Group Homework

Summary report for the Student Performance Factors dataset

林修平、許弘澤、郭育維、黃琮竣、蔡秉杰

2025-03-27

目錄

一、讀取資料 1
二、data information 4
三、data analysis - Backward elimination method 4
四、Data Preprocessing and Stepwise Selection with BIC for Student Performance Analysis 8
五、模型診斷 18
六、Results 20

一、讀取資料

```r
# R Interface to Python
library(reticulate)               # Make R and Python interoperable, allowing R to call Python code.
use_python("C:/Users/user/anaconda3/python.exe", required = TRUE)  # Finding Anaconda's Python path
library(Hmisc)                    # data analysis and report tools
library(ggplot2)                  # a system for creating graphics
library(tableone)                 # a tool for creating tableone
library(dplyr)
library(broom)
library(kableExtra)
# read dataset
Students_data <- read.csv("C:/Users/user/Downloads/StudentPerformanceFactors.csv")
# data description
latex(describe(Students_data), descript = "descriptive statistics", file = '', caption.placement = 'top'
```

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

Access_to_Resources

n	missing	distinct
6607	0	3

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

Extracurricular_Activities

n	missing	distinct
6607	0	2

Value	No	Yes
Frequency	2669	3938
Proportion	0.404	0.596

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	High	Low	Medium
Frequency	1319	1937	3351
Proportion	0.200	0.293	0.507

Internet_Access

n	missing	distinct
6607	0	2

Value	No	Yes
Frequency	499	6108
Proportion	0.076	0.924

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	High	Low	Medium
Frequency	1269	2672	2666
Proportion	0.192	0.404	0.404

Teacher_Quality

n	missing	distinct
6529	78	3

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

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	No	Yes
Frequency	5912	695
Proportion	0.895	0.105

Parental_Education_Level

n	missing	distinct
6517	90	3

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

Distance_from_Home

n	missing	distinct
6540	67	3

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

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

二、data information

1. Hours_Studies：學生每週花多少小時讀書（單位：hr/week）
2. Attendance：學生在課程上的出席率（單位：%）
3. Parental_Involvement：家長對小朋友教育的參與程度（順序尺度：High > Medium > Low）
4. Access_to_Resources：學生獲得的教育資源（順序尺度：High > Medium > Low）
5. Extracurricular_Activities：學生是否有參與課外活動（Yes, No）
6. Sleep_Hours：學生每天晚上睡多少小時（單位：hr/each night）
7. Previous_Scores：學生前幾次小考的成績
8. Motivation_Level：學生的學習動機（順序尺度：High > Medium > Low）
9. Internet_Access：學生在家是否有網路可以上網（Yes, No）
10. Tutoring_Sessions：學生每個月參加的輔導課程數
11. Family_Income：學生的家庭收入水平（順序尺度：High > Medium > Low）
12. Teacher_Quality：老師教學品質（順序尺度：High > Medium > Low）
13. School_Type：學生就讀的學校類型（Public, Private）
14. Peer_Influence：同儕對學生的學業影響（順序尺度：Positive > Neutral > Negative）
15. Physical_Activity：學生每週平均運動時數（單位：hr/week）
16. Learning_Disabilities：學生是否有學習障礙
17. Parental_Education_Level：家長的最高教育程度（順序尺度：Postgraduate > College > High School）
18. Distance_from_Home：學生從家裡到學校的距離（順序尺度：Far > Moderate > Near）
19. Gender：學生的生理性別
20. Exam_Score：學生的最終考試成績（因變數 Y）

三、data analysis - Backward elimination method

```r
# Missing value transfer to NA
Students_data[Students_data == "" | Students_data == " "] <- NA

#
categorical_vars <- c("Parental_Involvement",
                      "Access_to_Resources",
                      "Extracurricular_Activities",
                      "Motivation_Level",
                      "Internet_Access",
                      "Family_Income",
                      "Teacher_Quality",
                      "School_Type",
                      "Peer_Influence",
                      "Learning_Disabilities",
                      "Parental_Education_Level",
                      "Distance_from_Home",
                      "Gender")
```

表 1: Regression for coefficient

term	estimate	std.error	statistic	p.value
(Intercept)	41.829	0.338	123.710	0.000
Hours_Studied	0.295	0.004	68.009	0.000
Attendance	0.199	0.002	88.363	0.000
Parental_InvolvementLow	-1.983	0.075	-26.309	0.000
Parental_InvolvementMedium	-1.062	0.061	-17.557	0.000
Access_to_ResourcesLow	-2.063	0.075	-27.469	0.000
Access_to_ResourcesMedium	-1.010	0.060	-16.807	0.000
Extracurricular_ActivitiesYes	0.562	0.053	10.615	0.000
Sleep_Hours	-0.002	0.018	-0.112	0.911
Previous_Scores	0.049	0.002	27.071	0.000
Motivation_LevelLow	-1.062	0.075	-14.101	0.000
Motivation_LevelMedium	-0.542	0.069	-7.916	0.000
Internet_AccessYes	0.925	0.098	9.442	0.000
Tutoring_Sessions	0.498	0.021	23.699	0.000
Family_IncomeLow	-1.086	0.072	-15.115	0.000
Family_IncomeMedium	-0.591	0.072	-8.210	0.000
Teacher_QualityLow	-1.058	0.094	-11.207	0.000
Teacher_QualityMedium	-0.549	0.058	-9.453	0.000
School_TypePublic	0.033	0.056	0.579	0.563
Peer_InfluenceNeutral	0.522	0.070	7.415	0.000
Peer_InfluencePositive	1.027	0.070	14.653	0.000
Physical_Activity	0.187	0.025	7.385	0.000
Learning_DisabilitiesYes	-0.854	0.085	-10.074	0.000
Parental_Education_LevelHigh School	-0.486	0.060	-8.119	0.000
Parental_Education_LevelPostgraduate	0.503	0.075	6.734	0.000
Distance_from_HomeModerate	0.388	0.095	4.098	0.000
Distance_from_HomeNear	0.908	0.089	10.225	0.000
GenderMale	-0.042	0.053	-0.802	0.422

```r
full_model <- lm(Exam_Score ~ ., data = Students_data)
tidy(full_model) |>
  kbl(digits = 3, caption = "Regression for coefficient") |>
  kable_styling(latex_options = c("scale_down", "striped"))
```

```r
t(glance(full_model)) |>
  kbl(digits = 3, caption = "Full model statistic") |>
  kable_styling()
```

表 2: Full model statistic

r.squared	0.722
adj.r.squared	0.720
sigma	2.069
statistic	609.761
p.value	0.000
df	27.000
logLik	-13674.595
AIC	27407.191
BIC	27603.248
deviance	27194.880
df.residual	6350.000
nobs	6378.000

表 3: Regression for coefficient

term	estimate	std.error	statistic	p.value
(Intercept)	41.816	0.309	135.473	0
Hours_Studied	0.295	0.004	68.033	0
Attendance	0.199	0.002	88.401	0
Parental_InvolvementLow	-1.983	0.075	-26.322	0
Parental_InvolvementMedium	-1.064	0.060	-17.587	0
Access_to_ResourcesLow	-2.062	0.075	-27.470	0
Access_to_ResourcesMedium	-1.009	0.060	-16.805	0
Extracurricular_ActivitiesYes	0.562	0.053	10.615	0
Previous_Scores	0.049	0.002	27.096	0
Motivation_LevelLow	-1.062	0.075	-14.101	0
Motivation_LevelMedium	-0.542	0.068	-7.919	0
Internet_AccessYes	0.924	0.098	9.439	0
Tutoring_Sessions	0.498	0.021	23.713	0
Family_IncomeLow	-1.086	0.072	-15.130	0
Family_IncomeMedium	-0.591	0.072	-8.219	0
Teacher_QualityLow	-1.057	0.094	-11.207	0
Teacher_QualityMedium	-0.549	0.058	-9.454	0
Peer_InfluenceNeutral	0.521	0.070	7.396	0
Peer_InfluencePositive	1.026	0.070	14.651	0
Physical_Activity	0.186	0.025	7.375	0
Learning_DisabilitiesYes	-0.853	0.085	-10.066	0
Parental_Education_LevelHigh School	-0.486	0.060	-8.126	0
Parental_Education_LevelPostgraduate	0.502	0.075	6.723	0
Distance_from_HomeModerate	0.388	0.095	4.094	0
Distance_from_HomeNear	0.907	0.089	10.224	0

If we conduct backward selection (criterion of BIC)

```r
n <- nrow(Students_data)

invisible(capture.output({
  backward_model <- step(full_model, direction = "backward", k = log(n))
}))


tidy(backward_model) |>
  kbl(digits = 3, caption = "Regression for coefficient") |>
  kable_styling(latex_options = c("scale_down", "striped"))
```

```r
t(glance(backward_model)) |>
  kbl(digits = 3, caption = "Full model statistic") |>
  kable_styling()
```

```r
library(lmtest)
bptest(backward_model)
```

```
studentized Breusch-Pagan test
```

data: backward_model BP = 15.576, df = 24, p-value = 0.9028

```r
# Check the number of variables before selection (excluding the intercept)
length(coefficients(full_model)) - 1
```

表 4: Full model statistic

r.squared	0.722
adj.r.squared	0.721
sigma	2.069
statistic	686.156
p.value	0.000
df	24.000
logLik	-13675.096
AIC	27402.193
BIC	27577.969
deviance	27199.153
df.residual	6353.000
nobs	6378.000

```
[1] 27
```

```r
# Check the number of variables after selection
length(coefficients(backward_model)) - 1
```

```
[1] 24
```

```r
# Extract variable names from full_model
full_vars <- attr(terms(full_model), "term.labels")

# Extract variable names from backward_model
backward_model_variable_names <- attr(terms(backward_model), "term.labels")

# Identify removed variables
removed_vars_by_backward_model <- setdiff(full_vars, backward_model_variable_names)

# Display the removed variables
print(removed_vars_by_backward_model)
```

```
[1] "Sleep_Hours" "School_Type" "Gender"
```

四、Data Preprocessing and Stepwise Selection with BIC for Student Performance Analysis

```r
# Stepwise Selection with BIC
# direction = "both": indicates bidirectional selection (both forward and backward)
# k = log(n) is the penalty term for BIC
stepwise_model_BIC <- step(full_model, direction = "both", k = log(n))
```

```
Start:  AIC=9495.5
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

                             Df Sum of Sq   RSS    AIC
- Sleep_Hours                 1         0 27195 9486.7
```

```
- School_Type                 1         1 27196  9487.0
- Gender                      1         3 27198  9487.3
<none>                                    27195  9495.5
- Physical_Activity           1       234 27428  9541.2
- Internet_Access             1       382 27577  9575.6
- Learning_Disabilities       1       435 27630  9587.8
- Extracurricular_Activities  1       483 27677  9598.9
- Distance_from_Home          2       652 27847  9629.0
- Teacher_Quality             2       659 27854  9630.7
- Motivation_Level            2       864 28059  9677.3
- Parental_Education_Level    2       941 28136  9694.9
- Peer_Influence              2       955 28150  9698.0
- Family_Income               2      1011 28206  9710.6
- Tutoring_Sessions           1      2405 29600 10027.2
- Parental_Involvement        2      3059 30254 10157.8
- Previous_Scores             1      3139 30333 10183.3
- Access_to_Resources         2      3284 30479 10205.1
- Hours_Studied               1     19808 47003 12976.6
- Attendance                  1     33439 60634 14600.7

Step:  AIC=9486.71
Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement +
    Access_to_Resources + Extracurricular_Activities + 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

                             Df Sum of Sq   RSS     AIC
- School_Type                 1         1 27196  9478.3
- Gender                      1         3 27198  9478.6
<none>                                    27195  9486.7
+ Sleep_Hours                 1         0 27195  9495.5
- Physical_Activity           1       234 27429  9532.5
- Internet_Access             1       382 27577  9566.8
- Learning_Disabilities       1       435 27630  9579.1
- Extracurricular_Activities  1       483 27678  9590.1
- Distance_from_Home          2       652 27847  9620.2
- Teacher_Quality             2       659 27854  9621.9
- Motivation_Level            2       864 28059  9668.5
- Parental_Education_Level    2       941 28136  9686.1
- Peer_Influence              2       955 28150  9689.4
- Family_Income               2      1011 28206  9702.0
- Tutoring_Sessions           1      2406 29601 10018.6
- Parental_Involvement        2      3059 30254 10149.0
- Previous_Scores             1      3141 30336 10175.1
- Access_to_Resources         2      3285 30480 10196.5
- Hours_Studied               1     19811 47006 12968.2
- Attendance                  1     33455 60650 14593.6

Step:  AIC=9478.25
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 +
    Gender

                             Df Sum of Sq   RSS     AIC
- Gender                      1         3 27199  9470.1
<none>                                    27196  9478.3
+ School_Type                 1         1 27195  9486.7
+ Sleep_Hours                 1         0 27196  9487.0
- Physical_Activity           1       233 27430  9523.9
- Internet_Access             1       382 27579  9558.5
- Learning_Disabilities       1       435 27631  9570.7
- Extracurricular_Activities  1       483 27679  9581.7
- Distance_from_Home          2       652 27848  9611.7
- Teacher_Quality             2       659 27855  9613.3
- Motivation_Level            2       865 28061  9660.3
- Parental_Education_Level    2       940 28136  9677.4
- Peer_Influence              2       956 28152  9681.0
- Family_Income               2      1012 28209  9693.7
- Tutoring_Sessions           1      2407 29603 10010.2
- Parental_Involvement        2      3058 30254 10140.3
- Previous_Scores             1      3143 30340 10167.0
- Access_to_Resources         2      3284 30480 10187.7
- Hours_Studied               1     19811 47008 12959.7
- Attendance                  1     33460 60656 14585.5

Step:  AIC=9470.11
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

                             Df Sum of Sq   RSS     AIC
<none>                                    27199  9470.1
+ Gender                      1         3 27196  9478.3
+ School_Type                 1         1 27198  9478.6
+ Sleep_Hours                 1         0 27199  9478.9
- Physical_Activity           1       233 27432  9515.7
- Internet_Access             1       381 27581  9550.1
- Learning_Disabilities       1       434 27633  9562.2
- Extracurricular_Activities  1       482 27682  9573.4
- Distance_from_Home          2       652 27851  9603.6
- Teacher_Quality             2       659 27858  9605.3
- Motivation_Level            2       863 28063  9651.8
- Parental_Education_Level    2       941 28140  9669.3
- Peer_Influence              2       955 28154  9672.6
- Family_Income               2      1012 28211  9685.6
- Tutoring_Sessions           1      2408 29607 10002.3
- Parental_Involvement        2      3062 30261 10132.9
- Previous_Scores             1      3143 30342 10158.8
- Access_to_Resources         2      3283 30483 10179.4
- Hours_Studied               1     19816 47015 12951.9
- Attendance                  1     33458 60657 14576.8
```

表 5: Stepwise Selection (BIC) - Regression Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	41.816	0.309	135.473	0
Hours_Studied	0.295	0.004	68.033	0
Attendance	0.199	0.002	88.401	0
Parental_InvolvementLow	-1.983	0.075	-26.322	0
Parental_InvolvementMedium	-1.064	0.060	-17.587	0
Access_to_ResourcesLow	-2.062	0.075	-27.470	0
Access_to_ResourcesMedium	-1.009	0.060	-16.805	0
Extracurricular_ActivitiesYes	0.562	0.053	10.615	0
Previous_Scores	0.049	0.002	27.096	0
Motivation_LevelLow	-1.062	0.075	-14.101	0
Motivation_LevelMedium	-0.542	0.068	-7.919	0
Internet_AccessYes	0.924	0.098	9.439	0
Tutoring_Sessions	0.498	0.021	23.713	0
Family_IncomeLow	-1.086	0.072	-15.130	0
Family_IncomeMedium	-0.591	0.072	-8.219	0
Teacher_QualityLow	-1.057	0.094	-11.207	0
Teacher_QualityMedium	-0.549	0.058	-9.454	0
Peer_InfluenceNeutral	0.521	0.070	7.396	0
Peer_InfluencePositive	1.026	0.070	14.651	0
Physical_Activity	0.186	0.025	7.375	0
Learning_DisabilitiesYes	-0.853	0.085	-10.066	0
Parental_Education_LevelHigh School	-0.486	0.060	-8.126	0
Parental_Education_LevelPostgraduate	0.502	0.075	6.723	0
Distance_from_HomeModerate	0.388	0.095	4.094	0
Distance_from_HomeNear	0.907	0.089	10.224	0

```r
# Display regression coefficients after variable selection
tidy(stepwise_model_BIC) |>
  kbl(digits = 3, caption = "Stepwise Selection (BIC) - Regression Coefficients") |>
  kable_styling(latex_options = c("scale_down", "striped"))
```

```r
# Display stepwise_model_BIC statistics
t(glance(stepwise_model_BIC)) |>
  kbl(digits = 3, caption = "Stepwise Selection (BIC) - Model Statistics") |>
  kable_styling()
```

```r
# Check the number of variables before selection (excluding the intercept)
length(coefficients(full_model)) - 1
```

```
[1] 27
```

```r
# Check the number of variables after selection
length(coefficients(stepwise_model_BIC)) - 1
```

```
[1] 24
```

```r
# Extract variable names from full_model
full_vars <- attr(terms(full_model), "term.labels")

# Extract variable names from stepwise_model_BIC
stepwise_vars <- attr(terms(stepwise_model_BIC), "term.labels")
```

表 6: Stepwise Selection (BIC) - Model Statistics

r.squared	0.722
adj.r.squared	0.721
sigma	2.069
statistic	686.156
p.value	0.000
df	24.000
logLik	-13675.096
AIC	27402.193
BIC	27577.969
deviance	27199.153
df.residual	6353.000
nobs	6378.000

```r
# Identify removed variables
removed_vars <- setdiff(full_vars, stepwise_vars)

# Display the removed variables
print(removed_vars)
```

```
[1] "Sleep_Hours" "School_Type" "Gender"
```

```python
import pandas as pd     # Data analysis
from sklearn.preprocessing import LabelEncoder  # Encode categorical data

# read CSV
Students_data = pd.read_csv("C:/Users/user/Downloads/StudentPerformanceFactors.csv")

# Handling missing values
Students_data1 = Students_data.dropna().copy()
print(Students_data1.shape)
```

```
(6378, 20)
```

```python
# Define the variables that need to be encoded
label_mapping = {
    'Parental_Involvement': {'Low': 0, 'Medium': 1, 'High': 2},
    'Access_to_Resources': {'Low': 0, 'Medium': 1, 'High': 2},
    'Motivation_Level': {'Low': 0, 'Medium': 1, 'High': 2},
    'Family_Income': {'Low': 0, 'Medium': 1, 'High': 2},
    'Teacher_Quality': {'Low': 0, 'Medium': 1, 'High': 2},
    'Extracurricular_Activities': {'No': 0, 'Yes': 1},
    'Internet_Access': {'No': 0, 'Yes': 1},
    'Learning_Disabilities': {'No': 0, 'Yes': 1},
    'School_Type': {'Public': 0, 'Private': 1},
    'Gender': {'Female': 0, 'Male': 1},
    'Peer_Influence': {'Negative': 0, 'Neutral': 1, 'Positive': 2},
    'Distance_from_Home': {'Far': 0, 'Moderate': 1, 'Near': 2},
    'Parental_Education_Level': {'High School': 0, 'College': 1, 'Postgraduate': 2}
}

# Label Encoding
for column, mapping in label_mapping.items():
```

```
    # Ensure column names match those in the data
    if column in Students_data1.columns:
        # Convert category values to title case
        Students_data1[column] = Students_data1[column].str.strip().str.title()
        # Apply the mapping
        Students_data1[column] = Students_data1[column].map(mapping)
    else:
        print(f"Column {column} does not exist!")

# Check the data after encoding
print("\nData after encoding:")
```

```
Data after encoding:
```

```
print(Students_data.head())
```

```
   Hours_Studied  Attendance  ...  Gender Exam_Score
0             23          84  ...    Male         67
1             19          64  ...  Female         61
2             24          98  ...    Male         74
3             29          89  ...    Male         71
4             19          92  ...  Female         70

[5 rows x 20 columns]
```

```
# Confirm data types after encoding
print("\nData types after encoding:")
```

```
Data types after encoding:
```

```
print(Students_data1.dtypes)
```

```
Hours_Studied                 int64
Attendance                    int64
Parental_Involvement          int64
Access_to_Resources           int64
Extracurricular_Activities    int64
Sleep_Hours                   int64
Previous_Scores               int64
Motivation_Level              int64
Internet_Access               int64
Tutoring_Sessions             int64
Family_Income                 int64
Teacher_Quality               int64
School_Type                   int64
Peer_Influence                int64
Physical_Activity             int64
Learning_Disabilities         int64
Parental_Education_Level      int64
Distance_from_Home            int64
Gender                        int64
Exam_Score                    int64
dtype: object
```

```python
import numpy as np  # Numerical computing
import pandas as pd  # Data analysis
import statsmodels.api as sm # statistical modeling
from mlxtend.feature_selection import SequentialFeatureSelector as SFS # provides additional tools for M
from sklearn.linear_model import LinearRegression  # Linear Regression Package

# Define X and y
X = Students_data1.drop(columns=['Exam_Score'])  # Features
y = Students_data1['Exam_Score']  # Target

# Define a manual Stepwise Selection function (based on BIC)
# verbose=True, Display detailed information when running
def stepwise_selection_bic(X, y, verbose=True):
    included = []  # Initial list of variables
    best_bic = float('inf')  # Set initial BIC to infinity, In order to ensure that the first comparison

    while True:
        changed = False

        # Forward Selection: Attempt to add variables
        excluded = list(set(X.columns) - set(included)) # Find the variables that have not been selected
        new_bic = pd.Series(index=excluded, dtype=float) # Store BIC value after each unselected variabl
        for new_column in excluded:
            model = sm.OLS(y, sm.add_constant(X[included + [new_column]])).fit()
            new_bic[new_column] = model.bic
        best_new_bic = new_bic.min()  # Find the smallest (best) BIC value among all new models.
        # Find the variable corresponding to the minimum BIC, add it to the selection list and update it
        if best_new_bic < best_bic:
            best_feature = new_bic.idxmin()
            included.append(best_feature)
            best_bic = best_new_bic
            changed = True
            if verbose:
                print(f'Add variable: {best_feature}, BIC = {best_bic:.2f}')

        # Backward Selection: Attempt to remove variables
        # If there are already selected variables, then backward selection is performed.
        if included:
            model = sm.OLS(y, sm.add_constant(X[included])).fit() # Calculate BIC
            current_bic = model.bic
            # Test each selected variable
            for var in included:
                temp_included = included.copy()
                temp_included.remove(var)
                model = sm.OLS(y, sm.add_constant(X[temp_included])).fit()
                # If BIC after removal is less than the current best BIC, remove it from the selection l
                if model.bic < best_bic:
                    best_bic = model.bic
                    worst_feature = var
                    included.remove(worst_feature)
                    changed = True
                    if verbose:
                        print(f'Remove variable: {worst_feature}, BIC = {best_bic:.2f}')
```

```python
        if not changed:
            break

    return included

# Perform manual Stepwise Selection
print("\n=== Manual Implementation of Stepwise Selection (Based on BIC) ===")
```

```
=== Manual Implementation of Stepwise Selection (Based on BIC) ===
```

```python
selected_vars_manual = stepwise_selection_bic(X, y, verbose=True)
```

```
Add variable: Attendance, BIC = 32905.19
Add variable: Hours_Studied, BIC = 30615.37
Add variable: Access_to_Resources, BIC = 30178.22
Add variable: Parental_Involvement, BIC = 29729.30
Add variable: Previous_Scores, BIC = 29237.21
Add variable: Tutoring_Sessions, BIC = 28792.65
Add variable: Family_Income, BIC = 28605.88
Add variable: Peer_Influence, BIC = 28426.78
Add variable: Parental_Education_Level, BIC = 28250.86
Add variable: Motivation_Level, BIC = 28073.32
Add variable: Teacher_Quality, BIC = 27941.95
Add variable: Distance_from_Home, BIC = 27811.65
Add variable: Extracurricular_Activities, BIC = 27712.29
Add variable: Learning_Disabilities, BIC = 27625.25
Add variable: Internet_Access, BIC = 27548.42
Add variable: Physical_Activity, BIC = 27503.17
```

```python
# Display the final selected variables
print("\nFinal selected variables (Manual Implementation):")
```

```
Final selected variables (Manual Implementation):
```

```python
print(selected_vars_manual)
```

```
['Attendance', 'Hours_Studied', 'Access_to_Resources', 'Parental_Involvement', 'Previous_Scores', 'Tutor
```

```python
# Calculate the number of final variables
num_selected_vars_manual = len(selected_vars_manual)
print(f"\nNumber of remaining variables (Manual Implementation): {num_selected_vars_manual}")
```

```
Number of remaining variables (Manual Implementation): 16
```

```python
# Identify removed variables
all_vars = list(X.columns)
removed_vars_manual = list(set(all_vars) - set(selected_vars_manual))
print("\nRemoved variables (Manual Implementation):")
```

```
Removed variables (Manual Implementation):
```

```python
print(removed_vars_manual)
```

```
['Gender', 'School_Type', 'Sleep_Hours']
```

```python
# Fit the final model and display the summary
final_model_manual = sm.OLS(y, sm.add_constant(X[selected_vars_manual])).fit()
print("\nFinal model summary (Manual Implementation):")
```

```
Final model summary (Manual Implementation):
```

```python
print(final_model_manual.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Exam_Score   R-squared:                       0.721
Model:                            OLS   Adj. R-squared:                  0.721
Method:                 Least Squares   F-statistic:                     1030.
Date:                    , 27   2025   Prob (F-statistic):               0.00
Time:                        13:53:28   Log-Likelihood:                -13677.
No. Observations:                6378   AIC:                         2.739e+04
Df Residuals:                    6361   BIC:                         2.750e+04
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         33.9606      0.302    112.388      0.000      33.368      34.553
Attendance                     0.1988      0.002     88.502      0.000       0.194       0.203
Hours_Studied                  0.2948      0.004     68.060      0.000       0.286       0.303
Access_to_Resources            1.0292      0.037     27.721      0.000       0.956       1.102
Parental_Involvement           0.9970      0.037     26.700      0.000       0.924       1.070
Previous_Scores                0.0490      0.002     27.178      0.000       0.045       0.052
Tutoring_Sessions              0.4987      0.021     23.743      0.000       0.458       0.540
Family_Income                  0.5348      0.035     15.337      0.000       0.466       0.603
Peer_Influence                 0.5123      0.034     14.946      0.000       0.445       0.580
Parental_Education_Level       0.4929      0.033     14.837      0.000       0.428       0.558
Motivation_Level               0.5302      0.037     14.229      0.000       0.457       0.603
Teacher_Quality                0.5351      0.043     12.409      0.000       0.451       0.620
Distance_from_Home             0.4761      0.039     12.311      0.000       0.400       0.552
Extracurricular_Activities     0.5582      0.053     10.563      0.000       0.455       0.662
Learning_Disabilities         -0.8503      0.085    -10.044      0.000      -1.016      -0.684
Internet_Access                0.9212      0.098      9.418      0.000       0.729       1.113
Physical_Activity              0.1855      0.025      7.355      0.000       0.136       0.235
==============================================================================
Omnibus:                    10562.577   Durbin-Watson:                   2.010
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5034259.262
Skew:                          11.400   Prob(JB):                         0.00
Kurtosis:                     138.734   Cond. No.                     1.32e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.32e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```python
# use the command below to install package (run it once)
# !pip install mlxtend
```

```python
# Define a custom BIC scoring function
def bic_scorer(estimator, X, y):
    model = sm.OLS(y, sm.add_constant(X)).fit()
    return -model.bic  # Return negative BIC (because SFS maximizes the score)

# Define the regression model
lr = LinearRegression()

# Define Stepwise Selection (based on BIC)
sfs = SFS(lr,
          k_features='best',  # Automatically select the best number of variables
          forward=True,       # Forward selection
          floating=True,      # Allow backward removal (similar to R's "both")
          scoring=bic_scorer,  # Use the custom BIC scoring function
          cv=0)               # Do not use cross-validation

# Fit the data
print("\n=== Perform Stepwise Selection Using mlxtend (Based on BIC) ===")
```

```
=== Perform Stepwise Selection Using mlxtend (Based on BIC) ===
```

```python
sfs = sfs.fit(X, y)

# Display the final selected variables
selected_vars_mlxtend = list(sfs.k_feature_names_)
print("\nFinal selected variables (mlxtend):")
```

```
Final selected variables (mlxtend):
```

```python
print(selected_vars_mlxtend)
```

```
['Hours_Studied', 'Attendance', 'Parental_Involvement', 'Access_to_Resources', 'Extracurricular_Activiti
```

```python
# Calculate the number of final variables
num_selected_vars_mlxtend = len(selected_vars_mlxtend)
print(f"\nNumber of remaining variables (mlxtend): {num_selected_vars_mlxtend}")
```

```
Number of remaining variables (mlxtend): 16
```

```python
# Identify removed variables
removed_vars_mlxtend = list(set(all_vars) - set(selected_vars_mlxtend))
print("\nRemoved variables (mlxtend):")
```

```
Removed variables (mlxtend):
```

```python
print(removed_vars_mlxtend)
```

```
['Gender', 'School_Type', 'Sleep_Hours']
```

```python
# Fit the final model and display the summary
final_model_mlxtend = sm.OLS(y, sm.add_constant(X[selected_vars_mlxtend])).fit()
print("\nFinal model summary (mlxtend):")
```

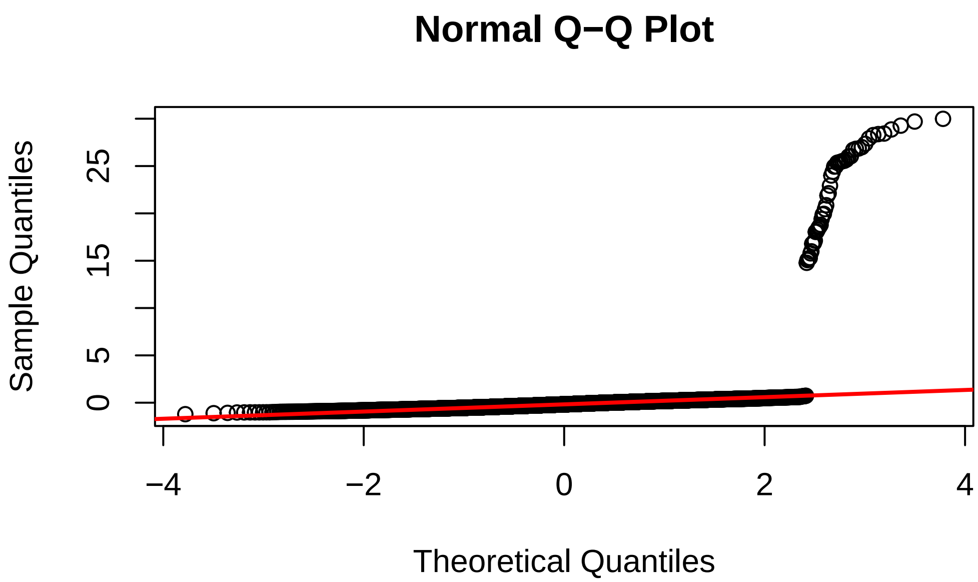
Final model summary (mlxtend):

```
print(final_model_mlxtend.summary())
```

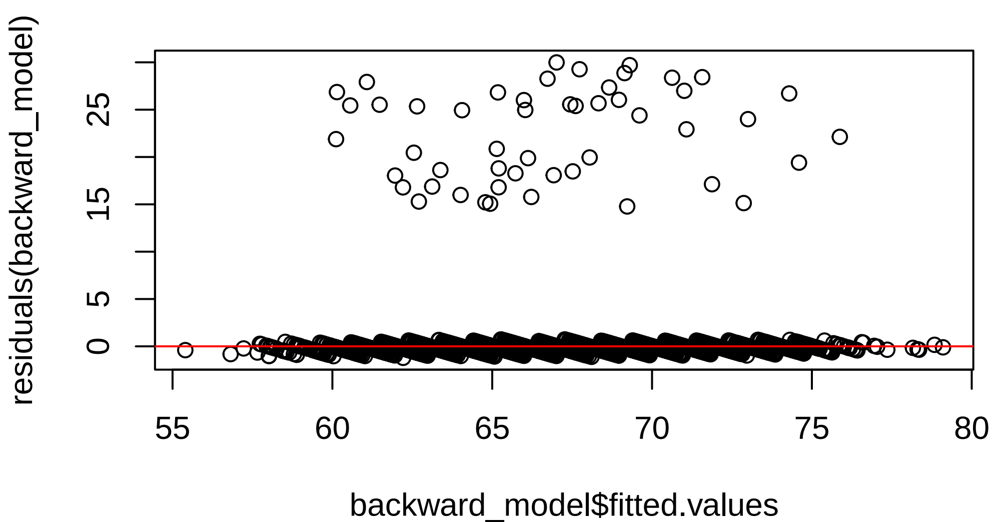
```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Exam_Score   R-squared:                       0.721
Model:                            OLS   Adj. R-squared:                  0.721
Method:                 Least Squares   F-statistic:                     1030.
Date:                   , 27    2025   Prob (F-statistic):              0.00
Time:                        13:53:29   Log-Likelihood:                -13677.
No. Observations:                6378   AIC:                         2.739e+04
Df Residuals:                    6361   BIC:                         2.750e+04
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         33.9606      0.302    112.388      0.000      33.368      34.553
Hours_Studied                  0.2948      0.004     68.060      0.000       0.286       0.303
Attendance                     0.1988      0.002     88.502      0.000       0.194       0.203
Parental_Involvement           0.9970      0.037     26.700      0.000       0.924       1.070
Access_to_Resources            1.0292      0.037     27.721      0.000       0.956       1.102
Extracurricular_Activities     0.5582      0.053     10.563      0.000       0.455       0.662
Previous_Scores                0.0490      0.002     27.178      0.000       0.045       0.052
Motivation_Level               0.5302      0.037     14.229      0.000       0.457       0.603
Internet_Access                0.9212      0.098      9.418      0.000       0.729       1.113
Tutoring_Sessions              0.4987      0.021     23.743      0.000       0.458       0.540
Family_Income                  0.5348      0.035     15.337      0.000       0.466       0.603
Teacher_Quality                0.5351      0.043     12.409      0.000       0.451       0.620
Peer_Influence                 0.5123      0.034     14.946      0.000       0.445       0.580
Physical_Activity              0.1855      0.025      7.355      0.000       0.136       0.235
Learning_Disabilities         -0.8503      0.085    -10.044      0.000      -1.016      -0.684
Parental_Education_Level       0.4929      0.033     14.837      0.000       0.428       0.558
Distance_from_Home             0.4761      0.039     12.311      0.000       0.400       0.552
==============================================================================
Omnibus:                    10562.577   Durbin-Watson:                   2.010
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5034259.262
Skew:                          11.400   Prob(JB):                         0.00
Kurtosis:                     138.734   Cond. No.                     1.32e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.32e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

五、模型診斷

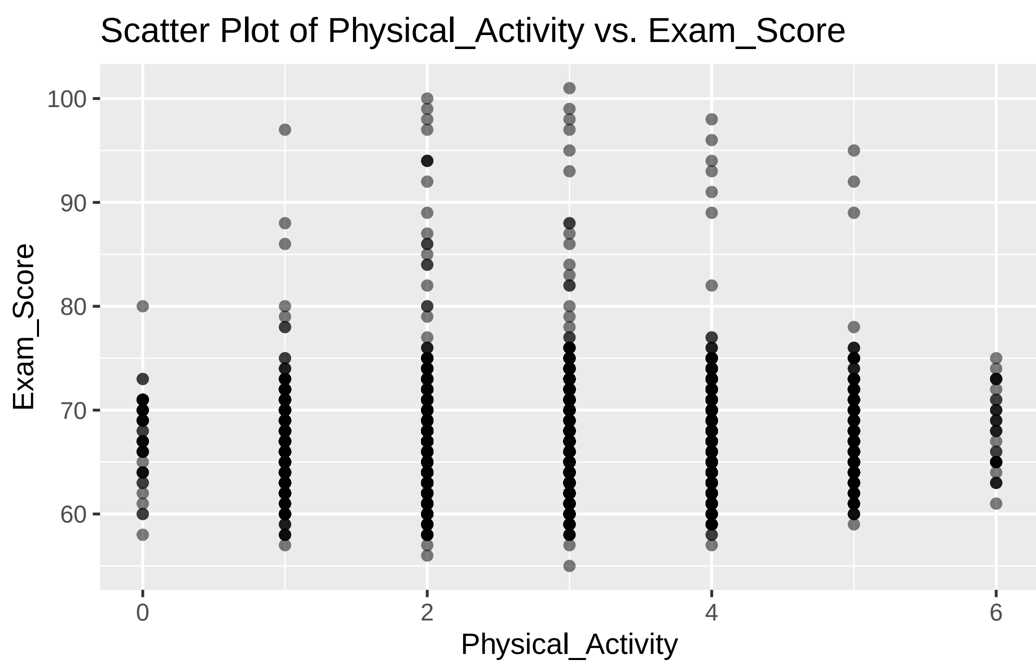
```
#
qqnorm(residuals(backward_model))
qqline(residuals(backward_model), col = "red", lwd = 2)
```




```r
#
plot(backward_model$fitted.values, residuals(backward_model))
abline(h = 0, col = "red")
```




```r
ggplot(Students_data, aes_string(x ="Physical_Activity", y = "Exam_Score")) +
  geom_point(alpha = 0.5) +
  ggtitle("Scatter Plot of Physical_Activity vs. Exam_Score")
```



1. QQ-plot顯示原始資料具有嚴重右偏的趨勢。
2. 用boxcox轉換無法有效解決違反模型假設的問題。
3. 以每週平均運動時間為例，運動時間並無和分數有明顯線性關係，
   雖然逐步回歸有將該解釋變數選入，但根據模型基礎假設，應該將此變數刪除。
4. 在逐步回歸模型的基礎下，考慮所有解釋變數對應變數的散佈圖、盒狀圖，
   將Physical_Activity、Previous_Scores兩個變數手動刪除

六、Results

1. 經由第三部分，Backward Selection Using BIC Strategy，最終刪除3個特徵變數，['School_Type'，'Sleep_Hours'，'Gender']。
2. 經由第四部份，Stepwise Selection Using BIC Strategy，最終也是刪除3個特徵變數，['School_Type'，'Sleep_Hours'，'Gender']。
3. 經由第五部份模型診斷，在逐步回歸模型的基礎下，考慮所有解釋變數對應變數的散佈圖、盒狀圖，將Physical_Activity、Previous_Scores兩個變數手動刪除。
4. 但是還是不符合其模型假設，因此該筆資料或許不適用於傳統回歸模型。