Final Project

Arvin Castelo, Semal Shastri

2024-12-10

Contents

---

1 Introduction

To explain which predictors are useful in determining exam scores.

---

2 Data Pre-processing

2.1 Data Cleaning

We observed entries with missing values in the dataset. These entries will be removed.

```
data <- data |>
  filter(Teacher_Quality != "",
         Distance_from_Home != "",
         Parental_Education_Level != "")
```

There are no duplicate entries in the document. The resulting dataset will be sampled for this study.

term	estimate	std.error	statistic	p.value
Hours_Studied	0.2800	0.0254	11.0364	0.0000
Attendance	0.1890	0.0121	15.6100	0.0000
Sleep_Hours	-0.1220	0.1144	-1.0666	0.2867
Previous_Scores	0.0430	0.0121	3.5597	0.0004
Tutoring_Sessions	0.4705	0.1376	3.4199	0.0007
Physical_Activity	0.4004	0.1590	2.5176	0.0121

2.2 Data Splitting

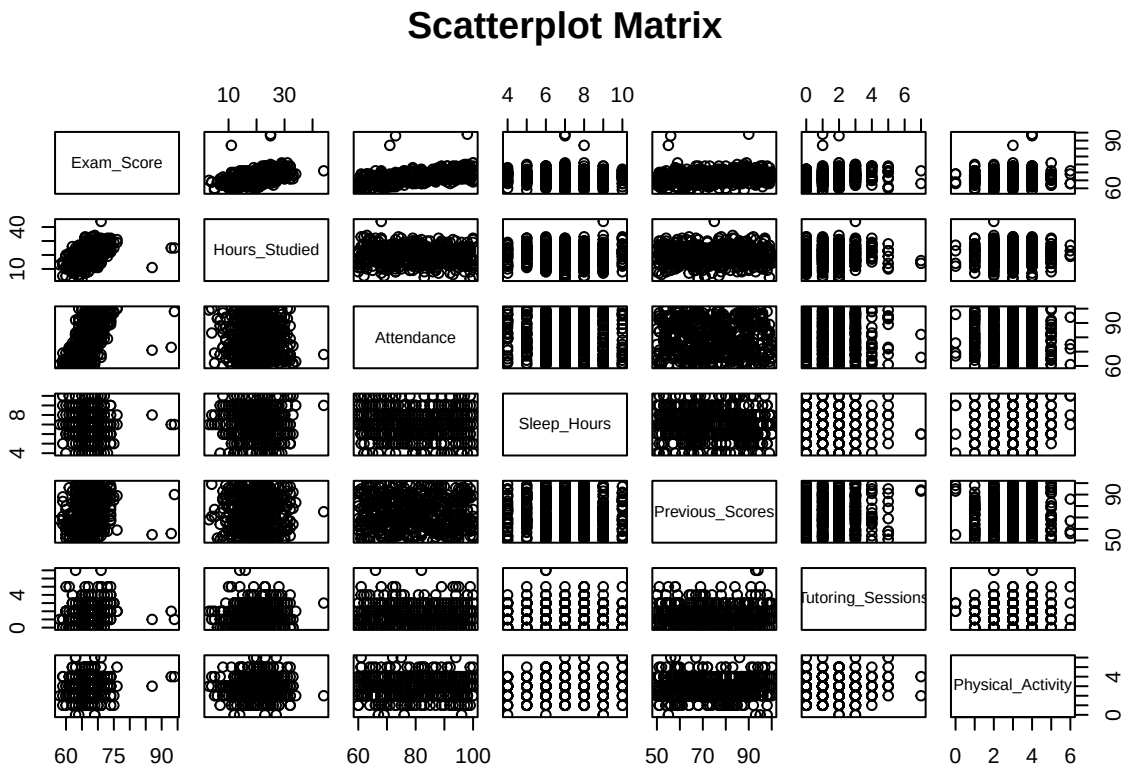
We used a separate train and test sets coming from this data. We used seed number 456 for this analysis.

---

3 Summary Statistics and Data Visualization

3.1 Numerical Variables

Scatterplots provide a good visualization of the relationships between the numerical predictors and the response variable.



Based on the scatterplot matrix and the regression models, the following observations between the dependent variable and each predictor can be drawn:

- Hours_Studied, Attendance, Previous_Scores, Tutoring_Sessions and Physical_Activity have significant

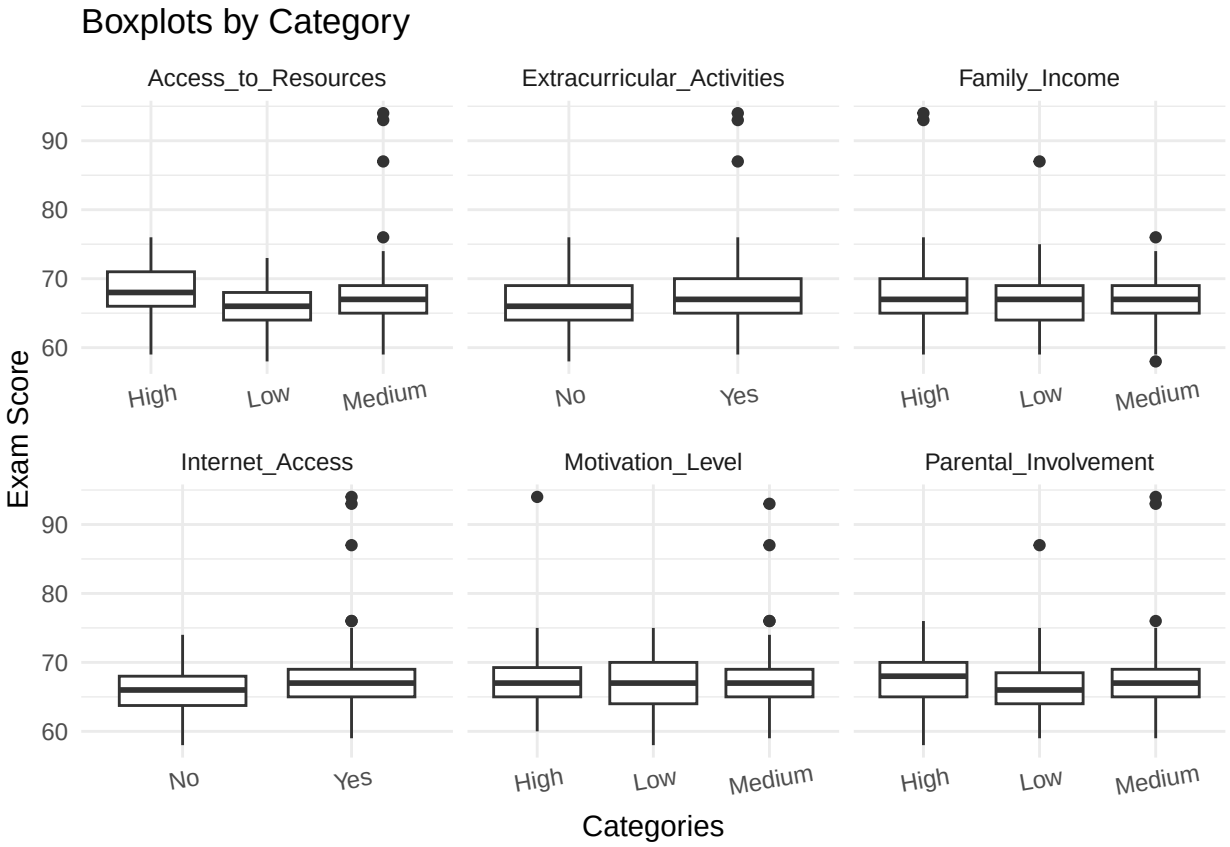
linear relationship with Exam_Scores , provided no other predictors are in the model

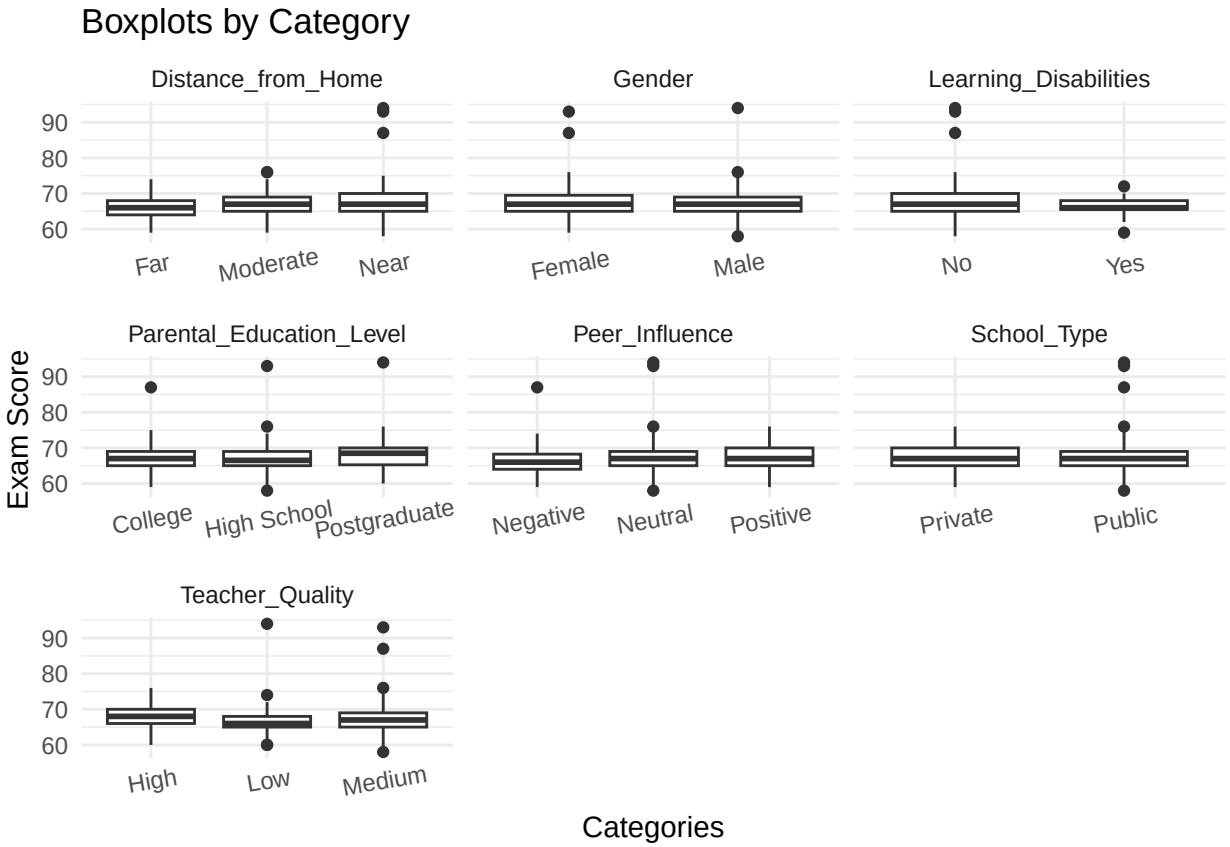
- On the other hand, Sleep_Hours have no significant linear relationship with Exam_Scores, when no other predictors are in the model

---

3.2 Categorical Variables

Boxplots provide a good visualization of the categorical data. It can indicate where the mean of the data is, where majority of the datapoints are, and outliers in the data.





Based on the plots, we can have the following insights:

- The mean level of Motivation_Level, School_Type, and Gender variables are the same across classes. The standard error of the distribution also appears similar between classes of these predictors. These plots may mean there is no significant difference in Exam_Scores between classes of these predictors.
- For the rest of the categorical predictors, there is visible difference in the mean or the standard error between classes.

These will be further tested using Single Linear Regression. Here, each categorical variable will be the single predictor in the regression model. This will only test whether there is significant difference in mean or intercept of the classes of these predictors:

Note: This summary is obtained by fitting a Single Linear Regression model per categorical predictor, but summarized in just 1 table for convenient comparison.

Based on these results, we can infer the following:

- Access_to_Resources and Distance_from_Home are significant categorical predictors since each dummy variable related to these predictors have low p-values
- On the other hand, Motivation_Level, Learning_Disabilities, and Gender are not significant predictors since each dummy variable related to these predictors have high p-values
- The remaining predictors have a dummy variable that has low p-value and another dummy variable with high p-value.

---

term	estimate	std.error	statistic	p.value
Parental_InvolvementLow	-1.2862	0.4780	-2.6911	0.0074
Parental_InvolvementMedium	-0.3294	0.3922	-0.8398	0.4014
Access_to_ResourcesLow	-2.3173	0.4638	-4.9962	0.0000
Access_to_ResourcesMedium	-1.1119	0.3801	-2.9251	0.0036
Extracurricular_ActivitiesYes	0.8117	0.3410	2.3807	0.0177
Motivation_LevelLow	-0.4737	0.4844	-0.9778	0.3286
Motivation_LevelMedium	-0.1493	0.4449	-0.3356	0.7373
Internet_AccessYes	0.9543	0.6253	1.5262	0.1276
Family_IncomeLow	-1.3347	0.4761	-2.8030	0.0053
Family_IncomeMedium	-0.7505	0.4821	-1.5566	0.1202
Teacher_QualityLow	-0.7475	0.6437	-1.1612	0.2461
Teacher_QualityMedium	-0.8933	0.3781	-2.3627	0.0185
School_TypePublic	-0.2512	0.3662	-0.6860	0.4931
Peer_InfluenceNeutral	1.0280	0.4610	2.2299	0.0262
Peer_InfluencePositive	0.7885	0.4562	1.7282	0.0846
Learning_DisabilitiesYes	-0.9287	0.6651	-1.3963	0.1632
Parental_Education_LevelHigh School	-0.3049	0.3799	-0.8024	0.4227
Parental_Education_LevelPostgraduate	1.1447	0.4944	2.3153	0.0210
Distance_from_HomeModerate	1.3486	0.5818	2.3181	0.0209
Distance_from_HomeNear	1.9403	0.5396	3.5961	0.0004
GenderMale	-0.1936	0.3433	-0.5640	0.5730

4 Multiple Linear Regression

4.1 Full Model

We will first consider the full model using all predictors.

```
##
## Call:
## lm(formula = Exam_Score ~ ., data = trainData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7316 -0.4807 -0.1519  0.2341 25.8693
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     40.810005   1.170006  34.880  < 2e-16 ***
## Hours_Studied                    0.299757   0.014588  20.549  < 2e-16 ***
## Attendance                       0.200853   0.007713  26.041  < 2e-16 ***
## Parental_InvolvementLow         -1.748667   0.247971  -7.052 6.31e-12 ***
## Parental_InvolvementMedium      -0.878136   0.203002  -4.326 1.86e-05 ***
## Access_to_ResourcesLow          -1.937808   0.248053  -7.812 3.68e-14 ***
## Access_to_ResourcesMedium       -0.703941   0.204831  -3.437 0.000641 ***
## Extracurricular_ActivitiesYes    0.757211   0.176367   4.293 2.14e-05 ***
## Sleep_Hours                      0.034418   0.059537   0.578 0.563470
## Previous_Scores                  0.042304   0.006304   6.711 5.56e-11 ***
```

```
## Motivation_LevelLow                  -1.089520   0.249205  -4.372 1.52e-05 ***
## Motivation_LevelMedium               -0.460565   0.228744  -2.013 0.044634 *
## Internet_AccessYes                    1.180319   0.321328   3.673 0.000267 ***
## Tutoring_Sessions                     0.455978   0.072797   6.264 8.47e-10 ***
## Family_IncomeLow                     -1.351811   0.246743  -5.479 6.99e-08 ***
## Family_IncomeMedium                  -0.950575   0.248907  -3.819 0.000152 ***
## Teacher_QualityLow                   -0.637784   0.333761  -1.911 0.056623 .
## Teacher_QualityMedium                -0.365224   0.195952  -1.864 0.062964 .
## School_TypePublic                     0.207426   0.189781   1.093 0.274962
## Peer_InfluenceNeutral                 0.440130   0.242210   1.817 0.069829 .
## Peer_InfluencePositive                0.662559   0.238535   2.778 0.005694 **
## Physical_Activity                     0.318773   0.082833   3.848 0.000135 ***
## Learning_DisabilitiesYes             -1.191282   0.347982  -3.423 0.000672 ***
## Parental_Education_LevelHigh School  -0.617984   0.197170  -3.134 0.001830 **
## Parental_Education_LevelPostgraduate  0.552663   0.257016   2.150 0.032038 *
## Distance_from_HomeModerate            0.540532   0.303718   1.780 0.075766 .
## Distance_from_HomeNear                1.233677   0.282488   4.367 1.55e-05 ***
## GenderMale                           -0.140958   0.175998  -0.801 0.423589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.918 on 472 degrees of freedom
## Multiple R-squared:  0.7589, Adjusted R-squared:  0.7451
## F-statistic: 55.03 on 27 and 472 DF,  p-value: < 2.2e-16
```

4.1.1 Model Evaluation

The Full model has achieved an R_a^2 of 0.75. This means the full model explains almost 75 % of the variability of the response variable.

Based on the p-value of the F-statistic ($< 2.2e^{-16}$) , at least one of the predictors has significant linear relationship with exam_score. This can also be seen by the number of variables that have below 5% p-value.

There are also a few predictors with p-value greater than 5%, which means these predictors do not have significant linear relationship with exam scores given that the other predictors are already in the model.

Thus, we will perform variable/model selection.

However, before doing so, we will evaluate the MSE of this full model on unseen data (test dataset). This test error will be used as the baseline performance.

The MSE of the full model is 4.7566.

4.2 Stepwise - AIC

Instead of modelling all possible combinations of the 19 variables, we will use stepwise selection to identify the best models. We prefer stepwise selection instead backward selection because it is possible for predictors already removed to still be added back if they can improve our evaluation metric. Likewise, it is preferable than forward selection for the similar reason. We will use AIC to select the best model.

The AIC or Akaike Information Criterion minimizes mean squared error while also penalizing complexity of the model. This is defined by the following formula:

$$AIC = n\ ln\left(\frac{SSE}{n}\right) + 2p$$

term	Full Model	Stepwise - AIC
(Intercept)	40.81000476	41.1801855
Hours_Studied	0.29975654	0.2985929
Attendance	0.20085256	0.2004258
Parental_InvolvementLow	-1.74866682	-1.7449619
Parental_InvolvementMedium	-0.87813570	-0.8855843
Access_to_ResourcesLow	-1.93780776	-1.9315804
Access_to_ResourcesMedium	-0.70394108	-0.7264337
Extracurricular_ActivitiesYes	0.75721119	0.7623519
Sleep_Hours	0.03441834	NA
Previous_Scores	0.04230428	0.0419970
Motivation_LevelLow	-1.08951959	-1.1039822
Motivation_LevelMedium	-0.46056501	-0.4483638
Internet_AccessYes	1.18031911	1.1806072
Tutoring_Sessions	0.45597793	0.4585026
Family_IncomeLow	-1.35181140	-1.3318762
Family_IncomeMedium	-0.95057519	-0.9607421
Teacher_QualityLow	-0.63778366	-0.6339541
Teacher_QualityMedium	-0.36522366	-0.3579991
School_TypePublic	0.20742609	NA
Peer_InfluenceNeutral	0.44013041	0.4356338
Peer_InfluencePositive	0.66255901	0.6595680
Physical_Activity	0.31877345	0.3219211
Learning_DisabilitiesYes	-1.19128166	-1.1665347
Parental_Education_LevelHigh School	-0.61798432	-0.6142759
Parental_Education_LevelPostgraduate	0.55266287	0.5380331
Distance_from_HomeModerate	0.54053191	0.5321244
Distance_from_HomeNear	1.23367651	1.2403961
GenderMale	-0.14095815	NA

where SSE is the sum of squared error of the model, n is the number of observations in the training data, and p is the number of predictors in the model.

- Of the 27 predictors in the Full model, 3 were removed by the criterion: Sleep_Hours, School_TypePublic, and Gender_Male.

4.2.1 Model Comparison

We will compare the full model with the model from stepwise selection using AIC.

```
## Analysis of Variance Table
##
## Model 1: Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement +
##     Access_to_Resources + Extracurricular_Activities + Previous_Scores +
##     Motivation_Level + Internet_Access + Tutoring_Sessions +
##     Family_Income + Teacher_Quality + Peer_Influence + Physical_Activity +
##     Learning_Disabilities + Parental_Education_Level + Distance_from_Home
## Model 2: Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement +
```

```
##     Access_to_Resources + Extracurricular_Activities + Sleep_Hours +
##     Previous_Scores + Motivation_Level + Internet_Access + Tutoring_Sessions +
##     Family_Income + Teacher_Quality + School_Type + Peer_Influence +
##     Physical_Activity + Learning_Disabilities + Parental_Education_Level +
##     Distance_from_Home + Gender
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    475 1743.9
## 2    472 1735.7  3    8.2301 0.746 0.5251
```

- Using the Analysis of Variance (Anova), we will compare the full model and reduced model under the main hypothesis that the coefficients of predictors removed from the full model are equal to 0. Since the p-value from the F-test is high, we fail to reject H_0 which means the reduced model is enough to predict Exam_Scores..
- We prefer the model from stepwise selection using AIC.

4.2.2 Model Evaluation

```
##
## Call:
## lm(formula = Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement +
##     Access_to_Resources + Extracurricular_Activities + Previous_Scores +
##     Motivation_Level + Internet_Access + Tutoring_Sessions +
##     Family_Income + Teacher_Quality + Peer_Influence + Physical_Activity +
##     Learning_Disabilities + Parental_Education_Level + Distance_from_Home,
##     data = trainData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5636 -0.4679 -0.1295  0.2099 25.9972
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       41.180185   1.031929  39.906  < 2e-16 ***
## Hours_Studied                      0.298593   0.014555  20.515  < 2e-16 ***
## Attendance                         0.200426   0.007663  26.155  < 2e-16 ***
## Parental_InvolvementLow           -1.744962   0.247688  -7.045 6.55e-12 ***
## Parental_InvolvementMedium        -0.885584   0.202437  -4.375 1.50e-05 ***
## Access_to_ResourcesLow            -1.931580   0.246473  -7.837 3.06e-14 ***
## Access_to_ResourcesMedium         -0.726434   0.203955  -3.562 0.000406 ***
## Extracurricular_ActivitiesYes      0.762352   0.175969   4.332 1.80e-05 ***
## Previous_Scores                    0.041997   0.006269   6.699 5.94e-11 ***
## Motivation_LevelLow               -1.103982   0.248522  -4.442 1.11e-05 ***
## Motivation_LevelMedium            -0.448364   0.227836  -1.968 0.049658 *
## Internet_AccessYes                 1.180607   0.320205   3.687 0.000253 ***
## Tutoring_Sessions                  0.458503   0.072609   6.315 6.22e-10 ***
## Family_IncomeLow                  -1.331876   0.246170  -5.410 9.99e-08 ***
## Family_IncomeMedium               -0.960742   0.248272  -3.870 0.000124 ***
## Teacher_QualityLow                -0.633954   0.333083  -1.903 0.057607 .
## Teacher_QualityMedium             -0.357999   0.195055  -1.835 0.067076 .
## Peer_InfluenceNeutral              0.435634   0.240794   1.809 0.071059 .
## Peer_InfluencePositive             0.659568   0.237976   2.772 0.005798 **
## Physical_Activity                  0.321921   0.082341   3.910 0.000106 ***
## Learning_DisabilitiesYes          -1.166535   0.346977  -3.362 0.000836 ***
## Parental_Education_LevelHigh School -0.614276   0.196739  -3.122 0.001904 **
```

model	adj_R2	train_MSE	test_MSE
Full Model	0.7451	3.4713	4.7566
Stepwise - AIC	0.7455	3.4878	4.7703

```
## Parental_Education_LevelPostgraduate  0.538033   0.256385   2.099 0.036386 *
## Distance_from_HomeModerate            0.532124   0.301773   1.763 0.078489 .
## Distance_from_HomeNear                1.240396   0.281189   4.411 1.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.916 on 475 degrees of freedom
## Multiple R-squared:  0.7578, Adjusted R-squared:  0.7455
## F-statistic: 61.91 on 24 and 475 DF,  p-value: < 2.2e-16
```

The model from stepwise selection using AIC has achieved an R_a^2 of 0.75. This means the full model explains almost 75 % of the variability of the response variable.

Based on the p-value of the F-statistic ($< 2.2e^{-16}$) , at least one of the predictors has significant linear relationship with exam_score. This can also be seen by the number of variables that have below 5% p-value. Only 4 predictors have p-values greater than 5% but all of these have at most 10% p-value.

The MSE of the reduced model (AIC) is 4.7703.

The model from stepwise selection achieved higher R_a^2 but slightly higher test MSE. Overall, the difference between the Full model and Stepwise - AIC model is not significant, as supported by anova.

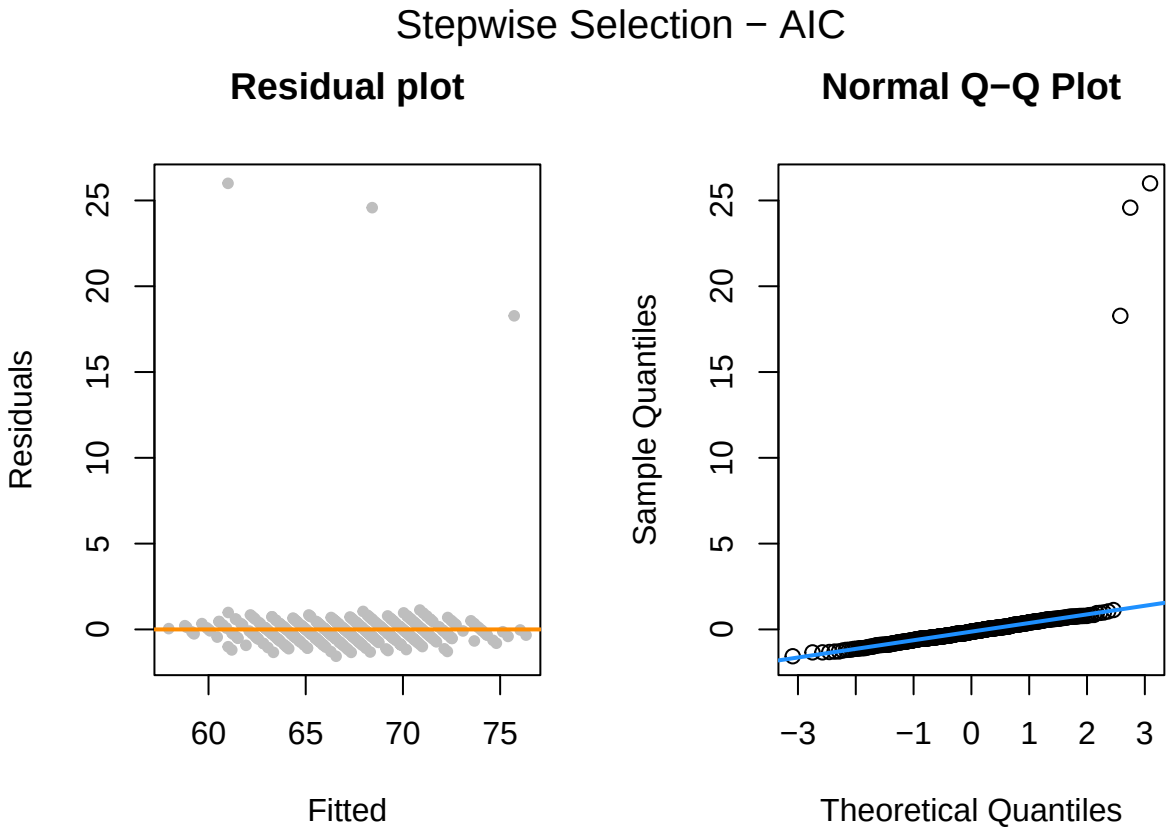
4.2.3 Model Diagnostics

We will check presence of collinearity in the model.

maxVIF
2.608632

- Since the maximum VIF among all predictors in the model is only 2.61, there is no significant collinearity in the model.

Let's check if the linearity, equal variance, and normality assumptions hold for the Stepwise - AIC model.



- The mean of the residuals seem to be 0 but there are a few data points with very high residual values.
- The outer regions in the graph have smaller spread than the spread of residuals in the middle.
- Majority of the data points seem to lie in the theoretical quantile of the Normal distribution, but this still needs to be verified.

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_stepwise_AIC
## BP = 20.37, df = 24, p-value = 0.6755
```

- From the results of the BP test, the p-value is 0.6755 which is higher than $\alpha = 0.05$. Thus, we fail to reject H_0 that the variance of the residuals is constant.

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_stepwise_AIC)
## W = 0.22102, p-value < 2.2e-16
```

- From the Shapiro-Wilk test, since the p-value is $< 2.2e^{-16}$, at $\alpha = 0.05$, we reject H_0 that the residuals are from the $Normal$ distribution.

We will explore the presence of influential points:

```
## 4435 2597 4261
##  357  402  413
```

- There are 3 influential points in the dataset based on cook's distance

```
##      Hours_Studied Attendance Parental_Involvement Exam_Score
## 4435            25         73               Medium         93
## 2597            11         71                  Low         87
## 4261            25         98               Medium         94
```

- All influential points are also considered outliers

4.3 Stepwise - AIC with no Influential Points

While we do not have sufficient background regarding the data collection process, we will still explore the impact of these influential points. These points will be removed, then we will refit the Stepwise - AIC model, evaluate, and perform model diagnostics.

```
##
## Call:
## lm(formula = formula(lm_stepwise_AIC), data = trainData_AIC_noInf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6710 -0.2438 -0.0100  0.2503  0.7951
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       40.827609   0.174413  234.09   <2e-16 ***
## Hours_Studied                      0.297285   0.002458  120.92   <2e-16 ***
## Attendance                         0.201249   0.001294  155.47   <2e-16 ***
## Parental_InvolvementLow           -1.964617   0.041837  -46.96   <2e-16 ***
## Parental_InvolvementMedium        -1.016828   0.034121  -29.80   <2e-16 ***
## Access_to_ResourcesLow            -1.940378   0.041518  -46.74   <2e-16 ***
## Access_to_ResourcesMedium         -0.979317   0.034409  -28.46   <2e-16 ***
## Extracurricular_ActivitiesYes      0.489947   0.029722   16.48   <2e-16 ***
## Previous_Scores                    0.049668   0.001061   46.83   <2e-16 ***
## Motivation_LevelLow               -0.981622   0.041983  -23.38   <2e-16 ***
## Motivation_LevelMedium            -0.450556   0.038618  -11.67   <2e-16 ***
## Internet_AccessYes                 1.009938   0.053946   18.72   <2e-16 ***
## Tutoring_Sessions                  0.499994   0.012240   40.85   <2e-16 ***
## Family_IncomeLow                  -0.946763   0.041767  -22.67   <2e-16 ***
## Family_IncomeMedium               -0.493528   0.042085  -11.73   <2e-16 ***
## Teacher_QualityLow                -1.018429   0.056607  -17.99   <2e-16 ***
## Teacher_QualityMedium             -0.539472   0.032883  -16.41   <2e-16 ***
## Peer_InfluenceNeutral              0.469782   0.040700   11.54   <2e-16 ***
## Peer_InfluencePositive             0.951114   0.040198   23.66   <2e-16 ***
## Physical_Activity                  0.237858   0.013887   17.13   <2e-16 ***
## Learning_DisabilitiesYes          -0.928465   0.058473  -15.88   <2e-16 ***
## Parental_Education_LevelHigh School -0.517486  0.033200  -15.59   <2e-16 ***
## Parental_Education_LevelPostgraduate  0.462199  0.043395  10.65   <2e-16 ***
## Distance_from_HomeModerate         0.568126   0.050828   11.18   <2e-16 ***
## Distance_from_HomeNear             1.051405   0.047384   22.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3227 on 472 degrees of freedom
## Multiple R-squared:  0.9909, Adjusted R-squared:  0.9905
## F-statistic:  2145 on 24 and 472 DF,  p-value: < 2.2e-16
```

term	Full Model	Stepwise - AIC	Stepwise - AIC - noInf
(Intercept)	40.81000476	41.1801855	40.82760906
Hours_Studied	0.29975654	0.2985929	0.29728531
Attendance	0.20085256	0.2004258	0.20124902
Parental_InvolvementLow	-1.74866682	-1.7449619	-1.96461711
Parental_InvolvementMedium	-0.87813570	-0.8855843	-1.01682775
Access_to_ResourcesLow	-1.93780776	-1.9315804	-1.94037756
Access_to_ResourcesMedium	-0.70394108	-0.7264337	-0.97931685
Extracurricular_ActivitiesYes	0.75721119	0.7623519	0.48994749
Sleep_Hours	0.03441834	NA	NA
Previous_Scores	0.04230428	0.0419970	0.04966763
Motivation_LevelLow	-1.08951959	-1.1039822	-0.98162163
Motivation_LevelMedium	-0.46056501	-0.4483638	-0.45055644
Internet_AccessYes	1.18031911	1.1806072	1.00993760
Tutoring_Sessions	0.45597793	0.4585026	0.49999379
Family_IncomeLow	-1.35181140	-1.3318762	-0.94676274
Family_IncomeMedium	-0.95057519	-0.9607421	-0.49352771
Teacher_QualityLow	-0.63778366	-0.6339541	-1.01842868
Teacher_QualityMedium	-0.36522366	-0.3579991	-0.53947216
School_TypePublic	0.20742609	NA	NA
Peer_InfluenceNeutral	0.44013041	0.4356338	0.46978220
Peer_InfluencePositive	0.66255901	0.6595680	0.95111432
Physical_Activity	0.31877345	0.3219211	0.23785794
Learning_DisabilitiesYes	-1.19128166	-1.1665347	-0.92846536
Parental_Education_LevelHigh School	-0.61798432	-0.6142759	-0.51748618
Parental_Education_LevelPostgraduate	0.55266287	0.5380331	0.46219865
Distance_from_HomeModerate	0.54053191	0.5321244	0.56812624
Distance_from_HomeNear	1.23367651	1.2403961	1.05140517
GenderMale	-0.14095815	NA	NA

model	adj_R2	train_MSE	test_MSE
Full Model	0.7451	3.4713	4.7566
Stepwise - AIC	0.7455	3.4878	4.7703
Stepwise - AIC - noInf	0.9905	0.0989	4.8330

4.3.1 Model Evaluation

The model from stepwise selection using AIC applied on the data without influential points has achieved an R_a^2 of 0.9904517. This means the full model explains almost 0.99 % of the variability of the response variable.

Based on the p-value of the F-statistic ($< 2.2e^{-16}$) , at least one of the predictors has significant linear relationship with exam_score. All predictors have below 5% p-value.

The MSE of the reduced model (AIC) with no Influential points is 4.833.

The model from stepwise selection without influential points achieved significantly higher R_a^2 and lower train MSE. However, the decrease in MSE may be attributed to the removal of influential points. Further, while it

is expected for test MSE to be greater than train MSE, the difference between the two suggest there may be overfitting in the model.

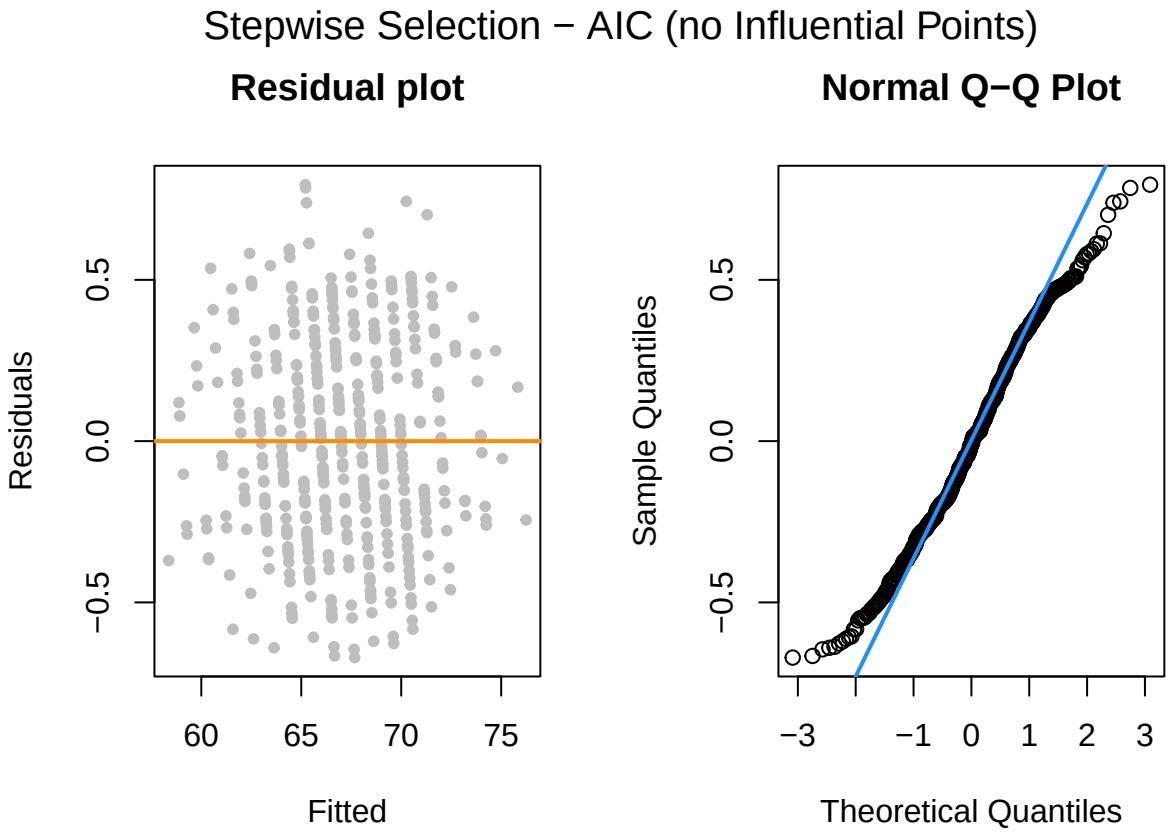
4.3.2 Model Diagnostics

We will check presence of collinearity in the model.

maxVIF
2.600575

- Since the maximum VIF among all predictors in the model is only 2.60, there is no significant collinearity in the model.

Let's check if the linearity, equal variance, and normality assumptions hold for the Stepwise - AIC model without the influential points.



- After removal of influential points, the mean of the residuals seem to be 0 at any region of the residual plot. Linearity seems to hold,
- The spread of the data points look constant throughout the residual plot.
- However, when influential points were removed, it is more observable that the empirical distribution of residuals do not follow the normal distribution.

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_stepwise_AIC_noInf
```

```
## BP = 9.3731, df = 24, p-value = 0.9967
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_stepwise_AIC_noInf)
## W = 0.98569, p-value = 8.351e-05
```

- Based on BP test, the equal variance assumption holds.
- Based on Shapiro-Wilk test, the Normality assumption does not hold.

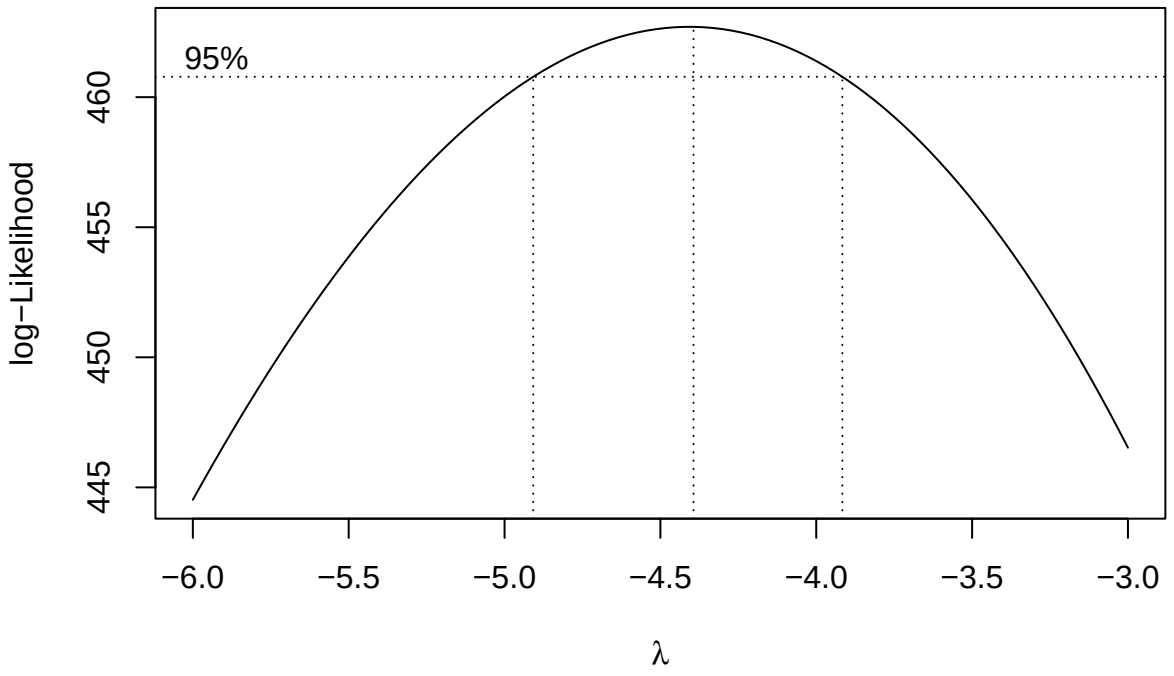
The influential points were removed to better see the model diagnostics of the current best model. It turns out, the normality assumption does not hold since the empirical distribution of the data does not follow the quantile plot of normal distribution. We will do response transformation using box-cox using the entire data (including influential points).

4.4 Stepwise - AIC with Box-Cox Transformation

4.4.1 Response Transformation - Box Cox

Since the spread of the residuals appear the same throughout the residual plot, and was also confirmed by the BP test, we will not use the variance stabilizing transformation methodologies. Instead, the Box-Cox transformation will be used as this often resolves the normality assumption.

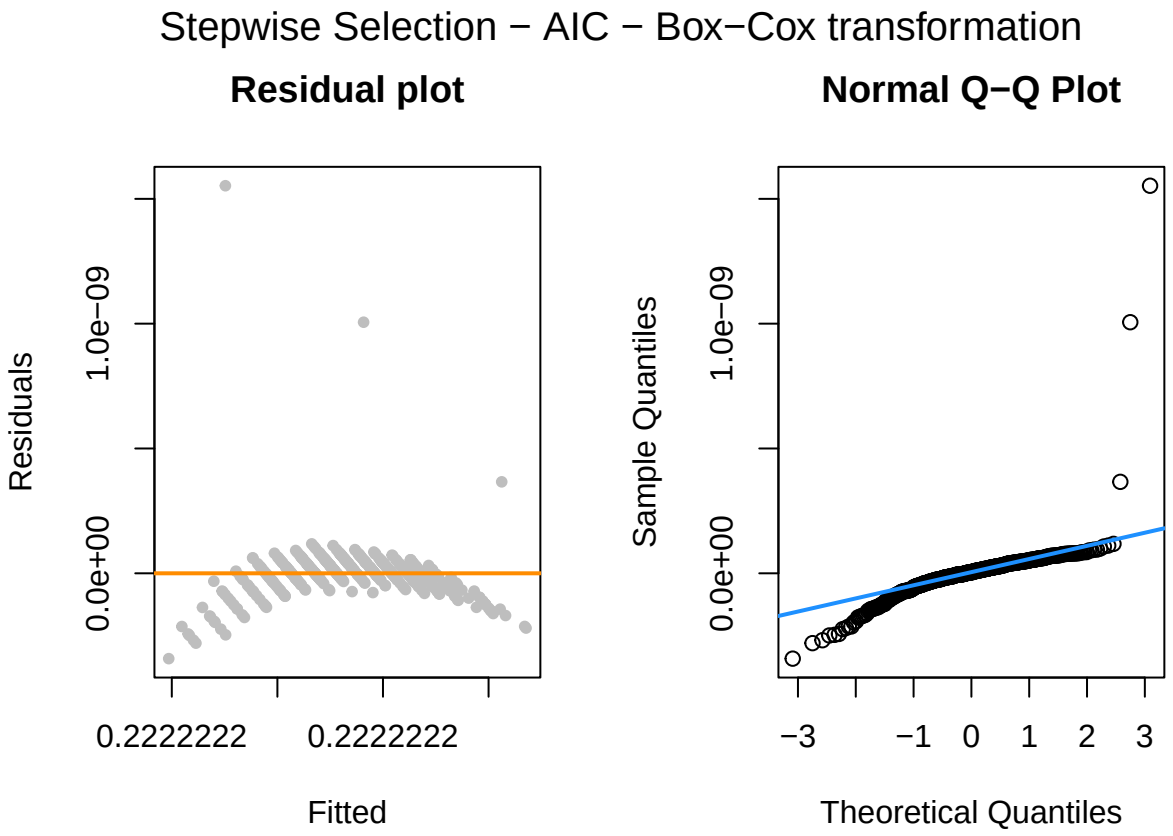
We will still consider the entire training data in this response transformation, since the influential points identified in the previous model are specific to the model only.



```r
lambda_boxcox = -4.5
```

```
##
## Call:
## lm(formula = sw_AIC_boxcox_formula, data = trainData)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.418e-10 -3.081e-11  4.990e-12  4.017e-11  1.552e-09
##
## Coefficients:
##                                    Estimate Std. Error    t value Pr(>|t|)
## (Intercept)                       2.222e-01  5.909e-11  3.761e+09  < 2e-16
## Hours_Studied                     2.786e-11  8.334e-13  3.342e+01  < 2e-16
## Attendance                        1.873e-11  4.388e-13  4.269e+01  < 2e-16
## Parental_InvolvementLow          -1.685e-10  1.418e-11 -1.188e+01  < 2e-16
## Parental_InvolvementMedium       -8.231e-11  1.159e-11 -7.100e+00 4.57e-12
## Access_to_ResourcesLow           -1.855e-10  1.411e-11 -1.314e+01  < 2e-16
## Access_to_ResourcesMedium        -6.919e-11  1.168e-11 -5.924e+00 6.02e-09
## Extracurricular_ActivitiesYes     5.983e-11  1.008e-11  5.938e+00 5.58e-09
## Previous_Scores                   4.118e-12  3.590e-13  1.147e+01  < 2e-16
## Motivation_LevelLow              -1.065e-10  1.423e-11 -7.485e+00 3.50e-13
## Motivation_LevelMedium           -4.203e-11  1.305e-11 -3.222e+00 0.001361
## Internet_AccessYes                1.077e-10  1.834e-11  5.876e+00 7.91e-09
## Tutoring_Sessions                 4.255e-11  4.158e-12  1.023e+01  < 2e-16
## Family_IncomeLow                 -8.779e-11  1.410e-11 -6.228e+00 1.04e-09
## Family_IncomeMedium              -5.274e-11  1.422e-11 -3.710e+00 0.000232
## Teacher_QualityLow               -8.597e-11  1.907e-11 -4.508e+00 8.27e-06
## Teacher_QualityMedium            -4.445e-11  1.117e-11 -3.979e+00 7.99e-05
## Peer_InfluenceNeutral             3.339e-11  1.379e-11  2.422e+00 0.015825
## Peer_InfluencePositive            6.932e-11  1.363e-11  5.087e+00 5.25e-07
## Physical_Activity                 2.591e-11  4.715e-12  5.495e+00 6.39e-08
## Learning_DisabilitiesYes         -8.134e-11  1.987e-11 -4.094e+00 4.99e-05
## Parental_Education_LevelHigh School -5.297e-11  1.127e-11 -4.702e+00 3.39e-06
## Parental_Education_LevelPostgraduate  3.475e-11  1.468e-11  2.367e+00 0.018332
## Distance_from_HomeModerate        5.837e-11  1.728e-11  3.378e+00 0.000790
## Distance_from_HomeNear            1.151e-10  1.610e-11  7.148e+00 3.34e-12
##
## (Intercept)                      ***
## Hours_Studied                    ***
## Attendance                       ***
## Parental_InvolvementLow          ***
## Parental_InvolvementMedium       ***
## Access_to_ResourcesLow           ***
## Access_to_ResourcesMedium        ***
## Extracurricular_ActivitiesYes    ***
## Previous_Scores                  ***
## Motivation_LevelLow              ***
## Motivation_LevelMedium           **
## Internet_AccessYes               ***
## Tutoring_Sessions                ***
## Family_IncomeLow                 ***
## Family_IncomeMedium              ***
## Teacher_QualityLow               ***
```

```
## Teacher_QualityMedium                ***
## Peer_InfluenceNeutral                *
## Peer_InfluencePositive               ***
## Physical_Activity                    ***
## Learning_DisabilitiesYes             ***
## Parental_Education_LevelHigh School  ***
## Parental_Education_LevelPostgraduate *
## Distance_from_HomeModerate           ***
## Distance_from_HomeNear               ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.097e-10 on 475 degrees of freedom
## Multiple R-squared:  0.8912, Adjusted R-squared:  0.8857
## F-statistic: 162.1 on 24 and 475 DF,  p-value: < 2.2e-16
```




```
# Test Equal Variance
bptest(lm_stepwise_AIC_boxcox)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_stepwise_AIC_boxcox
## BP = 20.951, df = 24, p-value = 0.6416
```

```
# Test Normality
shapiro.test(resid(lm_stepwise_AIC_boxcox))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_stepwise_AIC_boxcox)
## W = 0.53928, p-value < 2.2e-16
```

After box-cox transformation, the results show that the linearity assumption does not hold anymore. Further, it did not solve the normality assumption.

In the box-cox transformation formula, the lambda value of -4.5 is will cause the y values to shrink significantly. In the model summary, it is observable that the coefficients of the predictors became significantly small (at least 10^{-10}) and predictions are mainly driven by the intercept only. When further examined, a y value of 60 will be transformed to 0.222222 while a value of 100 will be transformed to 0.222222 as well. Since the range of values of y in the original data is only from 60 to 100, the transformed values have no variance anymore. This explains the bad model diagnostics.

The best model is still the model prior to box-cox transformation when influential points were removed.

Note that the R^2 and R^2_a of this model is 0.9901 and 0.9905, respectively. These suggest that the current model explains almost the entire variability of the response variable. This may also mean the model is overfitting the data since only 1% is not explained by the model, given the many predictors in the data.

In the anova between AIC and BIC above, the p-value of the f-test is small, which suggests the full model is better than the reduced model. But because the model assumptions were violated, and the box-cox transformation did not resolve this assumption, we will explore the backward selection model using BIC.

4.5 Stepwise - BIC

The BIC or Bayesian Information Criterion minimizes mean squared error while also penalizing complexity of the model. This is defined by the following formula:

$$BIC = n\ ln\left(\frac{SSE}{n}\right) + ln(n)p$$

where SSE is the sum of squared error of the model, n is the number of observations in the training data, and p is the number of predictors in the model.

Again, we will use stepwise selection using the entire dataset since the influential points identified in the previous models are specific to the model.

- Of the 27 predictors in the Full model, 3 were removed by AIC: Sleep_Hours, School_TypePublic, and Gender_Male. 4 predictors were further removed by BIC: 2 dummy variables related to Teacher_Quality and 2 dummy variables related to Peer_Influence.

4.5.1 Model Comparison

We will compare the model from stepwise selection using AIC with the model from stepwise selection using BIC.

```
## Analysis of Variance Table
##
## Model 1: Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement +
##     Access_to_Resources + Extracurricular_Activities + Previous_Scores +
##     Motivation_Level + Internet_Access + Tutoring_Sessions +
##     Family_Income + Physical_Activity + Learning_Disabilities +
##     Parental_Education_Level + Distance_from_Home
## Model 2: Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement +
```

term	Full Model	Stepwise - AIC	Stepwise - AIC - noInf	Stepw
(Intercept)	40.81000476	41.1801855	40.82760906	41.
Hours_Studied	0.29975654	0.2985929	0.29728531	0.
Attendance	0.20085256	0.2004258	0.20124902	0.
Parental_InvolvementLow	-1.74866682	-1.7449619	-1.96461711	-1.
Parental_InvolvementMedium	-0.87813570	-0.8855843	-1.01682775	-0.
Access_to_ResourcesLow	-1.93780776	-1.9315804	-1.94037756	-1.
Access_to_ResourcesMedium	-0.70394108	-0.7264337	-0.97931685	-0.
Extracurricular_ActivitiesYes	0.75721119	0.7623519	0.48994749	0.
Sleep_Hours	0.03441834	NA	NA	
Previous_Scores	0.04230428	0.0419970	0.04966763	0.
Motivation_LevelLow	-1.08951959	-1.1039822	-0.98162163	-1.
Motivation_LevelMedium	-0.46056501	-0.4483638	-0.45055644	-0.
Internet_AccessYes	1.18031911	1.1806072	1.00993760	1.
Tutoring_Sessions	0.45597793	0.4585026	0.49999379	0.
Family_IncomeLow	-1.35181140	-1.3318762	-0.94676274	-1.
Family_IncomeMedium	-0.95057519	-0.9607421	-0.49352771	-0.
Teacher_QualityLow	-0.63778366	-0.6339541	-1.01842868	
Teacher_QualityMedium	-0.36522366	-0.3579991	-0.53947216	
School_TypePublic	0.20742609	NA	NA	
Peer_InfluenceNeutral	0.44013041	0.4356338	0.46978220	
Peer_InfluencePositive	0.66255901	0.6595680	0.95111432	
Physical_Activity	0.31877345	0.3219211	0.23785794	0.
Learning_DisabilitiesYes	-1.19128166	-1.1665347	-0.92846536	-1.
Parental_Education_LevelHigh School	-0.61798432	-0.6142759	-0.51748618	-0.
Parental_Education_LevelPostgraduate	0.55266287	0.5380331	0.46219865	0.
Distance_from_HomeModerate	0.54053191	0.5321244	0.56812624	0.
Distance_from_HomeNear	1.23367651	1.2403961	1.05140517	1.
GenderMale	-0.14095815	NA	NA	

```
##     Access_to_Resources + Extracurricular_Activities + Previous_Scores +
##     Motivation_Level + Internet_Access + Tutoring_Sessions +
##     Family_Income + Teacher_Quality + Peer_Influence + Physical_Activity +
##     Learning_Disabilities + Parental_Education_Level + Distance_from_Home
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    479 1787.8
## 2    475 1743.9  4    43.887 2.9885 0.01866 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Using the Analysis of Variance (Anova), we will compare the full model (model from AIC) and reduced model (model from BIC) under the main hypothesis that the coefficients of predictors removed from the full model are equal to 0.
- Since the p-value from the F-test is less than 5%, we reject H_0. This means that there are predictors in the full model that were removed in the reduced mode that has significant linear relationship with Exam_Score. We still prefer the model from stepwise selection using AIC.

- Despite this result, the BIC model will be examined further as the model diagnostics from the model using AIC violated the normality assumption, which was not resolved by box-cox transformation.

4.5.2 Model Evaluation

```
##
## Call:
## lm(formula = Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement +
##     Access_to_Resources + Extracurricular_Activities + Previous_Scores +
##     Motivation_Level + Internet_Access + Tutoring_Sessions +
##     Family_Income + Physical_Activity + Learning_Disabilities +
##     Parental_Education_Level + Distance_from_Home, data = trainData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3264 -0.5731 -0.1166  0.3311 25.5684
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       41.297928   1.004513  41.112  < 2e-16 ***
## Hours_Studied                      0.302519   0.014623  20.688  < 2e-16 ***
## Attendance                         0.202027   0.007669  26.345  < 2e-16 ***
## Parental_InvolvementLow           -1.779626   0.249148  -7.143 3.42e-12 ***
## Parental_InvolvementMedium        -0.853270   0.203598  -4.191 3.31e-05 ***
## Access_to_ResourcesLow            -1.935943   0.246935  -7.840 2.95e-14 ***
## Access_to_ResourcesMedium         -0.764267   0.205058  -3.727 0.000217 ***
## Extracurricular_ActivitiesYes      0.715543   0.176177   4.061 5.70e-05 ***
## Previous_Scores                    0.041221   0.006282   6.561 1.39e-10 ***
## Motivation_LevelLow               -1.059273   0.249821  -4.240 2.68e-05 ***
## Motivation_LevelMedium            -0.445622   0.228480  -1.950 0.051713 .
## Internet_AccessYes                 1.203749   0.321708   3.742 0.000205 ***
## Tutoring_Sessions                  0.417912   0.072042   5.801 1.20e-08 ***
## Family_IncomeLow                  -1.350764   0.247435  -5.459 7.70e-08 ***
## Family_IncomeMedium               -0.959721   0.249429  -3.848 0.000135 ***
## Physical_Activity                  0.326309   0.082991   3.932 9.67e-05 ***
## Learning_DisabilitiesYes          -1.234322   0.348478  -3.542 0.000436 ***
## Parental_Education_LevelHigh School -0.644799   0.197619  -3.263 0.001182 **
## Parental_Education_LevelPostgraduate  0.513662   0.258314   1.989 0.047323 *
## Distance_from_HomeModerate         0.491786   0.303254   1.622 0.105526
## Distance_from_HomeNear             1.219088   0.283124   4.306 2.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.932 on 479 degrees of freedom
## Multiple R-squared:  0.7517, Adjusted R-squared:  0.7413
## F-statistic: 72.49 on 20 and 479 DF,  p-value: < 2.2e-16
```

The model from stepwise selection using BIC has achieved an R_a^2 of 0.74. This means the full model explains almost 74 % of the variability of the response variable.

Based on the p-value of the F-statistic ($< 2.2e^{-16}$) , at least one of the predictors has significant linear relationship with exam_score. This can also be seen by the number of variables that have below 5% p-value. Only 2 predictors have p-values greater than 5%, one of which has at most 10% p-value, and the other one at 10.6%.

The MSE of the full model is 4.9977.

model	adj_R2	train_MSE	test_MSE
Full Model	0.7451	3.4713	4.7566
Stepwise - AIC	0.7455	3.4878	4.7703
Stepwise - AIC - noInf	0.9905	0.0989	4.8330
Stepwise - BIC	0.7413	3.5756	4.9977

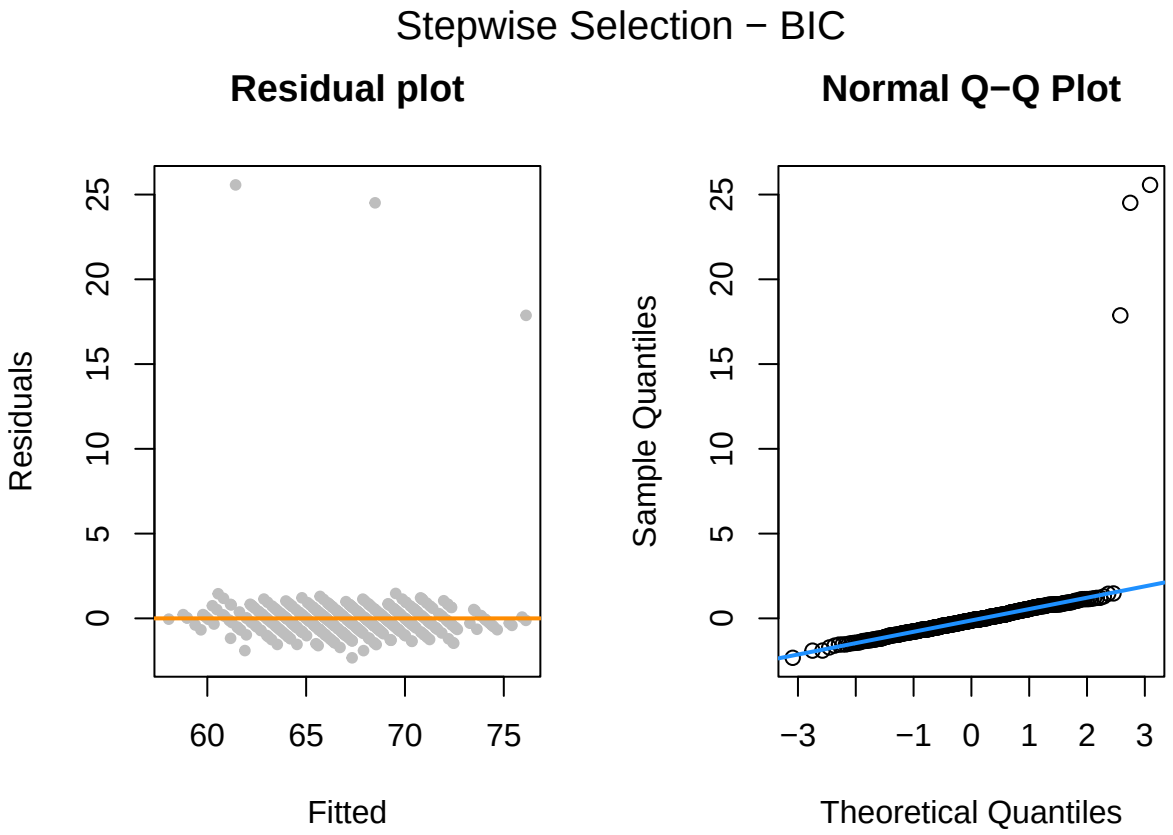
The model from stepwise selection achieved lower R_a^2 and slightly higher test MSE. As also shown by Anova, the model from AIC is still preferred as the predictors that were removed by BIC seem to have significant linear relationship with Exam_Score.

4.5.3 Model DIagnostics

We will check presence of collinearity in the model.

maxVIF
2.601464

- Since the maximum VIF among all predictors in the model is only 2.60, there is no significant collinearity in the model.



- The mean of the residuals seem to be 0 but there are a few data points with very high residual values.
- The outer regions in the graph have smaller spread than the spread of residuals in the middle.

Hours_Studied	Attendance	Parental_Involvement	Exam_Score
25	73	Medium	93
11	71	Low	87
25	98	Medium	94

- Majority of the data points seem to lie in the theoretical quantile of the Normal distribution, but this still needs to be verified.

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_stepwise_BIC
## BP = 17.159, df = 20, p-value = 0.6426
```

- From the results of the BP test, the p-value is 0.6426 which is higher than $\alpha = 0.05$. Thus, we fail to reject H_0 that the variance of the residuals is constant

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_stepwise_BIC)
## W = 0.28993, p-value < 2.2e-16
```

- From the Shapiro-Wilk test, since the p-value is $< 2.2e^{-16}$, at $\alpha = 0.05$, we reject H_0 that the residuals are from the $Normal$ distribution.

We will explore the presence of influential points:

```
## 4435 2597 4261
##  357  402  413
```

- There are 3 influential points in the dataset
- All influential points are also considered outliers

4.6 Stepwise - BIC with no Influential Points

Same as in AIC, while we do not have sufficient background regarding the data collection process, we will still explore the impact of these influential points. These points will be removed, then we will refit the Stepwise - BIC model, evaluate, and perform model diagnostics.

```
##
## Call:
## lm(formula = formula(lm_stepwise_BIC), data = trainData_BIC_noInf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51085 -0.40348 -0.00722  0.41304  1.31302
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     40.977374   0.291106 140.764  < 2e-16 ***
## Hours_Studied                    0.302721   0.004235  71.476  < 2e-16 ***
## Attendance                       0.203062   0.002222  91.371  < 2e-16 ***
## Parental_InvolvementLow         -2.021205   0.072128 -28.022  < 2e-16 ***
```

```
## Parental_InvolvementMedium           -0.975007   0.058827 -16.574  < 2e-16 ***
## Access_to_ResourcesLow               -1.951012   0.071291 -27.367  < 2e-16 ***
## Access_to_ResourcesMedium            -1.031919   0.059325 -17.394  < 2e-16 ***
## Extracurricular_ActivitiesYes         0.440963   0.051010   8.645  < 2e-16 ***
## Previous_Scores                       0.048132   0.001823  26.404  < 2e-16 ***
## Motivation_LevelLow                  -0.924377   0.072331 -12.780  < 2e-16 ***
## Motivation_LevelMedium               -0.466608   0.066292  -7.039 6.81e-12 ***
## Internet_AccessYes                    1.053420   0.092898  11.339  < 2e-16 ***
## Tutoring_Sessions                     0.439133   0.020809  21.103  < 2e-16 ***
## Family_IncomeLow                     -0.967384   0.072027 -13.431  < 2e-16 ***
## Family_IncomeMedium                  -0.490292   0.072539  -6.759 4.08e-11 ***
## Physical_Activity                     0.246243   0.023991  10.264  < 2e-16 ***
## Learning_DisabilitiesYes             -1.029391   0.100665 -10.226  < 2e-16 ***
## Parental_Education_LevelHigh School  -0.573284   0.057166 -10.028  < 2e-16 ***
## Parental_Education_LevelPostgraduate  0.430445   0.074970   5.742 1.67e-08 ***
## Distance_from_HomeModerate            0.525483   0.087557   6.002 3.88e-09 ***
## Distance_from_HomeNear                1.028480   0.081781  12.576  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5577 on 476 degrees of freedom
## Multiple R-squared:  0.9726, Adjusted R-squared:  0.9715
## F-statistic: 845.7 on 20 and 476 DF,  p-value: < 2.2e-16
```

```
gt(tidy_All |>
  dplyr::select(term, estimate, model) |>
  pivot_wider(names_from = model, values_from = estimate))
```

4.6.1 Model Evaluation

The model from stepwise selection using BIC applied on the data without influential points has achieved an R_a^2 of 0.97. This means the full model explains almost 97 % of the variability of the response variable.

Based on the p-value of the F-statistic ($< 2.2e^{-16}$) , at least one of the predictors has significant linear relationship with exam_score. All predictors have below 5% p-value.

The MSE of the full model is 5.1004.

The model from stepwise selection without influential points achieved significantly higher R_a^2 but slightly and lower MSE. However, the decrease in MSE may be attributed to the removal of influential points. Further, while it is expected for test MSE to be greater than train MSE, the difference between the two suggest there may be overfitting in the model.

4.6.2 Model Diagnostics

We will check presence of collinearity in the model.

```
flextable(tidy(vif(lm_stepwise_BIC_noInf)) |> summarise(maxVIF = max(x)))
```

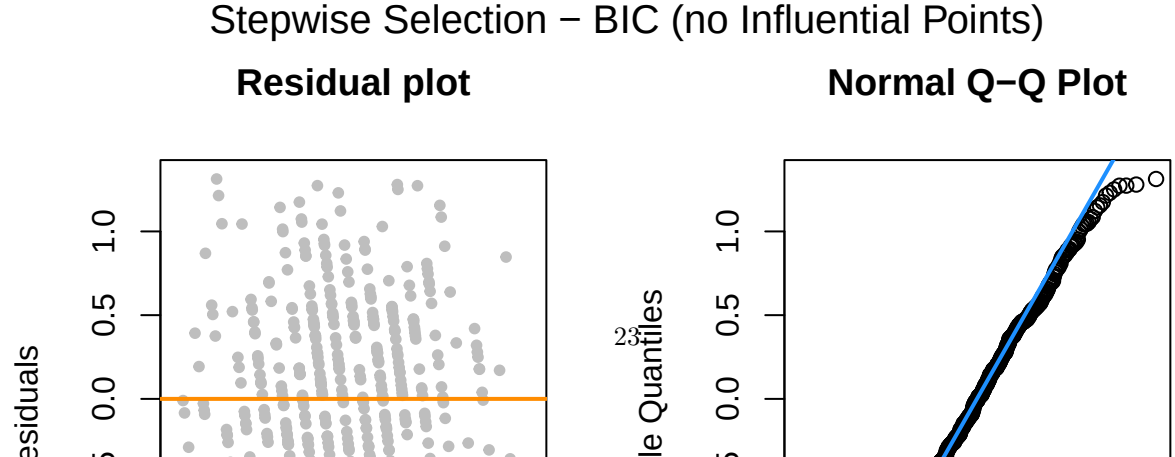
maxVIF
2.593299

- Since the maximum VIF among all predictors in the model is only 2.60, there is no significant collinearity in the model.

Let's check if the linearity, equal variance, and normality assumptions hold for the Stepwise - BIC model without the influential points.

term	Full Model	Stepwise - AIC	Stepwise - AIC - noInf	Stepw
(Intercept)	40.81000476	41.1801855	40.82760906	41.
Hours_Studied	0.29975654	0.2985929	0.29728531	0.
Attendance	0.20085256	0.2004258	0.20124902	0.
Parental_InvolvementLow	-1.74866682	-1.7449619	-1.96461711	-1.
Parental_InvolvementMedium	-0.87813570	-0.8855843	-1.01682775	-0.
Access_to_ResourcesLow	-1.93780776	-1.9315804	-1.94037756	-1.
Access_to_ResourcesMedium	-0.70394108	-0.7264337	-0.97931685	-0.
Extracurricular_ActivitiesYes	0.75721119	0.7623519	0.48994749	0.
Sleep_Hours	0.03441834	NA	NA	
Previous_Scores	0.04230428	0.0419970	0.04966763	0.
Motivation_LevelLow	-1.08951959	-1.1039822	-0.98162163	-1.
Motivation_LevelMedium	-0.46056501	-0.4483638	-0.45055644	-0.
Internet_AccessYes	1.18031911	1.1806072	1.00993760	1.
Tutoring_Sessions	0.45597793	0.4585026	0.49999379	0.
Family_IncomeLow	-1.35181140	-1.3318762	-0.94676274	-1.
Family_IncomeMedium	-0.95057519	-0.9607421	-0.49352771	-0.
Teacher_QualityLow	-0.63778366	-0.6339541	-1.01842868	
Teacher_QualityMedium	-0.36522366	-0.3579991	-0.53947216	
School_TypePublic	0.20742609	NA	NA	
Peer_InfluenceNeutral	0.44013041	0.4356338	0.46978220	
Peer_InfluencePositive	0.66255901	0.6595680	0.95111432	
Physical_Activity	0.31877345	0.3219211	0.23785794	0.
Learning_DisabilitiesYes	-1.19128166	-1.1665347	-0.92846536	-1.
Parental_Education_LevelHigh School	-0.61798432	-0.6142759	-0.51748618	-0.
Parental_Education_LevelPostgraduate	0.55266287	0.5380331	0.46219865	0.
Distance_from_HomeModerate	0.54053191	0.5321244	0.56812624	0.
Distance_from_HomeNear	1.23367651	1.2403961	1.05140517	1.
GenderMale	-0.14095815	NA	NA	

model	adj_R2	train_MSE	test_MSE
Full Model	0.7451	3.4713	4.7566
Stepwise - AIC	0.7455	3.4878	4.7703
Stepwise - AIC - noInf	0.9905	0.0989	4.8330
Stepwise - BIC	0.7413	3.5756	4.9977
Stepwise - BIC - noInf	0.9715	0.2979	3.5756



model	adj_R2	train_MSE	test_MSE
Full Model	0.7451	3.4713	4.7566
Stepwise - AIC	0.7455	3.4878	4.7703
Stepwise - AIC - noInf	0.9905	0.0989	4.8330
Stepwise - BIC	0.7413	3.5756	4.9977
Stepwise - BIC - noInf	0.9715	0.2979	3.5756

- After removal of influential points, the mean of the residuals seem to be 0 at any region of the residual plot. Linearity seems to hold,
- The spread of the data points look constant throughout the residual plot.
- When influential points were removed, it is more observable that the empirical distribution of residuals is close to the normal distribution.

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_stepwise_BIC_noInf
## BP = 24.043, df = 20, p-value = 0.2405
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_stepwise_BIC_noInf)
## W = 0.99553, p-value = 0.167
```

- Based on BP test, the equal variance assumption holds.
- Based on Shapiro-Wilk test, the Normality assumption already holds.

The influential points were removed to better see the model diagnostics of the current best model. It turns out, the linearity, equal variance and normality assumption already hold.

4.6.3 Test Results

Since the linearity, equal variance, and normality assumptions hold, this model will be further evaluated if it can generalize well on new data.

In the model evaluation summary above, we obtained the following results:

Without the influential points, the model achieved R_a^2 of 0.9714778 . This is lower than the R_a^2 obtained using AIC without the influential points. However, the difference between the train and test MSE is still significant. This shows the model may be overfitting the training data and cannot generalize well to unseen data.

Since we have already explored feature selection, we will now explore the shrinkage methods

---

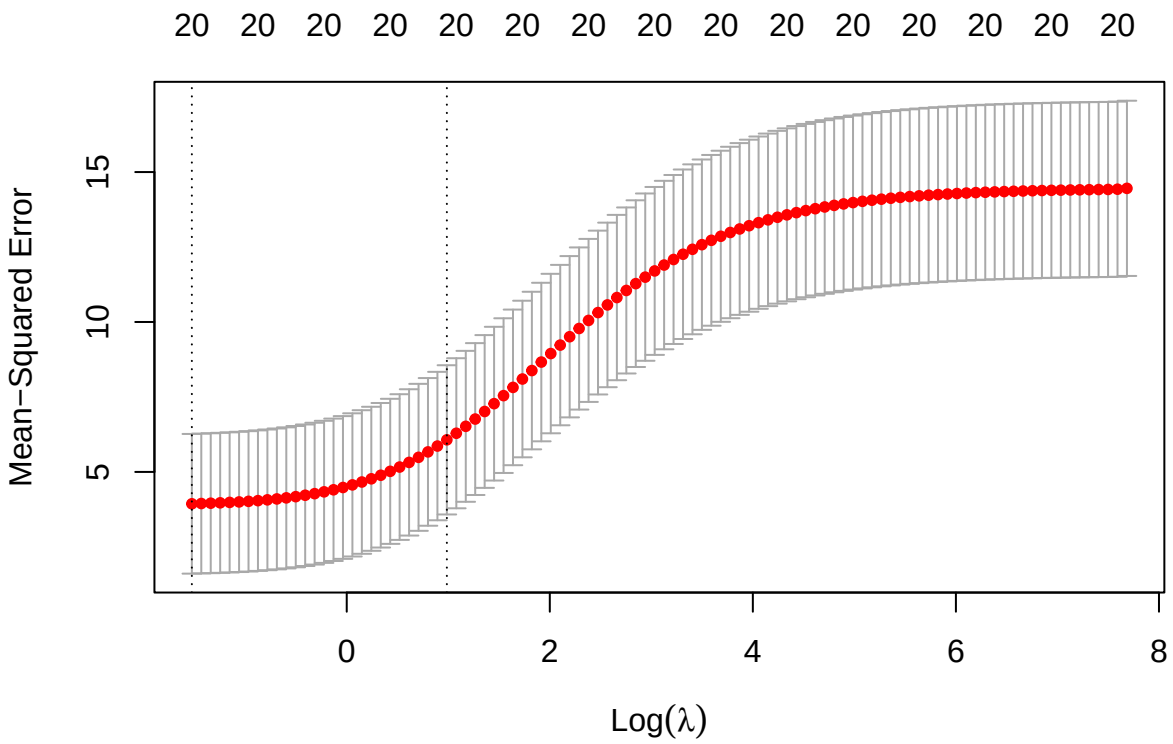
4.7 Ridge Regression

We will penalized multiple linear regression to the model that we obtained from BIC, since this model is the best model we obtained so far with good model diagnostics.

```
X_train <- model.matrix(formula(lm_stepwise_BIC),trainData)[, -1]
y_train <- trainData$Exam_Score
```

```
X_test <- model.matrix(formula(lm_stepwise_BIC),testData)[, -1]
y_test <- testData$Exam_Score
```

Plot below shows the MSE for different values of lambda, selected by glmnet cross-validation:



The best lambda for ridge regression is 0.2175 .

4.7.1 Model Evaluation

Since the glmnet package has different implementation from lm, the R_a^2, RSS and TSS will be computed manually.

There is a significant decrease in R_a^2 when the ridge penalty is applied to the model from BIC. Train and test MSE also worsened but still close to model without penalty. The MSE of the Ridge model is 5.1064.

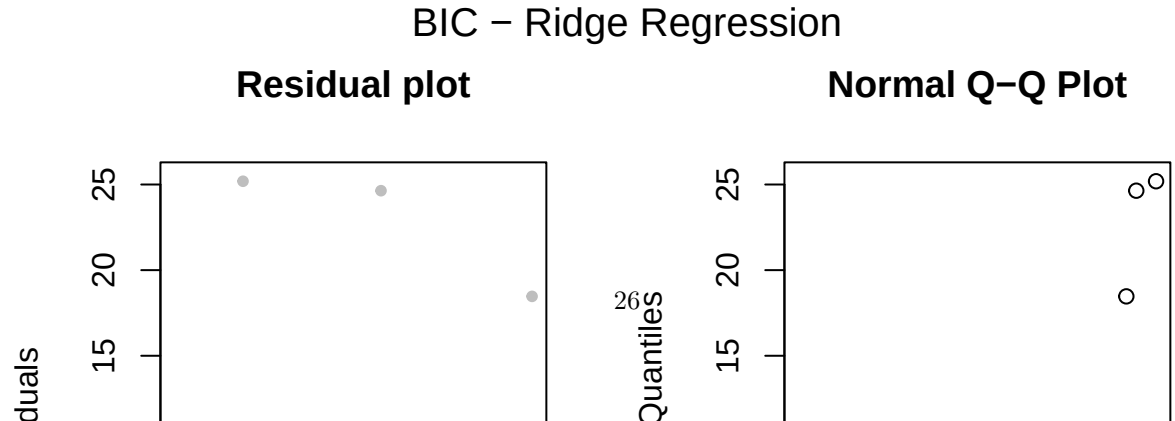
4.7.2 Model Diagnostics

There is no collinearity between the predictors, as already seen in the model diagnostics of the stepwise selection model using BIC.

Let's check if the linearity, equal variance, and normality assumptions hold for the penalized Stepwise - BIC model.

term	Full Model	Stepwise - AIC	Stepwise - AIC - noInf	Stepw
(Intercept)	40.81000476	41.1801855	40.82760906	41.
Hours_Studied	0.29975654	0.2985929	0.29728531	0.
Attendance	0.20085256	0.2004258	0.20124902	0.
Parental_InvolvementLow	-1.74866682	-1.7449619	-1.96461711	-1.
Parental_InvolvementMedium	-0.87813570	-0.8855843	-1.01682775	-0.
Access_to_ResourcesLow	-1.93780776	-1.9315804	-1.94037756	-1.
Access_to_ResourcesMedium	-0.70394108	-0.7264337	-0.97931685	-0.
Extracurricular_ActivitiesYes	0.75721119	0.7623519	0.48994749	0.
Sleep_Hours	0.03441834	NA	NA	
Previous_Scores	0.04230428	0.0419970	0.04966763	0.
Motivation_LevelLow	-1.08951959	-1.1039822	-0.98162163	-1.
Motivation_LevelMedium	-0.46056501	-0.4483638	-0.45055644	-0.
Internet_AccessYes	1.18031911	1.1806072	1.00993760	1.
Tutoring_Sessions	0.45597793	0.4585026	0.49999379	0.
Family_IncomeLow	-1.35181140	-1.3318762	-0.94676274	-1.
Family_IncomeMedium	-0.95057519	-0.9607421	-0.49352771	-0.
Teacher_QualityLow	-0.63778366	-0.6339541	-1.01842868	
Teacher_QualityMedium	-0.36522366	-0.3579991	-0.53947216	
School_TypePublic	0.20742609	NA	NA	
Peer_InfluenceNeutral	0.44013041	0.4356338	0.46978220	
Peer_InfluencePositive	0.66255901	0.6595680	0.95111432	
Physical_Activity	0.31877345	0.3219211	0.23785794	0.
Learning_DisabilitiesYes	-1.19128166	-1.1665347	-0.92846536	-1.
Parental_Education_LevelHigh School	-0.61798432	-0.6142759	-0.51748618	-0.
Parental_Education_LevelPostgraduate	0.55266287	0.5380331	0.46219865	0.
Distance_from_HomeModerate	0.54053191	0.5321244	0.56812624	0.
Distance_from_HomeNear	1.23367651	1.2403961	1.05140517	1.
GenderMale	-0.14095815	NA	NA	

model	adj_R2	train_MSE	test_MSE
Full Model	0.7451	3.4713	4.7566
Stepwise - AIC	0.7455	3.4878	4.7703
Stepwise - AIC - noInf	0.9905	0.0989	4.8330
Stepwise - BIC	0.7413	3.5756	4.9977
Stepwise - BIC - noInf	0.9715	0.2979	3.5756
BIC - Ridge	0.6381	3.6248	5.1064



- The mean of the residuals seem to be 0 but there are a few data points with very high residual values.
- The outer regions in the graph have smaller spread than the spread of residuals in the middle.
- Majority of the data points seem to lie in the theoretical quantile of the Normal distribution, but this still needs to be verified.

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(Ridge_Residuals ~ ., data = as.data.frame(X_train))
## BP = 17.159, df = 20, p-value = 0.6426
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Ridge_Residuals
## W = 0.30009, p-value < 2.2e-16
```

- From the results of the BP test, the p-value is 0.6426 which is higher than $\alpha = 0.05$. Thus, we fail to reject H_0 that the variance of the residuals is constant
- Based on Shapiro-Wilk test, however, the Normality assumption does not hold.
- From the Shapiro-Wilk test, since the p-value is $< 2.2e^{-16}$, at $\alpha = 0.05$, we reject H_0 that the residuals are from the $Normal$ distribution.

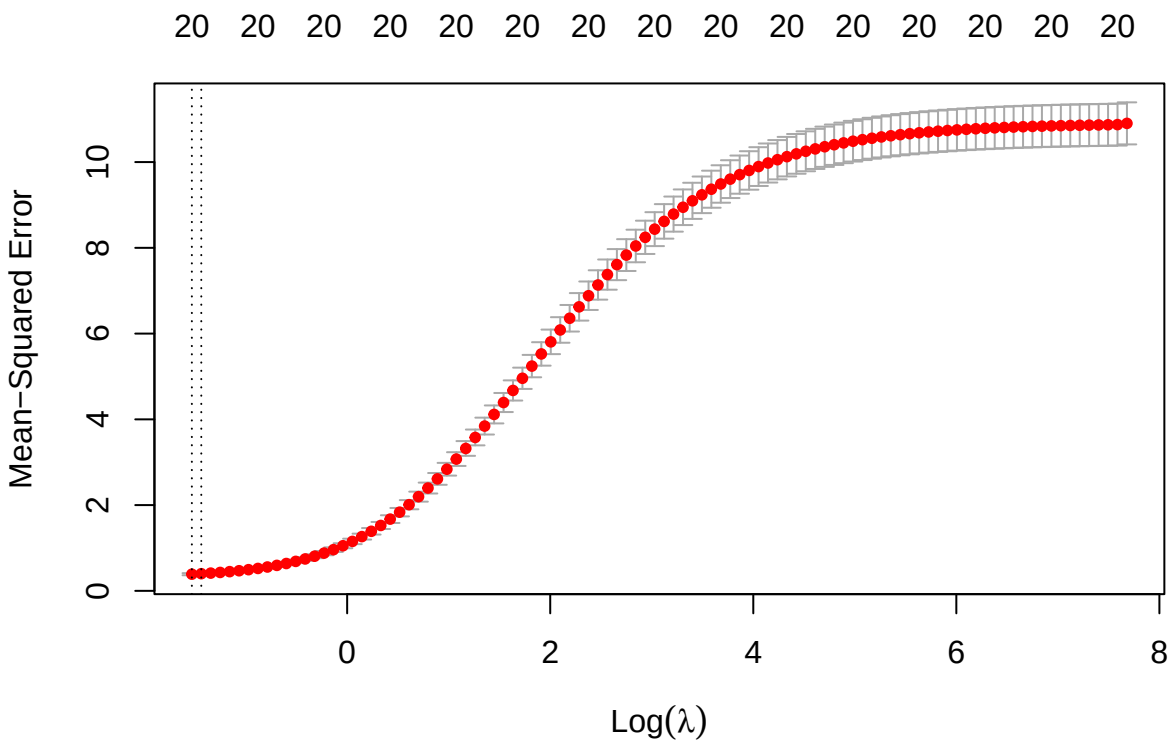
4.8 Ridge Regression with no Influential Points

We will penalize the model that we obtained from BIC using Ridge penalty, since this model is the best model we obtained so far with good model diagnostics, but we will remove influential points too.

```
X_train <- model.matrix(formula(lm_stepwise_BIC),trainData_BIC_noInf)[, -1]
y_train <- trainData_BIC_noInf$Exam_Score

X_test <- model.matrix(formula(lm_stepwise_BIC),testData)[, -1]
y_test <- testData$Exam_Score
```

Plot below shows the MSE for different values of lambda, selected by glmnet cross-validation:




```
bestlam_Ridge = lm_BIC_Ridgecv$lambda.min
```

The best lambda for ridge regression is 0.2166 .

```
gt(tidy_All |>
  dplyr::select(term, estimate, model) |>
  pivot_wider(names_from = model, values_from = estimate))
```

4.8.1 Model Evaluation

Since the glmnet package has different implementation from lm, the R_a^2, RSS and TSS will be computed manually.

The MSE of the Ridge model is 5.2312713.

The penalized model from stepwise selection using BIC, without influential points, achieved lower R_a^2 than the current best model. However, test MSE increased from 3.58 to 5.23.

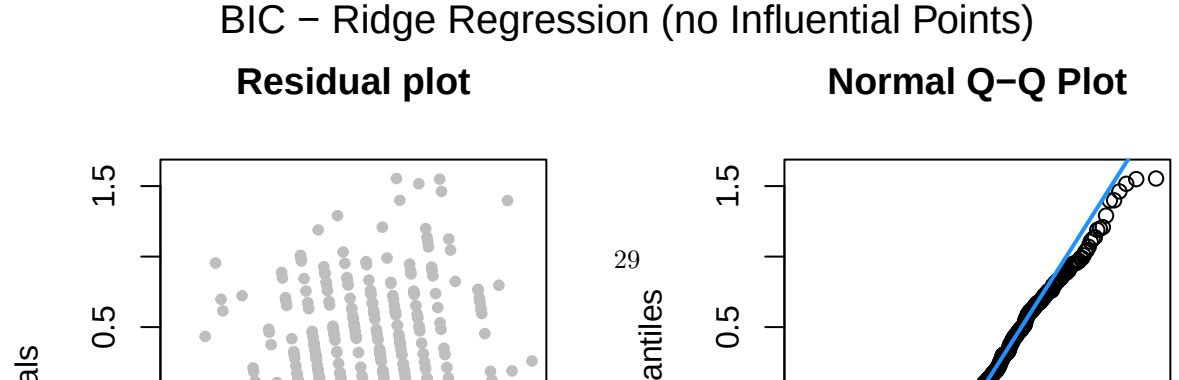
4.8.2 Model Diagnostics

There is no collinearity between the predictors, as already seen in the model diagnostics of the stepwise selection model using BIC.

Let's check if the linearity, equal variance, and normality assumptions hold for the penalized Stepwise - BIC model.

term	Full Model	Stepwise - AIC	Stepwise - AIC - noInf	Stepw
(Intercept)	40.81000476	41.1801855	40.82760906	41.
Hours_Studied	0.29975654	0.2985929	0.29728531	0.
Attendance	0.20085256	0.2004258	0.20124902	0.
Parental_InvolvementLow	-1.74866682	-1.7449619	-1.96461711	-1.
Parental_InvolvementMedium	-0.87813570	-0.8855843	-1.01682775	-0.
Access_to_ResourcesLow	-1.93780776	-1.9315804	-1.94037756	-1.
Access_to_ResourcesMedium	-0.70394108	-0.7264337	-0.97931685	-0.
Extracurricular_ActivitiesYes	0.75721119	0.7623519	0.48994749	0.
Sleep_Hours	0.03441834	NA	NA	
Previous_Scores	0.04230428	0.0419970	0.04966763	0.
Motivation_LevelLow	-1.08951959	-1.1039822	-0.98162163	-1.
Motivation_LevelMedium	-0.46056501	-0.4483638	-0.45055644	-0.
Internet_AccessYes	1.18031911	1.1806072	1.00993760	1.
Tutoring_Sessions	0.45597793	0.4585026	0.49999379	0.
Family_IncomeLow	-1.35181140	-1.3318762	-0.94676274	-1.
Family_IncomeMedium	-0.95057519	-0.9607421	-0.49352771	-0.
Teacher_QualityLow	-0.63778366	-0.6339541	-1.01842868	
Teacher_QualityMedium	-0.36522366	-0.3579991	-0.53947216	
School_TypePublic	0.20742609	NA	NA	
Peer_InfluenceNeutral	0.44013041	0.4356338	0.46978220	
Peer_InfluencePositive	0.66255901	0.6595680	0.95111432	
Physical_Activity	0.31877345	0.3219211	0.23785794	0.
Learning_DisabilitiesYes	-1.19128166	-1.1665347	-0.92846536	-1.
Parental_Education_LevelHigh School	-0.61798432	-0.6142759	-0.51748618	-0.
Parental_Education_LevelPostgraduate	0.55266287	0.5380331	0.46219865	0.
Distance_from_HomeModerate	0.54053191	0.5321244	0.56812624	0.
Distance_from_HomeNear	1.23367651	1.2403961	1.05140517	1.
GenderMale	-0.14095815	NA	NA	

model	adj_R2	train_MSE	test_MSE
Full Model	0.7451	3.4713	4.7566
Stepwise - AIC	0.7455	3.4878	4.7703
Stepwise - AIC - noInf	0.9905	0.0989	4.8330
Stepwise - BIC	0.7413	3.5756	4.9977
Stepwise - BIC - noInf	0.9715	0.2979	3.5756
BIC - Ridge	0.6381	3.6248	5.1064
BIC - Ridge - noInf	0.9637	0.3589	5.2313



model	adj_R2	train_MSE	test_MSE
Full Model	0.7451	3.4713	4.7566
Stepwise - AIC	0.7455	3.4878	4.7703
Stepwise - AIC - noInf	0.9905	0.0989	4.8330
Stepwise - BIC	0.7413	3.5756	4.9977
Stepwise - BIC - noInf	0.9715	0.2979	3.5756
BIC - Ridge	0.6381	3.6248	5.1064
BIC - Ridge - noInf	0.9637	0.3589	5.2313

- After removal of influential points, the mean of the residuals seem to be 0 at any region of the residual plot. Linearity seems to hold,
- The spread of the data points look constant throughout the residual plot.
- When influential points were removed, it is more observable that the empirical distribution of residuals is close to the normal distribution.

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(Ridge_Residuals ~ ., data = as.data.frame(X_train))
## BP = 24.043, df = 20, p-value = 0.2405
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Ridge_Residuals
## W = 0.99581, p-value = 0.2087
```

- Based on BP test, the equal variance assumption holds.
- Based on Shapiro-Wilk test, the Normality assumption already holds.

The influential points were removed to better see the model diagnostics of the current best model. It turns out, the linearity, equal variance and normality assumption still hold on the penalized model.

4.8.3 Test Results

In the model evaluation summary above, we obtained the following results:

Without the influential points, the penalized model achieved R_a^2 of 0.9637. This is lower than the R_a^2 obtained from stepwise selection using BIC, without the influential points. However, since the test MSE worsened when penalty was applied, the model without penalty is preferred.

---

5 Inference

Given the sampling distribution of the coefficients, the following are the bounds of the confidence interval for each coefficient at 95% confidence level.

term	estimate	p.value	CI_lower	CI_upper
(Intercept)	40.97737419	0.000000e+00	40.40536515	41.54938323
Hours_Studied	0.30272059	1.157699e-256	0.29439852	0.31104266
Attendance	0.20306174	5.887897e-304	0.19869484	0.20742863
Parental_InvolvementLow	-2.02120470	8.804987e-103	-2.16293306	-1.87947634
Parental_InvolvementMedium	-0.97500703	4.889220e-49	-1.09059980	-0.85941426
Access_to_ResourcesLow	-1.95101235	9.301946e-100	-2.09109647	-1.81092823
Access_to_ResourcesMedium	-1.03191906	8.123419e-53	-1.14849038	-0.91534774
Extracurricular_ActivitiesYes	0.44096347	8.281406e-17	0.34073128	0.54119567
Previous_Scores	0.04813242	2.729805e-95	0.04455052	0.05171433
Motivation_LevelLow	-0.92437660	2.321849e-32	-1.06650412	-0.78224907
Motivation_LevelMedium	-0.46660792	6.811238e-12	-0.59686927	-0.33634657
Internet_AccessYes	1.05341957	1.511832e-26	0.87087906	1.23596008
Tutoring_Sessions	0.43913345	2.880068e-70	0.39824442	0.48002249
Family_IncomeLow	-0.96738399	4.242564e-35	-1.10891430	-0.82585369
Family_IncomeMedium	-0.49029244	4.075771e-11	-0.63282856	-0.34775632
Physical_Activity	0.24624290	1.837957e-22	0.19910168	0.29338411
Learning_DisabilitiesYes	-1.02939128	2.536903e-22	-1.22719303	-0.83158953
Parental_Education_LevelHigh School	-0.57328413	1.335427e-21	-0.68561205	-0.46095620
Parental_Education_LevelPostgraduate	0.43044516	1.673408e-08	0.28313267	0.57775766
Distance_from_HomeModerate	0.52548309	3.875609e-09	0.35343763	0.69752856
Distance_from_HomeNear	1.02847953	1.620171e-31	0.86778380	1.18917526