A Multiple Linear Regression (MLR) Model for Automobile data set

1 Abstract

In this project, we fit a multiple linear regression model to the Automobile dataset provided by the UCI Machine Learning Repository. We consider 15 independent variables (predictors) for this model and, after evaluating the model, introduce the optimal model using Subset Selection and Stepwise Selection methods. R software is utilized for calculations in this project.

2 Dataset Description

The original data set contains 205 instances described by 25 attributes including 15 continuous and 10 categorical. After preprocess the data set in the following way: we neglect the 10 categorical attributes, and remove the instances with missing values, yielding a data set with 160 instances and 15 attributes. We select one of the 15 attributes as the response (price) and the others as the predictors: specifically we want to predict the price of an automobile based the other 14 attributes of it. The following displays details of the dataset.

```
> str(data_contin)
'data.frame':   160 obs. of  15 variables:
 $ price(Y)              : num  13950 17450 17710 23875 16430 ...
 $ highway-mpg(X1)       : num  30 22 25 20 29 29 28 28 53 43 ...
 $ city-mpg(X2)          : num  24 18 19 17 23 23 21 21 47 38 ...
 $ peak-rpm(X3)          : num  5500 5500 5500 5500 5800 5800 4250 4250 5100 5400 ...
 $ horsepower(X4)        : num  102 115 110 140 101 101 121 121 48 70 ...
 $ compression-ratio(X5) : num  10 8 8.5 8.3 8.8 8.8 9 9 9.5 9.6 ...
 $ stroke(X6)            : num  3.4 3.4 3.4 3.4 2.8 2.8 3.19 3.19 3.03 3.11 ...
 $ bore(X7)              : num  3.19 3.19 3.19 3.13 3.5 3.5 3.31 3.31 2.91 3.03 ...
 $ engine-size(X8)       : num  109 136 136 131 108 108 164 164 61 90 ...
 $ curb-weight(X9)       : num  2337 2824 2844 3086 2395 ...
 $ height(X10)           : num  54.3 54.3 55.7 55.9 54.3 54.3 54.3 54.3 53.2 52 ...
 $ width(X11)            : num  66.2 66.4 71.4 71.4 64.8 64.8 64.8 64.8 60.3 63.6 ...
 $ length(X12)           : num  177 177 193 193 177 ...
 $ wheel-base(X13)       : num  99.8 99.4 105.8 105.8 101.2 ...
 $ normalized-losses(X14): num  164 164 158 158 192 192 188 188 121 98 ...

> head(data_contin,n=5)
      Y X1 X2   X3  X4   X5  X6   X7  X8   X9  X10  X11   X12   X13 X14
1 13950 30 24 5500 102 10.0 3.4 3.19 109 2337 54.3 66.2 176.6  99.8 164
2 17450 22 18 5500 115  8.0 3.4 3.19 136 2824 54.3 66.4 176.6  99.4 164
3 17710 25 19 5500 110  8.5 3.4 3.19 136 2844 55.7 71.4 192.7 105.8 158
4 23875 20 17 5500 140  8.3 3.4 3.13 131 3086 55.9 71.4 192.7 105.8 158
5 16430 29 23 5800 101  8.8 2.8 3.50 108 2395 54.3 64.8 176.8 101.2 192
```

```
> tail(data_contin,n=5)
        Y X1 X2   X3  X4   X5   X6   X7  X8   X9  X10  X11   X12   X13 X14
156 16845 28 23 5400 114  9.5 3.15 3.78 141 2952 55.5 68.9 188.8 109.1  95
157 19045 25 19 5300 160  8.7 3.15 3.78 141 3049 55.5 68.8 188.8 109.1  95
158 21485 23 18 5500 134  8.8 2.87 3.58 173 3012 55.5 68.9 188.8 109.1  95
159 22470 27 26 4800 106 23.0 3.40 3.01 145 3217 55.5 68.9 188.8 109.1  95
160 22625 25 19 5400 114  9.5 3.15 3.78 141 3062 55.5 68.9 188.8 109.1  95
```

3 Initial Analysing and Transformations

In this section, for the initial analysis and examining the relationship between each independent variable (predictor) and the dependent variable, we first calculate the Pearson correlation coefficient for each pair of variables. Considering the correlation matrix, it is observed that high correlation exists among independent variables (X_1, X_2, X_4, X_7, X_8, X_9, X_{11}, X_{12}, X_{13}) and the dependent variable (Y). Additionally, there is high correlation among some independent variables.

```
> cor_mat(Correlation matrix)
15 x 15 Matrix of class "dtrMatrix"
        Y     X1     X2     X3     X4     X5     X6     X7     X8     X9    X10    X11    X12    X13    X14
Y   1.000 -0.718 -0.690 -0.174  0.759  0.211  0.159  0.535  0.842  0.894  0.248  0.843  0.760  0.735  0.200
X1      .  1.000  0.972 -0.034 -0.828  0.222 -0.014 -0.587 -0.711 -0.787 -0.222 -0.689 -0.718 -0.608 -0.190
X2      .      .  1.000 -0.055 -0.837  0.280 -0.021 -0.586 -0.696 -0.759 -0.195 -0.662 -0.717 -0.577 -0.237
X3      .      .      .  1.000  0.075 -0.419 -0.009 -0.316 -0.287 -0.262 -0.251 -0.236 -0.239 -0.292  0.241
X4      .      .      .      .  1.000 -0.163  0.149  0.557  0.810  0.788  0.032  0.679  0.667  0.514  0.291
X5      .      .      .      .      .  1.000  0.241  0.019  0.144  0.226  0.237  0.262  0.189  0.294 -0.130
X6      .      .      .      .      .      .  1.000 -0.106  0.297  0.172 -0.095  0.193  0.116  0.164  0.066
X7      .      .      .      .      .      .      .  1.000  0.597  0.647  0.262  0.575  0.649  0.581 -0.036
X8      .      .      .      .      .      .      .      .  1.000  0.889  0.116  0.780  0.727  0.650  0.204
X9      .      .      .      .      .      .      .      .      .  1.000  0.369  0.871  0.870  0.810  0.123
X10     .      .      .      .      .      .      .      .      .      .  1.000  0.298  0.505  0.559 -0.417
X11     .      .      .      .      .      .      .      .      .      .      .  1.000  0.839  0.816  0.105
X12     .      .      .      .      .      .      .      .      .      .      .      .  1.000  0.872  0.029
X13     .      .      .      .      .      .      .      .      .      .      .      .      .  1.000 -0.064
X14     .      .      .      .      .      .      .      .      .      .      .      .      .      .  1.000
```

Next, we fit a multiple linear regression model to the data, and the results are as follows.

```
summary(fit1)
Call:
lm(formula = Y ~ ., data = data_contin)
Residuals:
    Min      1Q  Median      3Q     Max
-5861.1 -1236.7  -213.4   898.0  7777.0
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.925e+04  1.564e+04  -3.788 0.000222 ***
X1          -6.750e+00  1.395e+02  -0.048 0.961487
X2           1.057e+01  1.549e+02   0.068 0.945678
X3           7.431e-01  5.651e-01   1.315 0.190559
X4           2.654e+01  1.645e+01   1.613 0.108937
X5           1.077e+02  7.694e+01   1.400 0.163674
X6          -1.847e+03  7.776e+02  -2.375 0.018834 *
X7          -1.828e+03  1.078e+03  -1.696 0.092029 .
X8           5.024e+01  1.852e+01   2.712 0.007489 **
X9           5.042e+00  1.596e+00   3.159 0.001926 **
X10          4.312e+01  1.360e+02   0.317 0.751621
X11          7.856e+02  2.314e+02   3.395 0.000884 ***
X12         -9.207e+01  4.668e+01  -1.973 0.050450 .
X13          1.813e+02  9.066e+01   2.000 0.047358 *
X14          8.261e+00  6.607e+00   1.250 0.213184
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 2371 on 145 degrees of freedom
Multiple R-squared:  0.8509,    Adjusted R-squared:  0.8365
F-statistic:  59.1 on 14 and 145 DF,  p-value: < 2.2e-16

> anova(fit1)
Analysis of Variance Table

Response: Y
           Df     Sum Sq    Mean Sq  F value    Pr(>F)
X1          1 2820869164 2820869164 501.7190 < 2.2e-16 ***
X2          1    6494644    6494644   1.1551   0.28426
X3          1  210978066  210978066  37.5245 8.066e-09 ***
X4          1  570651476  570651476 101.4959 < 2.2e-16 ***
X5          1  453038039  453038039  80.5772 1.331e-15 ***
X6          1     206824     206824   0.0368   0.84817
X7          1      62778      62778   0.0112   0.91599
X8          1  264684250  264684250  47.0767 1.838e-10 ***
X9          1  193764793  193764793  34.4630 2.839e-08 ***
X10         1    1038751    1038751   0.1848   0.66796
X11         1   90102292   90102292  16.0256 9.938e-05 ***
X12         1    8843029    8843029   1.5728   0.21182
X13         1   22286097   22286097   3.9638   0.04837 *
X14         1    8790010    8790010   1.5634   0.21318
Residuals 145  815249217    5622408
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

With a coefficient of determination (Multiple R-squared =0.8509), the fitted model seems to be a good fit. To further examine the fitted model, we plot the residuals against the fitted values and a QQ plot (Figure 1). Observing the residual plots and the lack of a specific pattern, the assumption of constant variance for this model is considered valid. Furthermore, examining the QQ plot and box plot indicates that the error distribution is approximately normal and has a mean of zero.

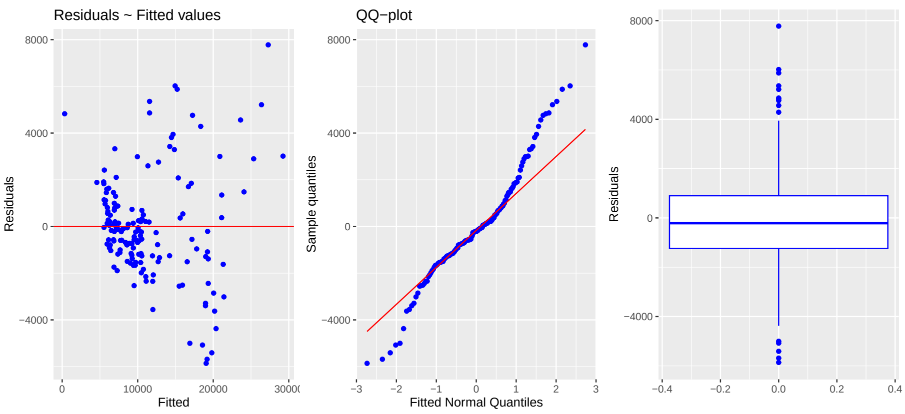

Figure 1: Residuals Analysis

Subsequently, to investigate the potential linear relationship between the dependent variable and each independent variable, scatter plots are drawn. By inspecting these plots in figure 2, linear relationships between some independent variables and the dependent variable are observed. Also, for some independent variables (X_1, X_2, X_{12}), their relationship with the dependent variable is transformed into a linear form using an appropriate transformation. Therefore, for variables (X_1, X_2), we consider the transformed relations: $X^* = 1/X$and for variable X_{12}, we consider the transformation $X^* = e^x$. Consequently, we proceeded to fit a multiple linear regression model to the transformed data, resulting in an improvement in the coefficient of determination (Multiple R-squared) compared to the initial model.

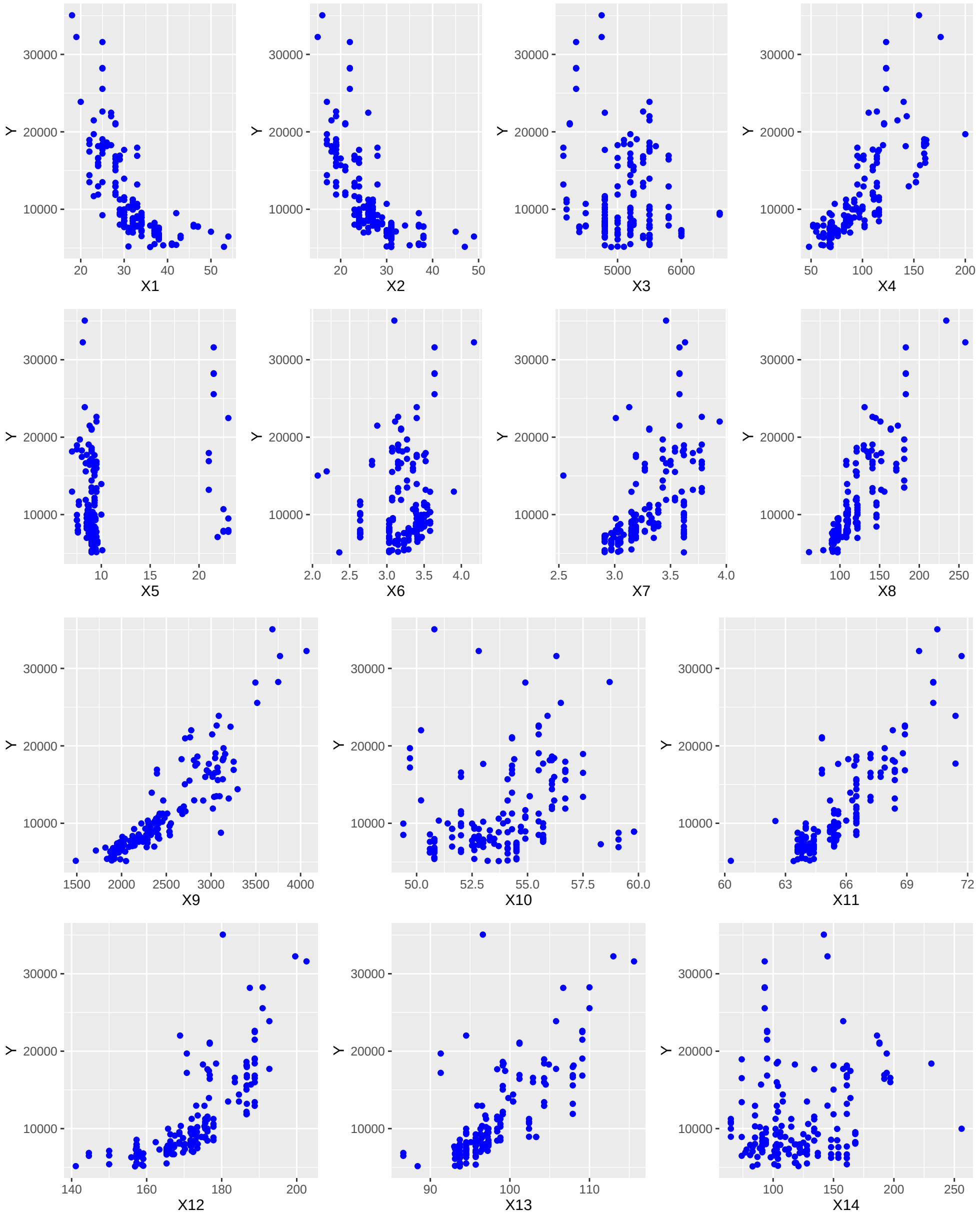

Figure 2: Scatter plots.

```
> fit2<-lm(Y ~.,data=data_trans)
> summary(fit2)
Call:
lm(formula = Y ~ ., data = data_trans)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-6264.7 -1275.7  -156.5   956.0  7346.6
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.956e+04  1.412e+04  -3.510 0.000596 ***
X1_star      1.398e+05  1.155e+05   1.211 0.227983
X2_star      2.721e+04  9.873e+04   0.276 0.783262
X3           6.280e-01  5.541e-01   1.133 0.258962
X4           1.115e+01  1.782e+01   0.626 0.532523
X5           1.836e+02  7.842e+01   2.342 0.020558 *
X6          -1.552e+03  7.730e+02  -2.007 0.046594 *
X7          -1.732e+03  1.050e+03  -1.650 0.101065
X8           5.228e+01  1.810e+01   2.889 0.004458 **
X9           2.935e+00  1.599e+00   1.835 0.068496 .
X10         -9.458e+00  1.313e+02  -0.072 0.942673
X11          5.699e+02  2.240e+02   2.544 0.012010 *
X12_star     6.588e-85  2.578e-85   2.555 0.011641 *
X13          8.800e+01  8.228e+01   1.069 0.286645
X14          7.961e+00  6.607e+00   1.205 0.230224
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 2324 on 145 degrees of freedom
Multiple R-squared:  0.8567,    Adjusted R-squared:  0.8429
F-statistic: 61.94 on 14 and 145 DF,  p-value: < 2.2e-16

> anova(fit2)
Analysis of Variance Table

Response: Y
           Df     Sum Sq    Mean Sq  F value    Pr(>F)
X1_star     1 3460957377 3460957377 640.7230 < 2.2e-16 ***
X2_star     1     110936     110936   0.0205 0.8862449
X3          1  193072024  193072024  35.7432 1.673e-08 ***
X4          1  157403410  157403410  29.1399 2.686e-07 ***
X5          1  496511573  496511573  91.9186 < 2.2e-16 ***
X6          1    1577818    1577818   0.2921 0.5897074
X7          1    1136014    1136014   0.2103 0.6472120
X8          1  151319920  151319920  28.0137 4.366e-07 ***
X9          1   96651113   96651113  17.8929 4.122e-05 ***
X10         1    1801190    1801190   0.3335 0.5645289
X11         1   71433351   71433351  13.2244 0.0003832 ***
X12_star    1   37685906   37685906   6.9767 0.0091632 **
X13         1    6319074    6319074   1.1698 0.2812291
X14         1    7841383    7841383   1.4517 0.2302236
Residuals 145  783238338    5401644
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

4 Outlier Observations

In regression analysis, the dataset may contain unusual cases referred to as outliers. These outliers can include large residuals and often have noticeable effects on the least squares regression function. Therefore, studying and deciding whether to retain or remove these data points are crucial. We use the studentized deleted residuals criterion to identify outlier observations, and, based on this criterion, observation 50 is identified as an outlier for Y, necessitating its removal, as shown in Figure 3.

```
Based on Bonferroni test of studentized deleted residuals,outliers are observations:
```

```
 id Y_outliers |t_Stud_Delet_Res|
 50      35056            4.270503
```

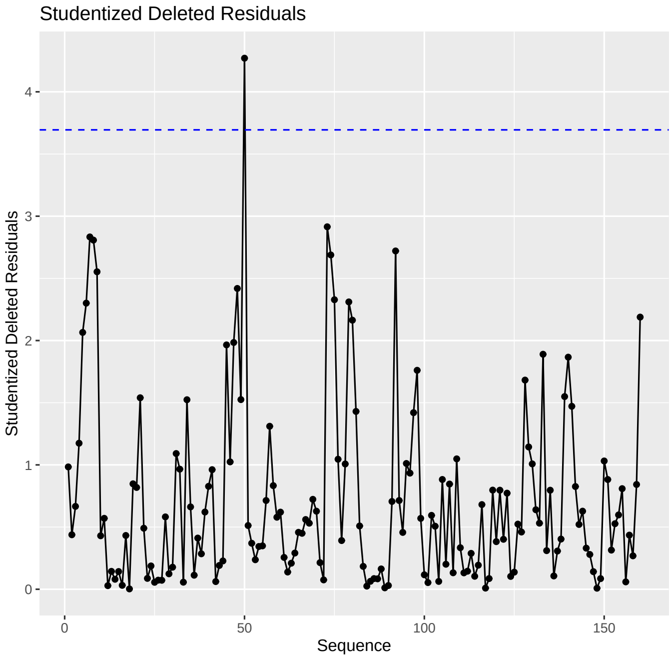

Figure 3: Studentized deleted residuals

The model's performance is re-evaluated using the new dataset after removing this observation, and the results are presented below. Examining these results indicates an improvement in the model's performance, considering the determination coefficient criterion compared to the previous model.

```
> fit3<-lm(Y~.,data=data_trans2)
> summary(fit3)
Call:
lm(formula = Y ~ ., data = data_trans2)
Residuals:
    Min      1Q  Median      3Q     Max
-5298.6 -1158.7  -160.2   866.4  6358.5
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.866e+04  1.335e+04  -3.645 0.000372 ***
X1_star      7.901e+03  1.134e+05   0.070 0.944576
X2_star      8.910e+04  9.445e+04   0.943 0.347110
X3           4.553e-01  5.254e-01   0.867 0.387640
X4           3.142e+01  1.750e+01   1.795 0.074734 .
X5           2.085e+02  7.437e+01   2.804 0.005745 **
X6          -1.081e+03  7.390e+02  -1.463 0.145540
X7          -1.291e+03  9.978e+02  -1.294 0.197859
X8           2.772e+01  1.805e+01   1.536 0.126751
X9           2.755e+00  1.512e+00   1.822 0.070529 .
X10          1.816e-02  1.242e+02   0.000 0.999884
X11          3.804e+02  2.164e+02   1.758 0.080899 .
X12_star     7.076e-85  2.440e-85   2.900 0.004318 **
X13          2.057e+02  8.253e+01   2.492 0.013829 *
X14          1.026e+01  6.269e+00   1.636 0.104048
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 2197 on 144 degrees of freedom
Multiple R-squared:  0.8583,    Adjusted R-squared:  0.8445
F-statistic: 62.29 on 14 and 144 DF,  p-value: < 2.2e-16

> anova(fit3)
Analysis of Variance Table

Response: Y
           Df     Sum Sq    Mean Sq  F value    Pr(>F)
X1_star     1 2956012232 2956012232 612.2978 < 2.2e-16 ***
X2_star     1    5522911    5522911   1.1440 0.2865991
X3          1  185579414  185579414  38.4403 5.637e-09 ***
X4          1  170914949  170914949  35.4027 1.947e-08 ***
X5          1  530630051  530630051 109.9128 < 2.2e-16 ***
X6          1    1786489    1786489   0.3700 0.5439369
X7          1    5744450    5744450   1.1899 0.2771747
X8          1   85112487   85112487  17.6299 4.678e-05 ***
X9          1  115513649  115513649  23.9271 2.643e-06 ***
X10         1     103948     103948   0.0215 0.8835456
X11         1   65332196   65332196  13.5327 0.0003304 ***
X12_star    1   45639120   45639120   9.4535 0.0025220 **
X13         1   29245255   29245255   6.0578 0.0150269 *
X14         1   12919543   12919543   2.6761 0.1040483
Residuals 144  695193982    4827736
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

5 Multicollinearity

Multicollinearity is a statistical issue arising due to high correlations among independent (predictor) variables in regression models. It complicates the interpretation of the model and introduces an overfitting problem. It is common for individuals to test for multicollinearity before selecting variables for inclusion in a regression model. To assess multicollinearity, we use the Variance Inflation Factor (VIF). For each independent variable, we calculate the VIF, and if the VIF value exceeds 10, multicollinearity is present, and we need to address this issue. Therefore, based on the VIF values, it is concluded that there is multicollinearity for variables (X_1, X_2, X_9). One approach to address multicollinearity is to use Regression Stepwise methods, which will be discussed in the next section.

```
> res_VIF
               VIF
X1_star  16.868394
X2_star  22.962424
X3        1.960606
X4        9.239406
X5        2.741599
X6        1.554847
X7        2.338202
X8        9.026536
X9       16.699418
X10       2.600649
X11       5.610467
X12_star  1.162254
X13       5.977023
X14       1.637262
```

6 Model and Variable Selection

One of the most critical statistical issues is the problem of selecting the best model among statistical linear models or choosing important variables in a linear model with a large number of predictor variables. Model selection is the process of choosing a model from a set of candidate models. Variable selection means choosing from among many variables to be included in a particular model, i.e., selecting relevant variables from a complete list of variables by excluding irrelevant or redundant variables. The goal of such selection is to determine a set of variables that provides the best fit for the model to make accurate predictions. In this section, to select the best model or important variables, we use Backward Stepwise Selection and Forward Stepwise Selection methods, Subset Selection methods, and choose a model that has smallest values of Mallows's C_p and Information Akaike Criterion (AIC), or equivalently largest Adjusted R-square . By comparing optimal models based on these three methods, the Subset Selection and Forward Stepwise Selection methods introduce the same model, which has smallest values of Mallows's C_p and AIC compared to the model introduced by Backward Stepwise Selection.

```
> ols_step_best_subset(fit3,details = F, pent=0.1,prem=0.3)
                              Best Subsets Regression
-----------------------------------------------------------------------------------
Model Index    Predictors
-----------------------------------------------------------------------------------
     1         X9
     2         X9 X11
     3         X4 X9 X11
     4         X4 X9 X11 X12_star
     5         X4 X9 X11 X12_star X14
     6         X4 X5 X9 X12_star X13 X14
     7         X4 X5 X9 X11 X12_star X13 X14
     8         X2_star X4 X5 X9 X11 X12_star X13 X14
     9         X2_star X4 X5 X7 X9 X11 X12_star X13 X14
    10         X2_star X4 X5 X6 X7 X9 X11 X12_star X13 X14
    11         X2_star X4 X5 X6 X7 X8 X9 X11 X12_star X13 X14
    12         X2_star X3 X4 X5 X6 X7 X8 X9 X11 X12_star X13 X14
    13         X1_star X2_star X3 X4 X5 X6 X7 X8 X9 X11 X12_star X13 X14
    14         X1_star X2_star X3 X4 X5 X6 X7 X8 X9 X10 X11 X12_star X13 X14
-----------------------------------------------------------------------------------
                                Subsets Regression Summary
------------------------------------------------------------------------------------
                        Adj.        Pred
Model    R-Square    R-Square    R-Square      C(p)         AIC
------------------------------------------------------------------------------------
  1        0.7984      0.7972       0.791     49.8023    2944.4640
  2        0.8160      0.8136      0.8027     33.9682    2931.9781
  3        0.8255      0.8221      0.8095     26.3520    2925.5826
  4        0.8354      0.8311     -2.0171     18.2792    2918.2855
  5        0.8402      0.8350     -1.8937     15.3231    2915.5035
  6        0.8474      0.8414     -1.5266     10.0698    2910.2350
  7        0.8510      0.8441     -2.1968      8.3638    2908.3890
  8        0.8539      0.8461     -2.2241      7.4850    2907.3358
  9        0.8549      0.8461     -2.0162      8.4208    2908.1922
 10        0.8557      0.8460     -3.0289      9.5844    2909.2876
 11        0.8575      0.8468     -2.6495      9.7769    2909.3148
 12        0.8583      0.8466     -2.4813     11.0049    2910.4646
 13        0.8583      0.8456     -2.5014     13.0000    2912.4592
 14        0.8583      0.8445     -2.7252     15.0000    2914.4592
------------------------------------------------------------------------------------
                                Subsets Regression Summary
```

```
-----------------------------------------------------------------------------------------------
Model     SBIC          SBC            MSEP               FPE            HSP          APC
-----------------------------------------------------------------------------------------------
  1     2492.2470    2953.6708    1001327540.1546    6376866.8350    40369.5583    0.2067
  2     2479.8134    2944.2537     920030161.2589    5895281.4565    37329.7392    0.1911
  3     2473.5109    2940.9271     878377766.0066    5662892.6141    35869.6489    0.1836
  4     2466.5632    2936.6989     833904740.9692    5408930.0605    34274.6644    0.1753
  5     2464.0345    2936.9858     814521798.8830    5315194.7036    33696.8085    0.1723
  6     2459.3661    2934.7863     783278494.1841    5142070.3610    32617.4574    0.1667
  7     2457.9687    2936.0092     769656223.0624    5082857.4861    32262.4487    0.1648
  8     2457.3780    2938.0249     760085160.5569    5049482.9456    32073.6573    0.1637
  9     2458.5283    2941.9501     759736462.3868    5076980.3994    32274.1038    0.1646
 10     2459.9142    2946.1144     760565434.9612    5112360.7279    32527.6253    0.1657
 11     2460.4308    2949.2105     756332128.3910    5113574.0060    32566.6213    0.1658
 12     2461.9278    2953.4293     757487128.6433    5151090.9248    32839.7537    0.1670
 13     2464.1317    2958.4928     762721559.3704    5216593.3140    33294.7309    0.1691
 14     2466.3401    2963.5617     768055276.4547    5283182.7751    33760.3915    0.1713
-----------------------------------------------------------------------------------------------
 AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria

> ols_step_forward_p(fit3, details = F, pent=0.1,prem=0.3)
                              Selection Summary
-----------------------------------------------------------------------------
        Variable                  Adj.
Step    Entered     R-Square    R-Square     C(p)         AIC         RMSE
-----------------------------------------------------------------------------
   1    X9            0.7984      0.7972    49.8023    2944.4640    2509.5121
   2    X11           0.8160      0.8136    33.9682    2931.9781    2405.4333
   3    X4            0.8255      0.8221    26.3520    2925.5826    2350.3034
   4    X12_star      0.8354      0.8311    18.2792    2918.2855    2289.9833
   5    X14           0.8402      0.8350    15.3231    2915.5035    2263.1648
   6    X13           0.8457      0.8396    11.7888    2911.9880    2231.5546
   7    X5            0.8510      0.8441     8.3638    2908.3890    2199.8562
   8    X2_star       0.8539      0.8461     7.4850    2907.3358    2186.0867
-----------------------------------------------------------------------------

> ols_step_backward_p(fit3, details = F, pent=0.1,prem=0.3)
                             Elimination Summary
-----------------------------------------------------------------------------
        Variable                  Adj.
Step    Removed     R-Square    R-Square     C(p)         AIC         RMSE
-----------------------------------------------------------------------------
   1    X10           0.8583      0.8456    13.0000    2912.4592    2189.6213
   2    X1_star       0.8583      0.8466    11.0049    2910.4646    2182.1467
   3    X3            0.8575      0.8468     9.7769    2909.3148    2180.5336
-----------------------------------------------------------------------------
```

Table 1: Model selection

method	Adj. R-Square	C(p)	AIC
Subset Selection	0.8461	7.4850	2907.3358
Forward stepwise selection	0.8461	7.4850	2907.3358
Backward stepwise selection	0.8468	9.7769	2909.3148

7 Conclusion

This project introduces an optimal multiple linear regression model for the Automobile dataset provided by the UCI Machine Learning Repository using Subset Selection and Stepwise Selection methods. Therefore, this model can be useful for predicting car prices using the relevant independent variables. Finally, it is worth mentioning that the optimality of the above model can be validated using alternative methods such as Cross-Validation and further improved by transformations and selecting more important variables.

References

[1] Xin Cai, Guang Lin & Jinglai Li (2021) Bayesian inverse regression for supervised dimension reduction with small datasets, *Journal of Statistical Computation and Simulation*, **91:14**, 2817-2832, DOI: 10.1080/00949655.2021.1909025

[2] Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005) Applied Linear Statistical Models. 5th Edition, McGraw-Hill, Irwin, New York.