統諮課堂作業0321

Group 3

2025-03-26

目錄

I.Variable Definition

Variable	Data Type	Definition
Hours Studied	Numericeric	學生每週花多少小時讀書, 即單位為’ hr/week’
Attendance	Numericeric	學生在課程上的出席率, 即單位為%
Parental Involvement	Factor	家長的對小朋友教育的參與程度, 此處以「順序尺度」呈現, High > Medium > Low
Access to Resources	Factor	學生獲得的教育資源, 此處以「順序尺度」呈現, High > Medium > Low
Extracurricular Activities	Factor	學生是否有參與課外活動, 以Yes, No呈現
Sleep Hours	Numericeric	學生每天晚上睡多少小時, 即單位為’ hr/each night’
Previous Scores	Numericeric	學生前幾次小考的成績
Motivation Level	Factor	學生的學習程度, 此處以「順序尺度」呈現, High > Medium > Low
Internet Access	Factor	學生在家是否有網路可以上網, 以Yes, No呈現
Tutoring Sessions	Numericeric	學生每個月的輔導課程數

Variable	Data Type	Definition
Family Income	Factor	學生的家庭收入水平, 此處以「順序尺度」呈現, High > Medium > Low
Teacher Quality	Factor	老師教學品質, 此處以「順序尺度」呈現, High > Medium > Low
School Type	Factor	學生就讀的學校的類型, 分公立、私立
Peer Influence	Factor	同儕對學生的學業影響, 此處以「順序尺度」呈現, Positive > Neutral > Negative
Physical Activity	Numericeric	學生平均每週的運動時數, 即單位為’ hr/week’
Learning Disabilities	Factor	學生是否有學習障礙
Parental Education Level	Factor	家長的最高教育水平, 此處以「順序尺度」呈現, Postgraduate > College < High School
Distance from Home	Factor	學生從家裡到學校的距離, 此處以「順序尺度」呈現, Far > Moderate > Near
Gender	Factor	學生的生理性別
Exam Score	Numericeric	學生最終的考試成績, 即Y

II.Data Description

stu
20 Variables 6607 Observations

Hours_Studied

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
6607	0	41	0.997	19.98	20	6.748	10	12	16	20	24	28	30

```
lowest :  1  2  3  4  5, highest: 37 38 39 43 44
```

Attendance

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
6607	0	41	0.999	79.98	80	13.33	62	64	70	80	90	96	98

```
lowest :  60  61  62  63  64, highest:  96  97  98  99 100
```

Parental_Involvement

n	missing	distinct
6607	0	3

Value	Low	Medium	High
Frequency	1337	3362	1908
Proportion	0.202	0.509	0.289

Access_to_Resources

n	missing	distinct
6607	0	3

Value	Low	Medium	High
Frequency	1313	3319	1975
Proportion	0.199	0.502	0.299

Extracurricular_Activities

n	missing	distinct
6607	0	2

Value	Yes	No
Frequency	3938	2669
Proportion	0.596	0.404

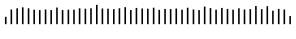
Sleep_Hours

n	missing	distinct	Info	Mean	pMedian	Gmd
6607	0	7	0.96	7.029	7	1.642

Value	4	5	6	7	8	9	10
Frequency	309	695	1376	1741	1399	775	312
Proportion	0.047	0.105	0.208	0.264	0.212	0.117	0.047

For the frequency table, variable is rounded to the nearest 0

Previous_Scores



n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
6607	0	51	1	75.07	75	16.62	53	55	63	75	88	95	97

lowest : 50 51 52 53 54, highest: 96 97 98 99 100

Motivation_Level

n	missing	distinct
6607	0	3

Value	Low	Medium	High
Frequency	1937	3351	1319
Proportion	0.293	0.507	0.200

Internet_Access

n	missing	distinct
6607	0	2

Value	Yes	No
Frequency	6108	499
Proportion	0.924	0.076

Tutoring_Sessions

n	missing	distinct	Info	Mean	pMedian	Gmd
6607	0	9	0.934	1.494	1.5	1.327

Value	0	1	2	3	4	5	6	7	8
Frequency	1513	2179	1649	836	301	103	18	7	1
Proportion	0.229	0.330	0.250	0.127	0.046	0.016	0.003	0.001	0.000

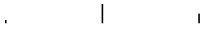
For the frequency table, variable is rounded to the nearest 0

Family_Income

n	missing	distinct
6607	0	3

Value	Low	Medium	High
Frequency	2672	2666	1269
Proportion	0.404	0.404	0.192

Teacher_Quality



n	missing	distinct
6529	78	3

Value	Low	Medium	High
Frequency	657	3925	1947
Proportion	0.101	0.601	0.298

School_Type

n	missing	distinct
6607	0	2

Value	Private	Public
Frequency	2009	4598
Proportion	0.304	0.696

Peer_Influence

n	missing	distinct
6607	0	3

Value	Negative	Neutral	Positive
Frequency	1377	2592	2638
Proportion	0.208	0.392	0.399

Physical_Activity

n	missing	distinct	Info	Mean	pMedian	Gmd
6607	0	7	0.914	2.968	3	1.118

Value	0	1	2	3	4	5	6
Frequency	46	421	1627	2545	1575	361	32
Proportion	0.007	0.064	0.246	0.385	0.238	0.055	0.005

For the frequency table, variable is rounded to the nearest 0

Learning_Disabilities

n	missing	distinct
6607	0	2

Value	Yes	No
Frequency	695	5912
Proportion	0.105	0.895

Parental_Education_Level

n	missing	distinct
6517	90	3

Value	High School	College	Postgraduate
Frequency	3223	1989	1305
Proportion	0.495	0.305	0.200

Distance_from_Home

n	missing	distinct
6540	67	3

Value	Near	Moderate	Far
Frequency	3884	1998	658
Proportion	0.594	0.306	0.101

Gender

n	missing	distinct
6607	0	2

Value	Female	Male
Frequency	2793	3814
Proportion	0.423	0.577

Exam_Score

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
6607	0	45	0.992	67.24	67	4.055	62	63	65	67	69	72	73

lowest : 55 56 57 58 59, highest: 97 98 99 100 101

III.Model

Step1. Linear Model

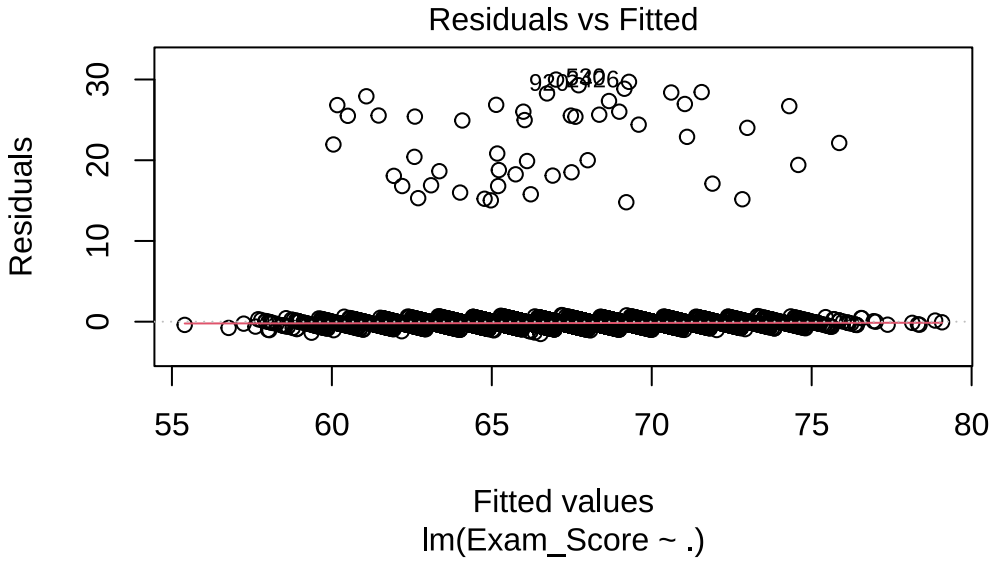
A linear regression model was constructed with Exam Score as the response variable (Y), and the other 19 variables as explanatory variables (X).

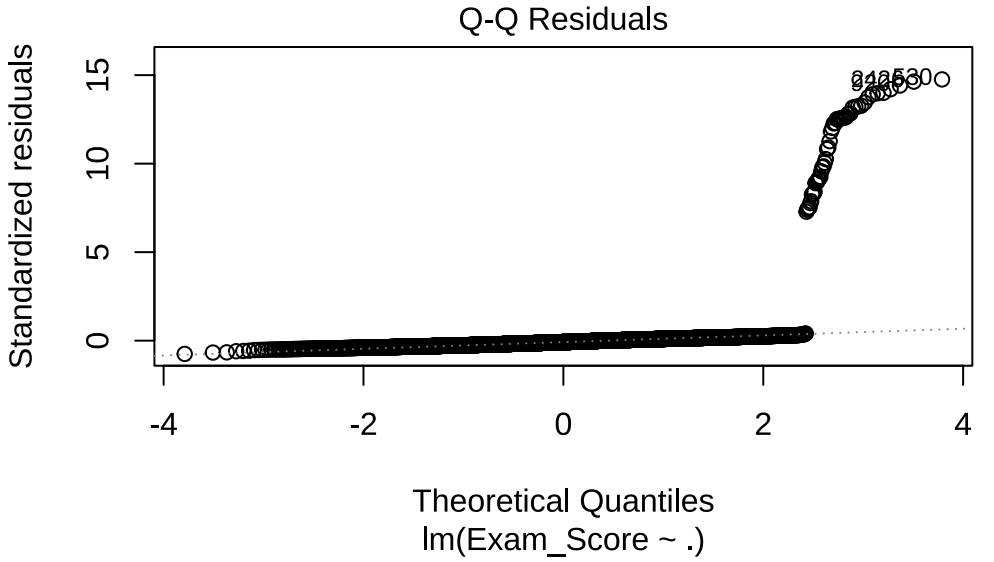
```
Adjusted R-squared: 0.7263

F-statistic: 650.4 on 27 and 6579 DF; p-value: 0

Residual Summary:
```

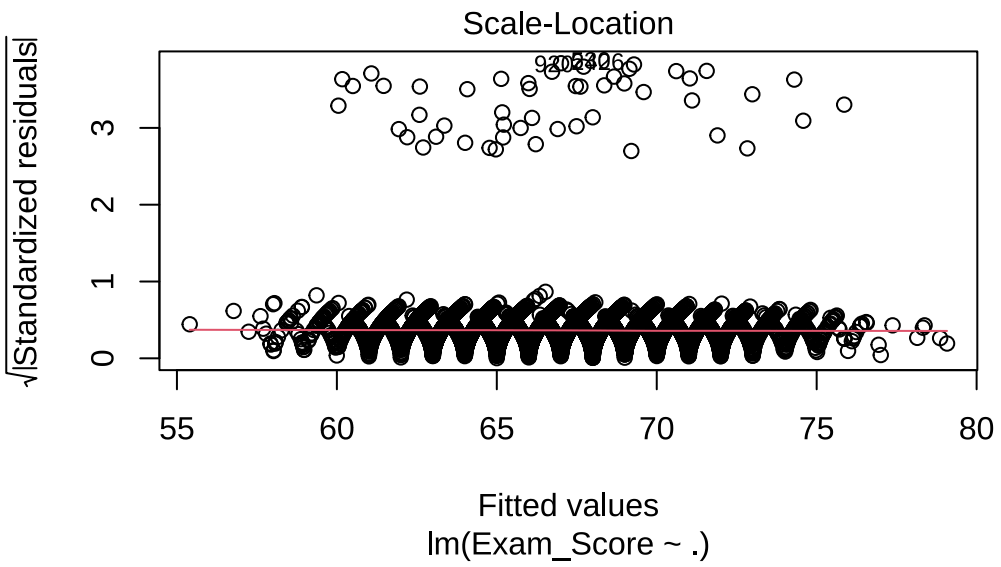
```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.52270 -0.42602 -0.16729  0.00000  0.09091 29.99177
```

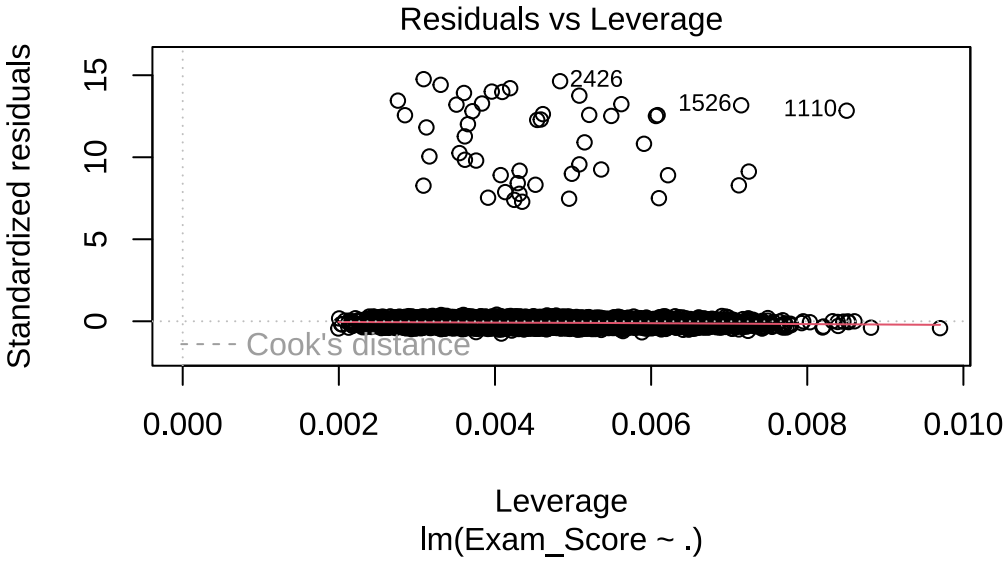
Coefficients (Estimates & p-values):

	Estimate	p_value
(Intercept)	35.619060614	0.000000e+00
Hours_Studied	0.294976449	0.000000e+00
Attendance	0.198849460	0.000000e+00
Parental_InvolvementMedium	0.925566641	3.962569e-44
Parental_InvolvementHigh	1.982722206	6.683184e-155
Access_to_ResourcesMedium	1.049555459	5.065110e-55
Access_to_ResourcesHigh	2.054329747	3.011753e-166
Extracurricular_ActivitiesNo	-0.557811038	1.788403e-27
Sleep_Hours	-0.001466809	9.315710e-01
Previous_Scores	0.048792455	8.444032e-163
Motivation_LevelMedium	0.519638420	5.728821e-19
Motivation_LevelHigh	1.063536536	1.309114e-47
Internet_AccessNo	-0.933385477	1.144466e-22
Tutoring_Sessions	0.497350406	4.036368e-126
Family_IncomeMedium	0.493391481	1.230038e-18
Family_IncomeHigh	1.079871155	1.750982e-53
Teacher_QualityMedium	0.505577436	3.470588e-09
Teacher_QualityHigh	1.051017507	2.661415e-30
School_TypePublic	0.032254745	5.541900e-01
Peer_InfluenceNeutral	0.520105143	2.376748e-14
Peer_InfluencePositive	1.026805726	5.325666e-51
Physical_Activity	0.189109353	9.757056e-15
Learning_DisabilitiesNo	0.856425990	1.718736e-25
Parental_Education_LevelCollege	0.487764100	3.562081e-17
Parental_Education_LevelPostgraduate	0.978247297	3.426240e-48
Distance_from_HomeModerate	-0.516801805	2.776901e-20
Distance_from_HomeFar	-0.905341575	5.825709e-26
GenderMale	-0.040095198	4.294255e-01


Residuals vs Fitted
Residuals
0
10
20
30
55
60
65
70
75
80
Fitted values
lm(Exam_Score ~ .)


Q-Q Residuals
Standardized residuals
0
5
10
15
-4
-2
0
2
4
Theoretical Quantiles
lm(Exam_Score ~ .)





Step2. Variable Selection

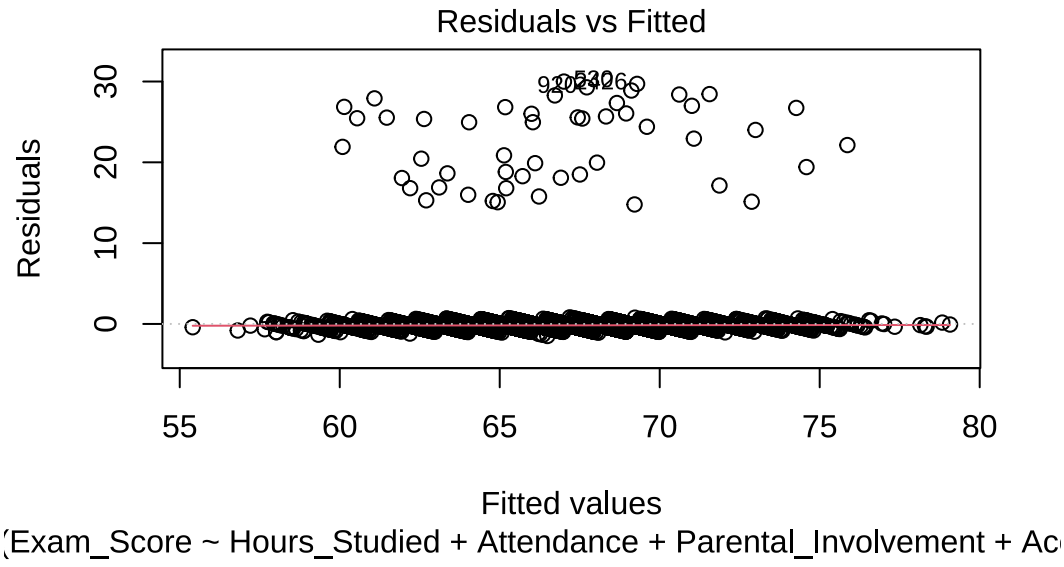
We used the backward selection method for variable selection. The remaining variables are : Hours Studied, Attendance, Parental Involvement, Access to Resources, Extracurricular Activities, Previous Scores, Motivation Level, Internet Access, Tutoring Sessions, Family Income, Teacher Quality, Peer Influence, Physical Activity, Learning Disabilities, Parental Education Level, and Distance from Home.

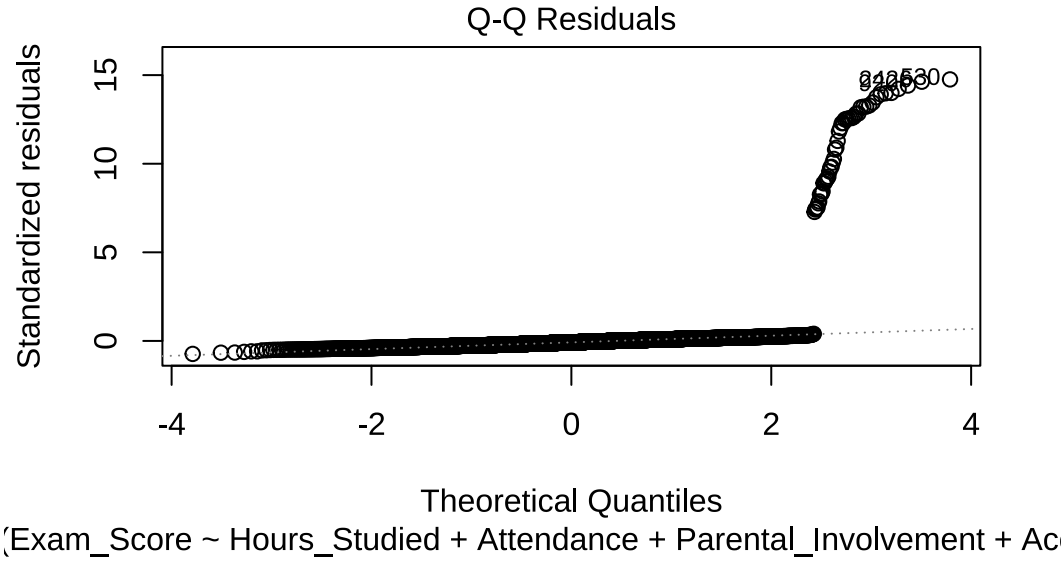
Step3. Linear Model - After variable selection

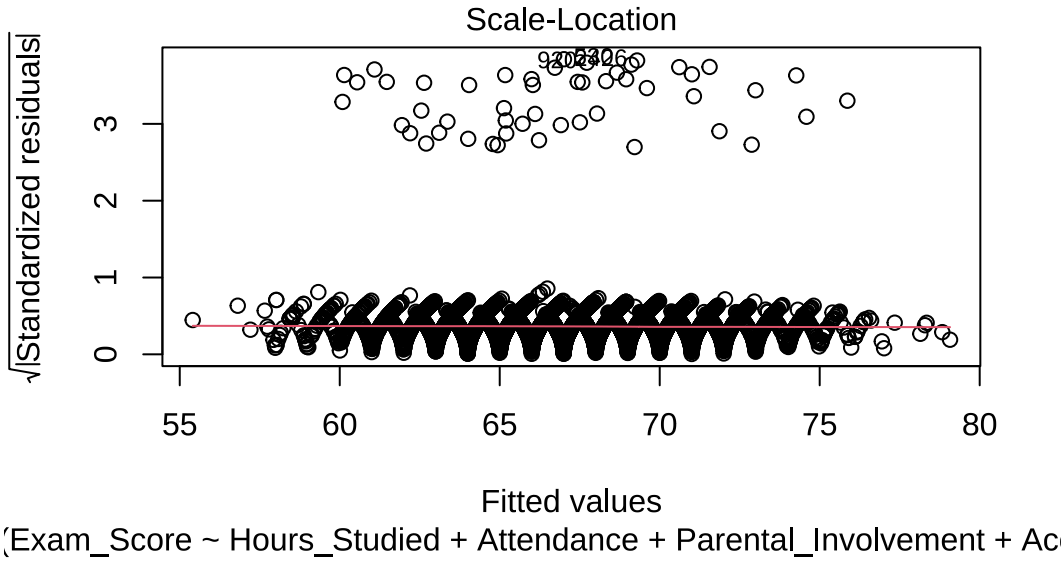
After the variable selection process, we refitted the model which is called "lm_2".

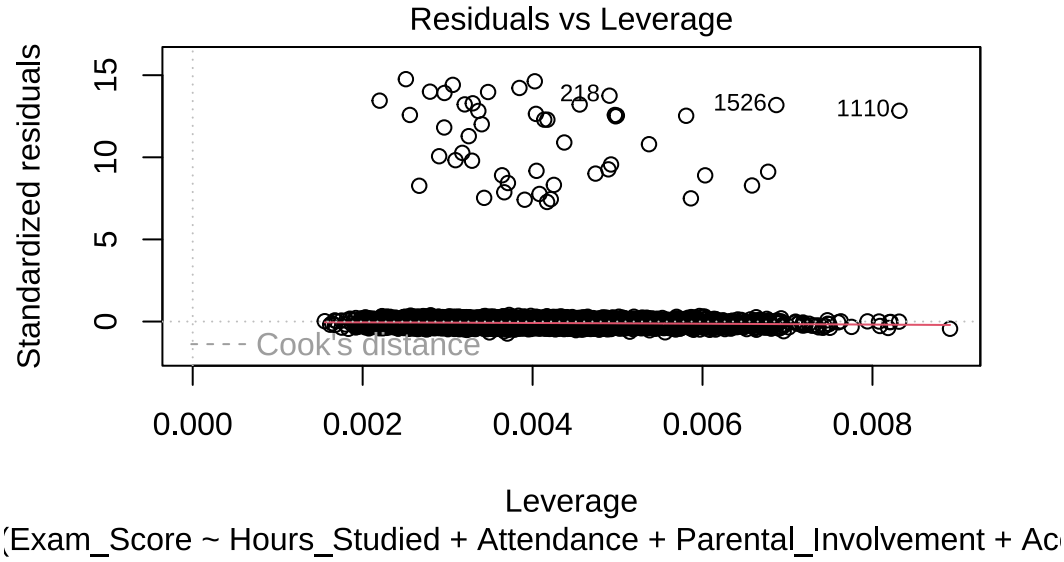
```
  Adjusted R-squared: 0.7264

  F-statistic: 731.89 on 24 and 6582 DF; p-value: 0

  Residual Summary:

    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.48894 -0.42580 -0.16839  0.00000  0.09231 29.99329


  Coefficients (Estimates & p-values):
```

	Estimate	p_value
(Intercept)	35.619060614	0.000000e+00
Hours_Studied	0.294976449	0.000000e+00
Attendance	0.198849460	0.000000e+00
Parental_InvolvementMedium	0.925566641	3.962569e-44
Parental_InvolvementHigh	1.982722206	6.683184e-155
Access_to_ResourcesMedium	1.049555459	5.065110e-55
Access_to_ResourcesHigh	2.054329747	3.011753e-166
Extracurricular_ActivitiesNo	-0.557811038	1.788403e-27
Sleep_Hours	-0.001466809	9.315710e-01
Previous_Scores	0.048792455	8.444032e-163
Motivation_LevelMedium	0.519638420	5.728821e-19
Motivation_LevelHigh	1.063536536	1.309114e-47
Internet_AccessNo	-0.933385477	1.144466e-22
Tutoring_Sessions	0.497350406	4.036368e-126
Family_IncomeMedium	0.493391481	1.230038e-18
Family_IncomeHigh	1.079871155	1.750982e-53
Teacher_QualityMedium	0.505577436	3.470588e-09
Teacher_QualityHigh	1.051017507	2.661415e-30
School_TypePublic	0.032254745	5.541900e-01
Peer_InfluenceNeutral	0.520105143	2.376748e-14
Peer_InfluencePositive	1.026805726	5.325666e-51
Physical_Activity	0.189109353	9.757056e-15
Learning_DisabilitiesNo	0.856425990	1.718736e-25
Parental_Education_LevelCollege	0.487764100	3.562081e-17
Parental_Education_LevelPostgraduate	0.978247297	3.426240e-48
Distance_from_HomeModerate	-0.516801805	2.776901e-20
Distance_from_HomeFar	-0.905341575	5.825709e-26
GenderMale	-0.040095198	4.294255e-01


Residuals vs Fitted
Residuals
0
10
20
30
55
60
65
70
75
80
Fitted values
(Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement + Ac


Q-Q Residuals
Standardized residuals
0
5
10
15
-4
-2
0
2
4
Theoretical Quantiles
(Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement + Ac





Step4. Detect outliers

Although outliers were detected, we decided to retain them to preserve data integrity. Belows are the outliers we detected：

```
 [1]   95  218  405  530  559  561  638  771  837  920 1100 1108 1110 1352 1526
[16] 1608 1845 1864 2077 2293 2422 2426 2514 2543 2596 2688 2905 2955 3125 3142
[31] 3365 3458 3580 3925 3933 4193 4255 4298 4356 4406 4532 4584 4667 4780 5126
[46] 5967 5990 6348 6394 6523
```

Step5-1. Multicollinearity Diagnosis

```
                              GVIF Df GVIF^(1/(2*Df))
Hours_Studied             1.002926  1        1.001462
Attendance                1.004823  1        1.002409
Parental_Involvement      1.007073  2        1.001764
Access_to_Resources       1.008729  2        1.002175
Extracurricular_Activities 1.004070  1        1.002033
Previous_Scores           1.005423  1        1.002708
Motivation_Level          1.007960  2        1.001984
Internet_Access           1.002720  1        1.001359
Tutoring_Sessions         1.001780  1        1.000889
Family_Income             1.007607  2        1.001896
Teacher_Quality           1.007034  2        1.001754
Peer_Influence            1.007599  2        1.001894
Physical_Activity         1.007018  1        1.003503
Learning_Disabilities     1.002500  1        1.001249
Parental_Education_Level  1.007528  2        1.001877
Distance_from_Home        1.004814  2        1.001201
```

The VIF values of most variables are close to 1, which means that their multicollinearity issues are relatively small.

Step5-2. Heteroscedasticity Test

```
    studentized Breusch-Pagan test

data:  lm_2
BP = 15.433, df = 24, p-value = 0.9074
```

Since the p-value = 0.4865 > 0.05, there is no evidence to suggest the presence of heteroscedasticity.

Step5-3. Comparision

```
  Model Adjusted_R_Squared      BIC
1    lm          0.7263446 28366.52
2  lm_2          0.7264284 28341.13
```

IV. Conclusion

Although the explanatory power of the two models (as measured by Adjusted R-squared) is nearly the same, The second model (lm_2), which is the model obtained after variable selection, has a lower BIC value.
This suggests that it achieves a better balance between model complexity and goodness of fit. Therefore, from the perspective of statistical model selection, lm_2 is considered the better model.