Student performance analysis

Tshepang Mahlatji

2024-02-01

Introduction

This R Project focuses on the analysis of student performance, exploring the influence of various factors on academic outcomes. The project encompasses data loading, exploration, linear regression modeling, and diagnostic checks to gain insights into the determinants of students' academic success.

Variables of Interest

The dataset used in this analysis includes the following variables:

1. **Hours Studied:** The number of hours students spend on studying.
2. **Previous Scores:** Students' performance in previous assessments.
3. **Extracurricular Activities:** Involvement in activities beyond the regular curriculum.
4. **Sleep Hours:** The amount of sleep students get regularly.
5. **Sample Question Papers Practiced:** The number of practice question papers students have worked on.
6. **Performance Index:** The overall performance index representing academic achievements.

Throughout the project, we will explore relationships between these variables and the students' performance index. The analysis includes building linear regression models to understand the impact of different factors and conducting diagnostic checks to ensure the validity of the models.

The subsequent sections will provide a detailed walkthrough of the data loading process, exploration of variables, linear regression modeling, and the diagnostic checks performed to draw meaningful conclusions about the factors influencing student performance.

```
head(data)
```

```
##   Hours.Studied Previous.Scores Extracurricular.Activities Sleep.Hours
## 1             7              99                        Yes           9
## 2             4              82                         No           4
## 3             8              51                        Yes           7
## 4             5              52                        Yes           5
## 5             7              75                         No           8
## 6             3              78                         No           9
##   Sample.Question.Papers.Practiced Performance.Index
## 1                                1                91
## 2                                2                65
## 3                                2                45
## 4                                2                36
## 5                                5                66
## 6                                6                61
```

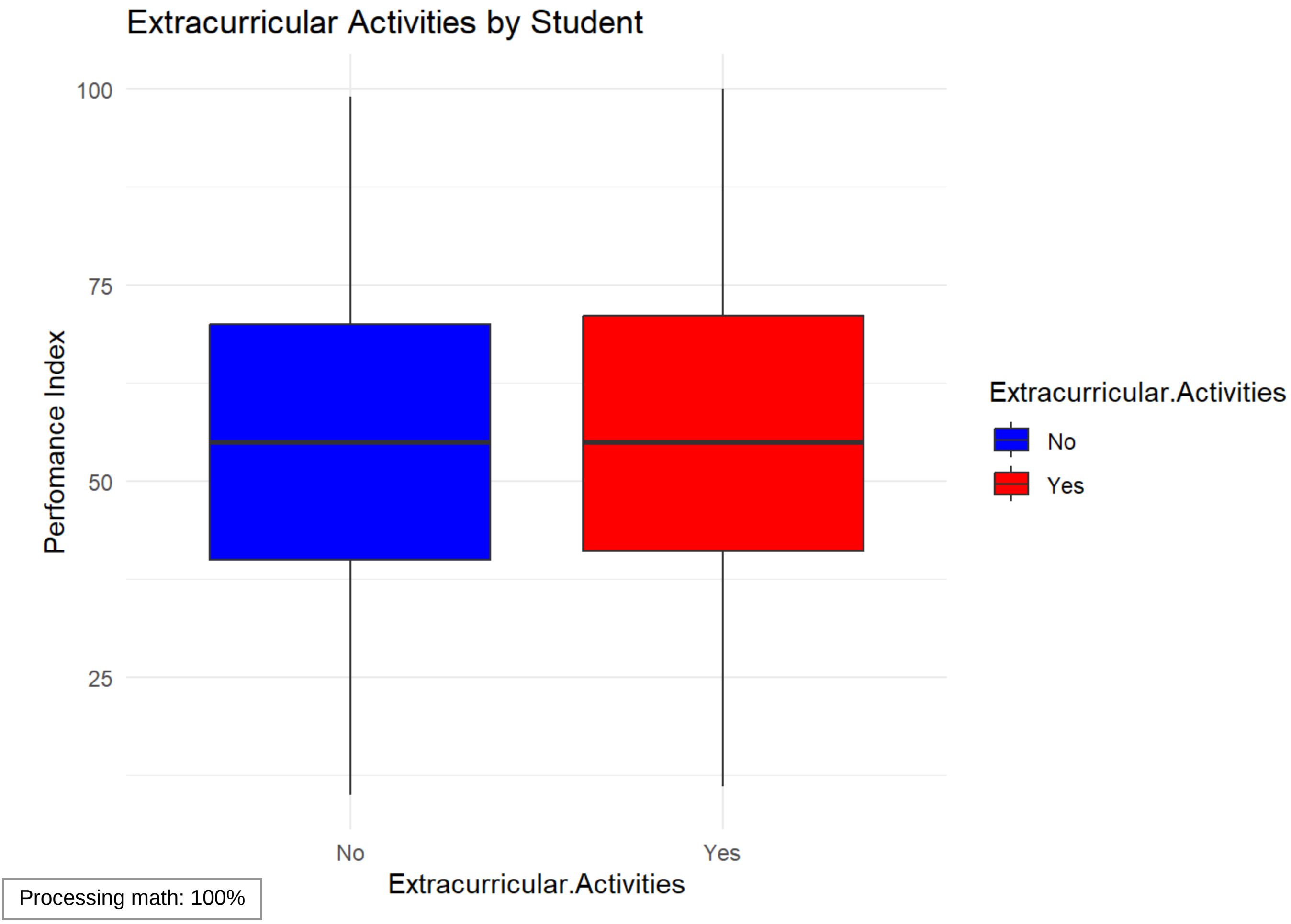
```
colnames(data)
```

```
## [1] "Hours.Studied"                    "Previous.Scores"
## [3] "Extracurricular.Activities"       "Sleep.Hours"
## [5] "Sample.Question.Papers.Practiced" "Performance.Index"
```

```
summary(data)
```

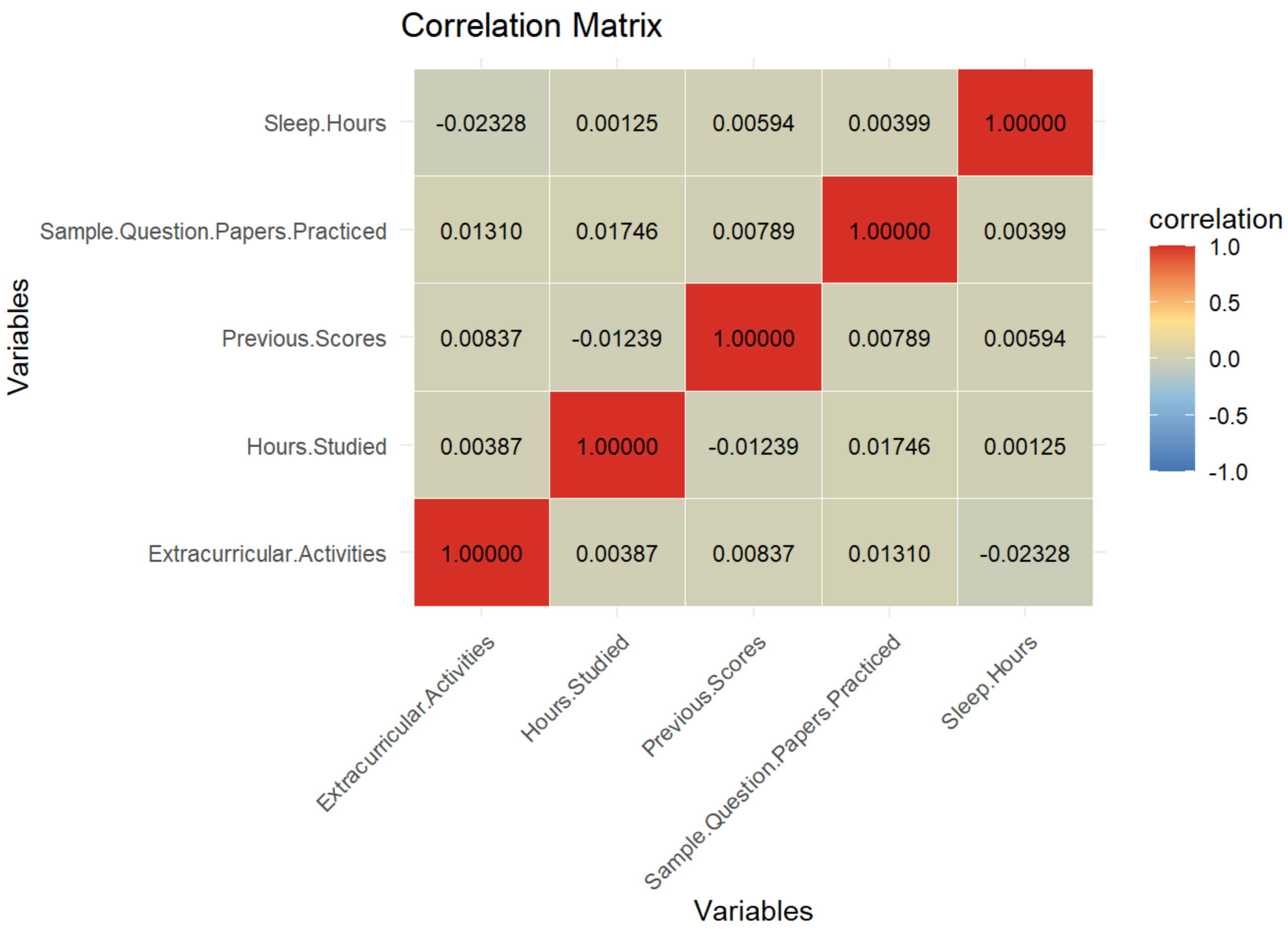
```
##  Hours.Studied   Previous.Scores Extracurricular.Activities  Sleep.Hours
##  Min.   :1.000   Min.   :40.00   Length:10000               Min.   :4.000
##  1st Qu.:3.000   1st Qu.:54.00   Class :character           1st Qu.:5.000
##  Median :5.000   Median :69.00   Mode  :character           Median :7.000
##  Mean   :4.993   Mean   :69.45                              Mean   :6.531
##  3rd Qu.:7.000   3rd Qu.:85.00                              3rd Qu.:8.000
##  Max.   :9.000   Max.   :99.00                              Max.   :9.000
##  Sample.Question.Papers.Practiced Performance.Index
##  Min.   :0.000                    Min.   : 10.00
##  1st Qu.:2.000                    1st Qu.: 40.00
##  Median :5.000                    Median : 55.00
##  Mean   :4.583                    Mean   : 55.22
##  3rd Qu.:7.000                    3rd Qu.: 71.00
##  Max.   :9.000                    Max.   :100.00
```

```
ggplot(data, aes(x = Extracurricular.Activities, y = Performance.Index, fill = Extracurricular.Act
ivities)) +
  geom_boxplot() +
  labs(title = "Extracurricular Activities by Student", y = "Perfomance Index") +
  scale_fill_manual(values = c("blue", "red")) + ### Adding color to the plot
  theme_minimal()
```



Processing math: 100%

```
data$Extracurricular.Activities = ifelse(data$Extracurricular.Activities == 'Yes', 1, 0)
Y = data$Performance.Index
X = subset(data,select = -Performance.Index)


correlation_matrix = cor(X)
correlation_long = correlation_matrix %>%
  as.data.frame() %>%
  rownames_to_column(var = "variable1") %>%
  gather(variable2, correlation, -variable1)
ggplot(correlation_long, aes(x = variable1, y = variable2, fill = correlation, label = sprintf("%.
5f", correlation))) +
  geom_tile(color = "white") +
  geom_text(size = 3, color = "black", show.legend = FALSE) +  # Add text labels
  scale_fill_gradientn(colors = c("#4575b4", "#91bfdb", "#fee08b", "#d73027"),
                       limits = c(-1, 1),
                       guide = "colorbar") +
  theme_minimal() +
  labs(title = "Correlation Matrix", x = "Variables", y = "Variables") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Multicollinearity is a concern in regression analysis when independent variables are highly correlated, potentially leading to unstable coefficient estimates. This analysis focuses on visualizing the correlation matrix of explanatory variables to identify any signs of multicollinearity.From the plot the relationship between independent variable is between -0.02 and 0.03, which indicates a weak relationship between the variables.

```
set.seed(42)
index <- sample(1:nrow(data), 0.8 * nrow(data))
X_train <- X[index, ]
X_test <- X[-index, ]
Y_train <- Y[index]
Y_test <- Y[-index]

model <- lm(Y_train ~ ., data = X_train)

coefficients <- coef(model)[-1]  # Exclude the intercept
coefficients
```

```
##                    Hours.Studied                  Previous.Scores 
##                        2.8519917                        1.0184417 
##         Extracurricular.Activities                      Sleep.Hours 
##                        0.6234627                        0.4783107 
## Sample.Question.Papers.Practiced 
##                        0.1966157
```

```
variable_names <- names(coefficients) # Get variable names
```

```
# Display coefficients with variable names
print(paste("Intercept:", coef(model)[1]))
```

```
## [1] "Intercept: -34.0772245100225"
```

```
cat("Coefficients:\n")
```

```
## Coefficients:
```

```
for (i in seq_along(coefficients)) {
  cat(variable_names[i], ":", coefficients[i], "\n")
}
```

```
## Hours.Studied : 2.851992 
## Previous.Scores : 1.018442 
## Extracurricular.Activities : 0.6234627 
## Sleep.Hours : 0.4783107 
## Sample.Question.Papers.Practiced : 0.1966157
```

```
print(paste("Intercept:", coef(model)[1]))
```

```
## [1] "Intercept: -34.0772245100225"
```

```
Y_pred <- predict(model, newdata = X_test)
r_squared <- summary(model)$r.squared
cat("R-squared:", round(r_squared, 5), "\n")
```

```
## R-squared: 0.98876
```

R^2 is 0.98876, which implies that the 98.87% of performance index in final grade is explained by 5 independent variables

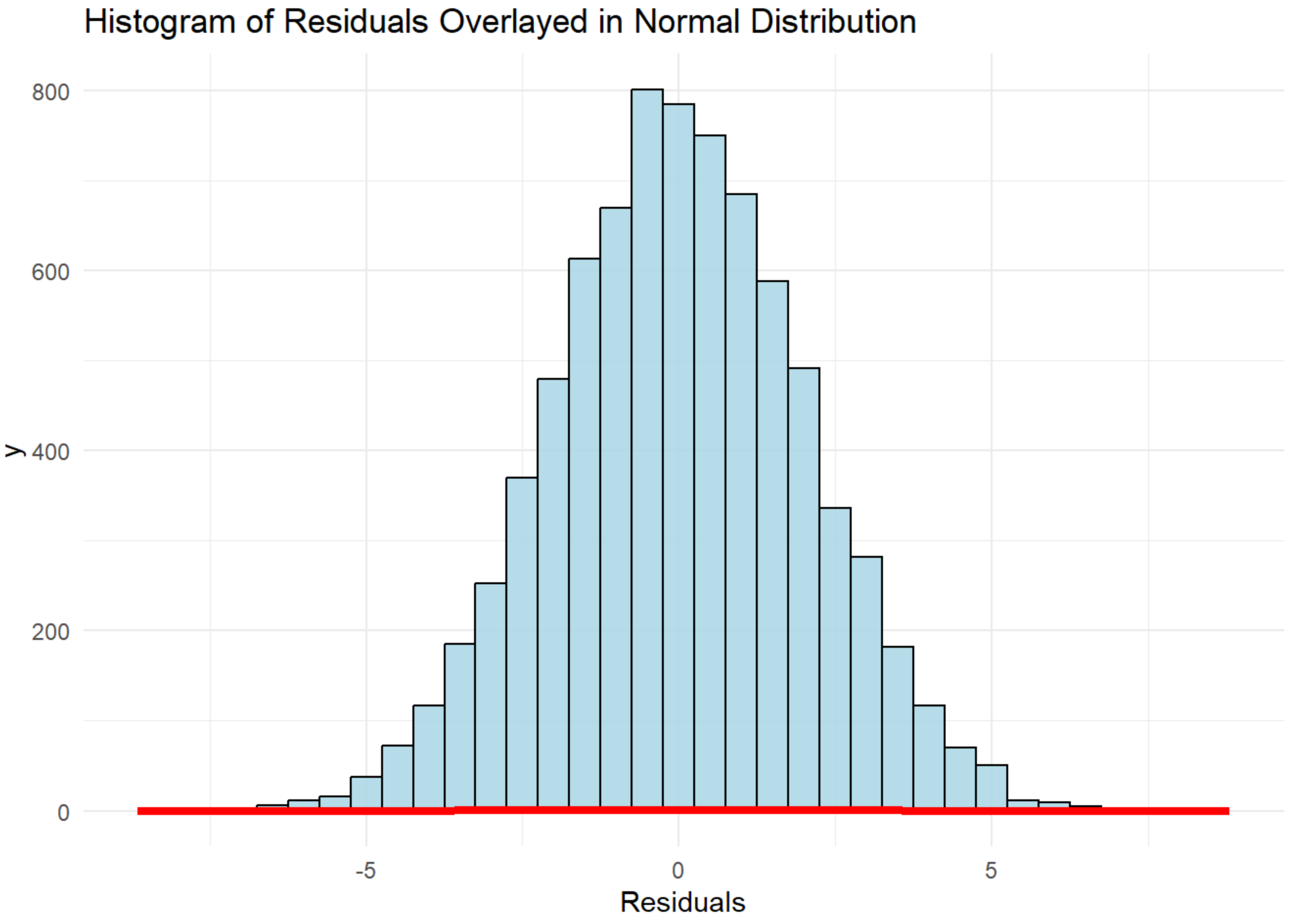
```
summary(model)
```

```
##
## Call:
## lm(formula = Y_train ~ ., data = X_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6457 -1.3716 -0.0379  1.3680  8.7976
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -34.077225   0.142017 -239.95   <2e-16 ***
## Hours.Studied                      2.851992   0.008818  323.44   <2e-16 ***
## Previous.Scores                    1.018442   0.001317  773.50   <2e-16 ***
## Extracurricular.Activities         0.623463   0.045727   13.63   <2e-16 ***
## Sleep.Hours                        0.478311   0.013450   35.56   <2e-16 ***
## Sample.Question.Papers.Practiced   0.196616   0.007969   24.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.044 on 7994 degrees of freedom
## Multiple R-squared:  0.9888, Adjusted R-squared:  0.9888
## F-statistic: 1.407e+05 on 5 and 7994 DF,  p-value: < 2.2e-16
```

The p-values associated with the explanatory variables in our regression analysis play a crucial role in assessing their statistical significance. A p-value less than 0.05 is commonly used as a threshold for significance. In our analysis, all p-values for the explanatory variables are less than 0.05, indicating that they are statistically significant in explaining the performance index.

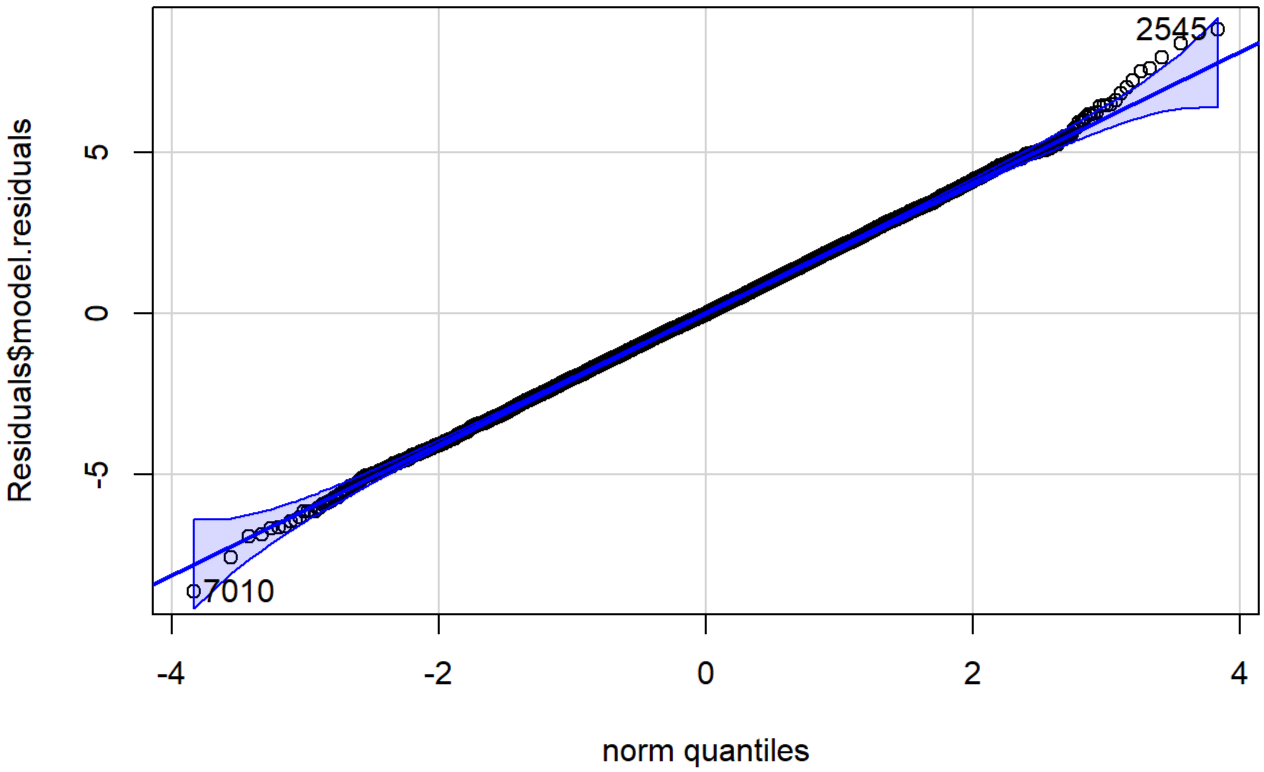
```
mean_residuals = mean(model$residuals, na.rm=FALSE)
sd_residuals = sd(model$residuals)

ggplot(model, aes(x = model$residuals)) +
  geom_histogram(binwidth = 0.5, fill = "lightblue", color = "black", alpha = 0.9) +
  stat_function(fun = dnorm, args = list(mean = mean_residuals, sd = sd_residuals),
                color = "red", linewidth = 1.5)+
  labs(title = "Histogram of Residuals Overlayed in Normal Distribution", x = "Residuals") +
  theme_minimal()+
  xlim(range(model$residuals))
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



```
Residuals = data.frame(model$residuals)
mu = mean(Residuals$model.residuals)
sigma = sd(Residuals$model.residuals)
```

```
qqPlot(Residuals$model.residuals)
```



```
## [1] 2545 7010
```

The alignment of the QQ plot of residuals with the 45-degree line provides evidence supporting the assumption of normality. This strengthens the validity of our linear regression analysis, suggesting that the model is an appropriate representation of the relationship between the variables

```
result = ks.test(Residuals$model.residuals, "pnorm", mean = mu, sd = sigma)
```

```
## Warning in ks.test.default(Residuals$model.residuals, "pnorm", mean = mu, :
## ties should not be present for the Kolmogorov-Smirnov test
```

```
cat("Kolmogorov-Smirnov Test Statistic:", result$statistic, "\n")
```

```
## Kolmogorov-Smirnov Test Statistic: 0.007589643
```

```
cat("P-value:", result$p.value, "\n")
```

```
## P-value: 0.746113
```

```r
alpha <- 0.05
if (result$p.value < alpha) {
  cat("Reject the null hypothesis: The residuals are not normally distributed.\n")
} else {
  cat("Fail to reject the null hypothesis: The residuals appear to be normally distributed.\n")
}
```

```
## Fail to reject the null hypothesis: The residuals appear to be normally distributed.
```

The Kolmogorov-Smirnov test is commonly used to assess the normality of residuals. The null hypothesis of this test is that the data follows a normal distribution. In our analysis, the p-value from the Kolmogorov-Smirnov test is 0.746113. With a p-value greater than 0.05, we fail to reject the null hypothesis, suggesting that our residuals likely follow a normal distribution.