0.908       MSE                10.593
Pred R-Squared          0.907       MAE                 2.431
--------------------------------------------------------------
```

- Stepwise summary:

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	motor_UPDRS	addition	0.897	0.897	655.5130	31133.6489	3.4319
2	sex	addition	0.902	0.902	361.7040	30837.1672	3.3535
3	age	addition	0.905	0.904	199.5800	30682.3511	3.3092
4	RPDE	addition	0.905	0.905	169.1760	30652.8881	3.3006
5	Shimmer.APQ11	addition	0.906	0.906	130.1450	30614.8051	3.2897
6	Shimmer.APQ5	addition	0.906	0.906	95.2200	30580.4973	3.2798
7	PPE	addition	0.906	0.906	83.9760	30569.4197	3.2764
8	HNR	addition	0.907	0.907	64.3750	30550.0274	3.2707
9	Jitter.Abs	addition	0.907	0.907	56.3740	30542.0960	3.2682
10	Shimmer	addition	0.907	0.907	47.5120	30533.2908	3.2655
11	test_time	addition	0.907	0.907	40.7420	30526.5523	3.2633
12	Jitter	addition	0.907	0.907	35.5770	30521.4038	3.2616
13	Jitter.RAP	addition	0.908	0.907	19.6960	30505.5249	3.2569
14	DFA	addition	0.908	0.908	12.7380	30498.5475	3.2547

- Model summary:

R	0.953	RMSE	3.255
R-Squared	0.908	Coef. Var	11.217
Adj. R-Squared	0.908	MSE	10.593
Pred R-Squared	0.907	MAE	2.431

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	610043.669	14	43574.548	4113.466	0.0000
Residual	61959.317	5849	10.593		
Total	672002.986	5863			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	1.911	1.046		1.828	0.068	-0.139	3.961
motor_UPDRS	1.226	0.006	0.931	215.774	0.000	1.215	1.238
sex	-1.403	0.102	-0.061	-13.692	0.000	-1.604	-1.202
age	0.068	0.005	0.056	13.330	0.000	0.058	0.078
RPDE	3.219	0.603	0.030	5.337	0.000	2.036	4.401
Shimmer.APQ11	-33.614	6.833	-0.063	-4.919	0.000	-47.010	-20.218
Shimmer.APQ5	90.382	16.251	0.141	5.562	0.000	58.523	122.241
PPE	-3.812	0.954	-0.033	-3.995	0.000	-5.682	-1.941
HNR	-0.098	0.023	-0.039	-4.318	0.000	-0.143	-0.054
Jitter.Abs	14032.325	3087.240	0.047	4.545	0.000	7980.192	20084.457
Shimmer	-47.224	10.465	-0.114	-4.512	0.000	-67.740	-26.708
test_time	0.003	0.001	0.013	3.180	0.001	0.001	0.004
Jitter	-271.965	52.637	-0.143	-5.167	0.000	-375.152	-168.778
Jitter.RAP	385.396	87.157	0.113	4.422	0.000	214.537	556.255
DFA	-2.123	0.709	-0.014	-2.994	0.003	-3.514	-0.733

c. Model equation:

The **final model after 14 steps of stepwise selection consists of 14 independent variables**. A total of 6 variables are not selected and they are ‘Jitter.DDP’, ‘Shimmer.dB’, ‘Shimmer.DDA’, ‘Shimmer.APQ3’, ‘Jitter.PPQ5’ and ‘NHR’.

The **equation for the model** is written below based on the beta values generated from the model summary:

> **total_UPDRS** = 1.911 + 0.068***age** – 1.403***sex** + 0.003***test_time** + 1.226***motor_UPDRS** – 271.965***Jitter** + 14032.325* **Jitter.Abs** + 385.396* **Jitter.RAP** - 47.224***Shimmer** + 90.382***Shimmer.APQ5** - 33.614***Shimmer.APQ11** - 0.098* **HNR** + 3.219***RPDE** - 2.123***DFA** – 3.812***PPE**

4. Comparing the models:

The models generated using the above methods are compared on various parameters to identify the best model. The **parameters used for the comparison are Adj R^2, RMSE, MAE, MSE, AIC and BIC**. While AIC and BIC are computed using AIC() and BIC() functions on R respectively, the remaining parameters are obtained from the model summary attached in the previous sections.

The table below summarizes the results for the three models:

	Backward elimination model (model 1)	Forward selection model (model 2)	Stepwise selection model (model 3)
Adj R^2	0.908	0.908	0.908
RMSE	3.302	3.302	3.302
MAE	2.449	2.449	2.449
MSE	10.906	10.906	10.906
AIC	30498.55	30498.55	30498.55
BIC	30605.37	30605.37	30605.37

Since the three methods have resulted in the same model, it is safe to assume that the model generated is the **best possible fit for the criteria defined**. It can also be observed from the above table that all the parameters have the same values for each model as they are representing the same independent and dependent variables.

Note:

In general, we always prefer models with the highest R^2 value and low values for RMSE, MAE, MSE, AIC and BIC.

5. **Issues with the variable selection:**

Although these selection methods are quick and easy, they have issues of their own. Some of the issues are as follows:

- Miss suppressor relations: Sometimes a predictor, which although has no correlation with a target variable, can help in predicting the target variable by complementing with parts of another predictor which doesn't help in explaining the target variable. These variables are called suppressor variables that enhance a model, and the forward selection or backward elimination can miss on these relations giving us not the most efficient model.
- Miss complementary variables: Complimentary variables are variables that are negatively correlated with each other. Often there can be such a pair in the list of predictors which together explain the target variable better thus enhancing the model. These relations can be missed while using a forward selection or backward elimination.
- High probability of type 1 error: Since the stepwise selection method performs a large number of T-tests at every step, there is a very high probability of type 1 error happening.
- Lack of flexibility: In forward and backward methods, once a variable is selected or eliminated it cannot be undone thus lacking flexibility. This leads to a serious efficiency issue of the model as these variables can either become significant at later steps of building the model or sometimes the variable can become insignificant after adding another variable.
- Instability for low sample size: There can be instability in choosing variables using the stepwise method if the sample size is small. Hence to overcome this, this method should always be used with a dataset having at least 50 events per variable.

- Too many variables and collinearity issues: These methods fail to provide the best model if there are too many candidate variables involved and also are bad at handling collinearity.

Hence these methods should only be used as a guide towards building the final model.

6. Hypothetical question – What if we have too many predictors?

While predictors help predict the target variable, having too many of them can lead to various issues. Three such issues are listed below:

- Overfitting:
  Overfitting is a problem caused in models when the model is very complicated and fits all the data very well. By complicated we mean including too many variables. Including too many variables **can retain some of the noisy variables** and when the model trains against this data and fits too close, the model becomes overfitted and cannot validate for data from new sources.
- Collinearity:
  Including too many variables in a model can lead to collinearity issues. At times a pair of predictors can be spotted **having a high correlation** with each other. Having these pairs in the model can lead to inconsistent results or fluctuations. Hence, it's not a good practice to construct a complex model.
- Reduced statistical power:
  Having too many variables in a model can **reduce the statistical power of the predictors and increase the chances of errors** in a model. So, unless the sample size is large, having too many variables in a regression model will affect the inference giving us a false sense of understanding.

7. **Multicollinearity:**

a. Introduction:

Multicollinearity is the state where there is a high correlation between two or more predictors that are used for a model. In simpler words, if one predictor can predict the other predictor there is a correlation between the two. Using such variables in a linear regression model can lead to redundancy issues thus giving skewed results.

On a high level, there are two types of multicollinearities:

- Data-based multicollinearity: This type of error is caused by poor design experiments or data collection methods. So before using data, it is important to identify if there are any high correlations.
- Structural multicollinearity: This is caused when the analyst while constructing a model creates new variables which are highly correlated with already existing predictors.

Since in regression models, our aim is to determine the effect of each predictor on the target variable, having multicollinearity can lead to a wrong interpretation as we do not know the clear effect of each individual variable. Hence it is very important to eliminate multicollinearity in regression models.

b. Methods to identify multicollinearity:

Three methods to identify multicollinearity are as follows:

- Correlation matrix:
  By generating a correlation matrix, we obtain the correlation coefficient values for all potential predictors being used. Using these values, one can spot pairs of variables that are having high correlation and thus can remove one of them from the model.
- Variance Inflation Factor:
  The variance inflation factor is yet another method where a VIF test is run on the model in which multicollinearity is being assessed. After obtaining the results, we look for variables that have a VIF value greater than 5. The

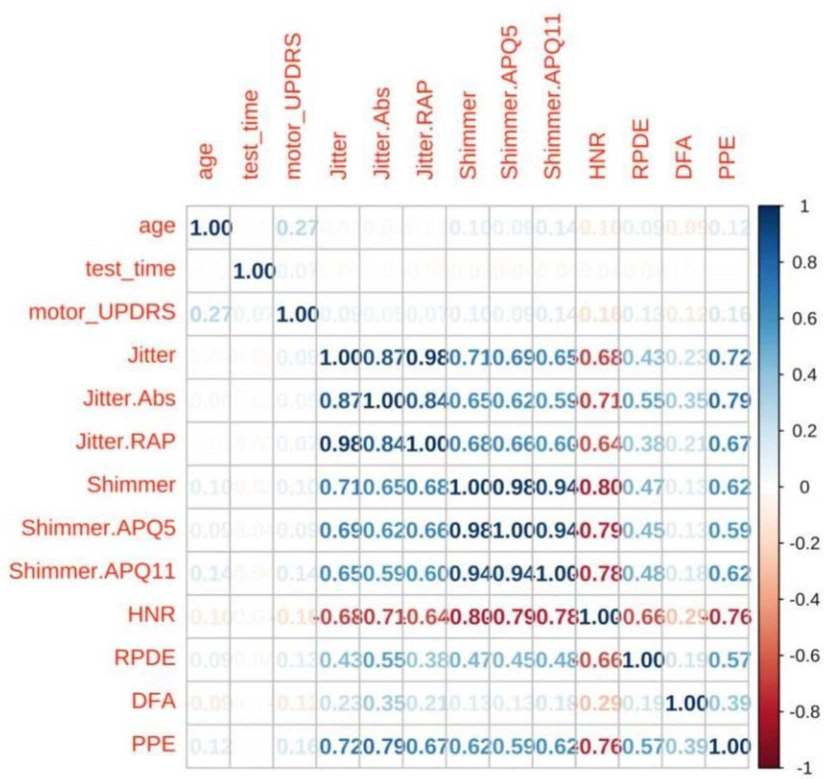
variables that are identified are the ones with high correlation with another predictor in the model.

- <u>High Standard errors:</u>
  This is a very quick way to identify collinearity. After generating the model, the variables with high standard error for their coefficients are noted. These variables are the reason for multicollinearity in the model. It is best practice to confirm the results generated by this method with another method.

c. <u>Implementing the above methods:</u>

The above methods are implemented one by one to identify the variables which are posing the issue of multicollinearity.

- Correlation matrix:
  The correlation matrix has been created for the variables in the model to observe collinearity in the model. This has been achieved using the 'cor()' function on R. The generated result is as follows:



From the above matrix, we can see a **high correlation between Jitter and Jitter.RAP**. Also, a **high correlation can be observed between Jitter and Jitter.Abs**.

Other interesting pairs are Shimmer variables. **High correlation is noted among different Shimmer variables**. HNR can also be spotted with a high correlation with Shimmer and other variables of Shimmer.

- Variance Inflation Factor:

  Variance inflation factor has been calculated for the final model generated using the selection methods. The result generated after running the vif() function for the final model is as follows:

```
> vif(finalmodel)
          age           sex     test_time   motor_UPDRS        Jitter    Jitter.Abs    Jitter.RAP       Shimmer
     1.130542      1.260827      1.014316      1.182252     48.555480      6.828097     41.083911     40.448104
 Shimmer.APQ5 Shimmer.APQ11           HNR          RPDE           DFA           PPE
    40.567483     10.320015      5.285742      2.055809      1.398114      4.214126
```

From the above output, we can notice that **Jitter, Jitter.Abs, Jitter.RAP, Shimmer, Shimmer.APQ5, Shimmer.APQ11**. Since HNR has a score of 5.28 which is very close to 5, the collinearity threat posed by this variable is not significant.

- High Standard Errors for coefficients:

  After generating the model, the standard errors are checked for high values. The result for the model summary is as follows:

```
Coefficients:
                   Estimate    Std. Error t value              Pr(>|t|)
(Intercept)       1.9112051     1.0456829   1.828              0.06764 .
age               0.0683262     0.0051258  13.330 < 0.0000000000000002 ***
sex              -1.4031464     0.1024760 -13.692 < 0.0000000000000002 ***
test_time         0.0025476     0.0008012   3.180              0.00148 **
motor_UPDRS       1.2264538     0.0056840 215.774 < 0.0000000000000002 ***
Jitter         -271.9649284    52.6366422  -5.167         0.0000002460 ***
Jitter.Abs    14032.3245541  3087.2403356   4.545         0.0000055971 ***
Jitter.RAP      385.3960796    87.1567020   4.422         0.0000099619 ***
Shimmer         -47.2241852    10.4652654  -4.512         0.0000065332 ***
Shimmer.APQ5     90.3819937    16.2513691   5.562         0.0000000279 ***
Shimmer.APQ11   -33.6136516     6.8334796  -4.919         0.0000008938 ***
HNR              -0.0982873     0.0227629  -4.318         0.0000160146 ***
RPDE              3.2186516     0.6031372   5.337         0.0000000983 ***
DFA              -2.1232970     0.7092906  -2.994              0.00277 **
PPE              -3.8118009     0.9540747  -3.995         0.0000654098 ***
---
```

From the above results, it can be quickly noted that **the three Jitter and the three Shimmer variables have high standard errors**. Hence these variables cause multicollinearity in the model.

Conclusion:

Upon analyzing the selected variables using the three methods listed above, a common inference can be made. The three Jitter and three Shimmer variables are having high collinearity among themselves. The results are summarized below:

S.no.	Variables	Parameters		
		Correlation Coefficient	VIF	Standard Error
1.	Jitter	High with 2 and 3	48.55	52.636
2.	Jitter.Abs	High with 1 and 2	6.828	3087.24
3.	Jitter.RAP	High with 1 and 3	41.08	87.156
4.	Shimmer	High with 5 and 6	40.44	10.46
5.	Shimmer.APQ5	High with 4 and 6	40.56	16.25
6.	Shimmer.APQ11	High with 4 and 5	10.32	6.3

Note: s.no of the variables are mentioned under the correlation coefficient column.

Hence the final model can **drop Jitter.Abs, Jitter.RAP, Shimmer.APQ5 and Shimmer.APQ11** from the model and **retain Jitter and Shimmer** to remove multicollinearity.