OLS Model Portuguese

Gabe Vasquez

2025-09-12

Contents

Summary **1**

Distribution of Response Variables **2**

Fit OLS Model **5**

Assumptions Check **8**
Normality check 8
Constant Variance Check 10

Fit an OLS Model without Zero GPA Students **23**
Fit OLS Model 2 25
Assumptions Check 28
Constant Variance Check 30

Summary

We fitted an ordinary least squares (OLS) regression model to the data using a subset of predictors selected based on Akaike Information Criterion (AIC) minimization, aiming to model students' academic performance as measured by GPA. However, diagnostic checks indicate that two key assumptions of the classical linear regression model—normality of residuals and homoscedasticity—are violated.

The assumption of normality fails because the distribution of the response variable, GPA, is notably left-skewed and exhibits a slightly bimodal structure. This skewness is primarily driven by a cluster of students with GPA values equal to 0 (n = 16), which act as extreme outliers and create a long left tail. Since OLS estimates are sensitive to non-normality and outliers, especially when inference (e.g., hypothesis testing and confidence intervals) is the goal, this raises concerns about the validity of statistical inference under the model.

Moreover, the ***Residuals vs. Fitted*** values plot reveals a clear pattern of heteroscedasticity—that is, the variance of the residuals is not constant across fitted values. This violates the assumption of constant variance (homoscedasticity), which is critical for the unbiased estimation of standard errors in OLS. Heteroscedasticity is mathematically characterized by $Var(\epsilon_i|x_i) \neq \sigma^2$ where ϵ_i is the residual for observation 'i', and x_i is the vector of predictors.

Furthermore, when evaluating categorical predictors, we observed significant differences in residual variance across groups. For example, Levene's Test indicated statistically significant heterogeneity in variances across categories of some predictors, reinforcing the presence of group-specific heteroscedasticity. Together, these findings suggest that the OLS model does not adequately meet its underlying assumptions and that alternative modeling approaches—such as quantile regression, which is more robust to skewed distributions and heteroscedasticity—may be more appropriate for this data.

NOTE I ran another OLS model without the students with 0 GPA values. The model definitely improved. The constant variance almost seemed to pass. However, the ***Residuals vs. Fitted*** plot still showed some insignificant variance and one of the predictors shows very noticeable inconsistent variance. One of the other predictors passed the ***LeveneTest()***. but it's a categorical variable has two imbalanced levels where one level has much more students than the other.

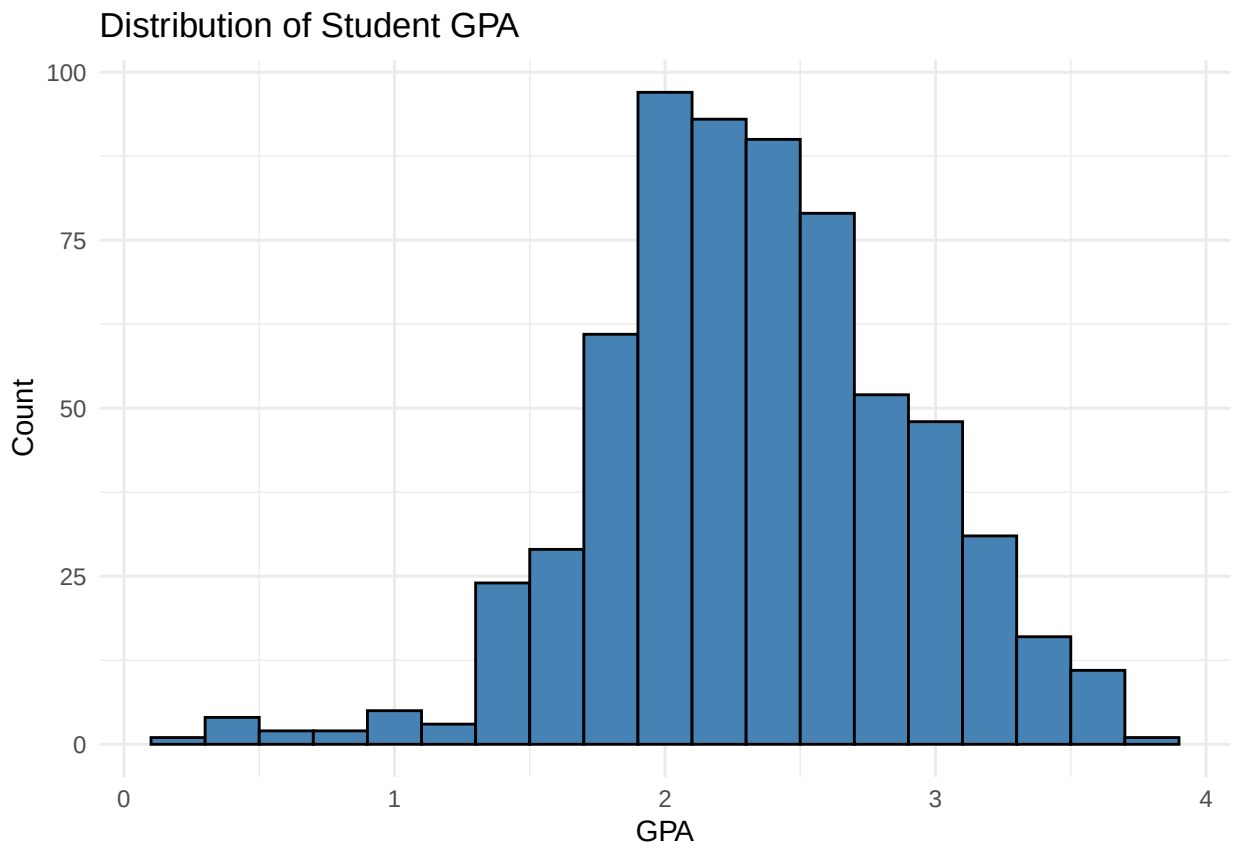
```
# Read in dataset
df <- read.csv("student_data.csv")

# factor categorical variables
df <- df %>%
  mutate(across(where(is.character), as.factor)) %>%
  dplyr::select(-G1, -G2, -G3)

# factor 'failures'
df$failures <- as.factor(df$failures)

#str(df)
names(df)
```

```
##  [1] "school"     "sex"        "age"        "address"    "famsize"
##  [6] "Pstatus"    "Medu"       "Fedu"       "Mjob"       "Fjob"
## [11] "reason"     "guardian"   "traveltime" "studytime"  "failures"
## [16] "schoolsup"  "famsup"     "paid"       "activities" "nursery"
## [21] "higher"     "internet"   "romantic"   "famrel"     "freetime"
## [26] "goout"      "Dalc"       "Walc"       "health"     "absences"
## [31] "GPA"
```

Distribution of Response Variables

Student GPA's Distribution

```
# Distribution of the response variable GPA
ggplot(df, aes(x = GPA)) +
  geom_histogram(binwidth = 0.2, fill = "steelblue", color = "black") +
  labs(title = "Distribution of Student GPA", x = "GPA", y = "Count") +
  theme_minimal()
```

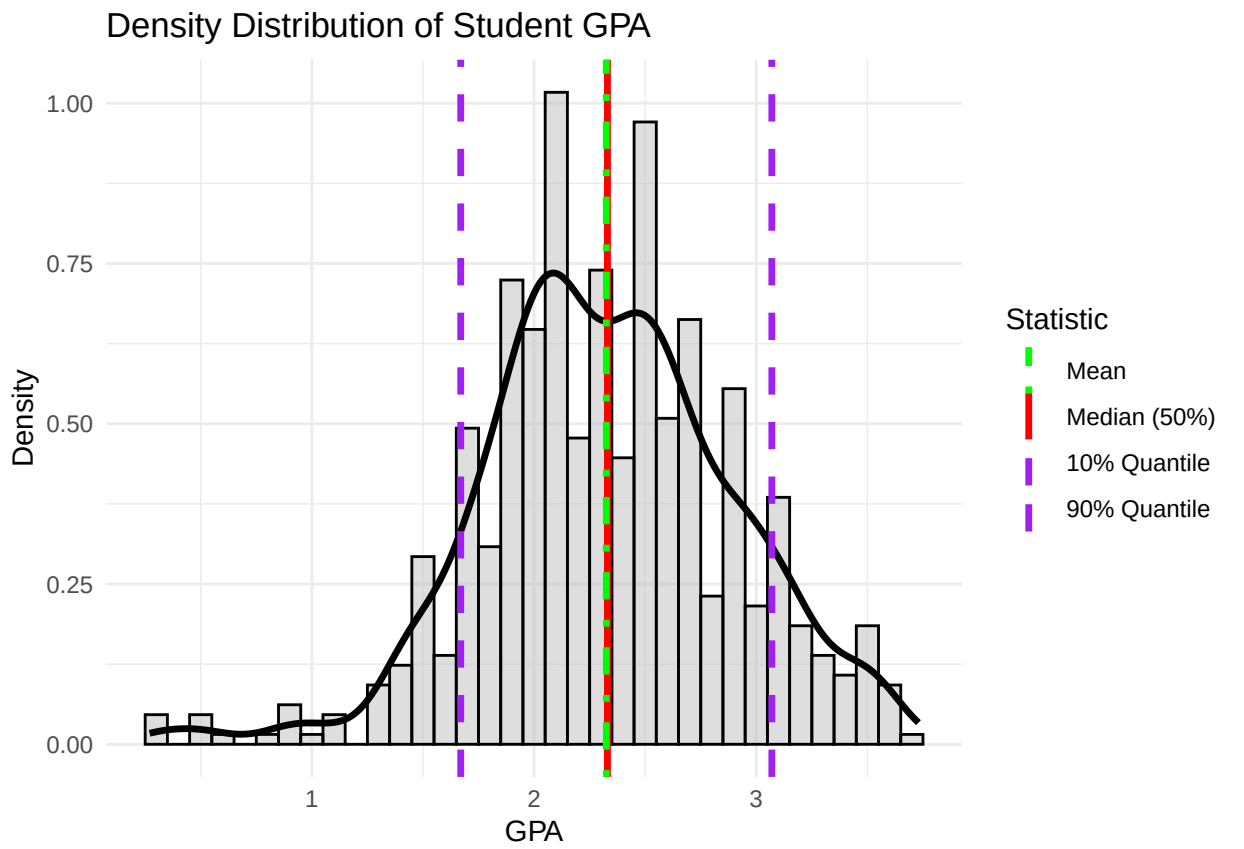



```r
# Compute quantiles and mean
gpa_quantiles <- quantile(df$GPA, probs = c(0.1, 0.5, 0.9))
gpa_mean <- mean(df$GPA)

# Create a data frame for vertical lines
vline_data <- data.frame(
  xintercept = c(
    gpa_quantiles["10%"], gpa_quantiles["50%"], gpa_quantiles["90%"], gpa_mean
  ),
  Label = factor(c(
    "10% Quantile", "Median (50%)", "90% Quantile", "Mean"
  ),
  levels = c("Mean", "Median (50%)", "10% Quantile", "25% Quantile", "75% Quantile", "90% Quantile"))
)

# Plot
ggplot(df, aes(x = GPA)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.1, fill = "gray", color = "black", alpha = 0.5) +
  geom_density(color = "black", size = 1.2) +

  # Add vlines with mapped aesthetics for legend
  geom_vline(data = vline_data, aes(xintercept = xintercept, color = Label, linetype = Label), size = 1

  scale_color_manual(values = c(
    "Mean" = "green",
    "Median (50%)" = "red",
```

```
    "10% Quantile" = "purple",
    "90% Quantile" = "purple"
  )) +

  scale_linetype_manual(values = c(
    "Mean" = "dotdash",
    "Median (50%)" = "solid",
    "10% Quantile" = "dashed",
    "90% Quantile" = "dashed"
  )) +

  labs(
    title = "Density Distribution of Student GPA",
    x = "GPA",
    y = "Density",
    color = "Statistic",
    linetype = "Statistic"
  ) +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
round(mean(df$GPA), 3)
```

```
## [1] 2.325
```

```
round(median(df$GPA), 3)
```

```
## [1] 2.33
```

GPA mean $<$ median $\rightarrow$ GPA distribution follows more of a left-skewed distribution. But, they're so close that we could almost say this distribution is normal. Later, a Shaprio-Wilk's test will tell if the OLS model's distribution is normal or not.

Fit OLS Model

```
# Initialize models
null_model <- lm(GPA ~ 1, data = df)
full_model <- lm(GPA ~ ., data = df)
```

```
# Variable selection
stepAIC_model <- stepAIC(null_model, scope = list(lower = null_model, upper = full_model), direction =
summary(stepAIC_model)
```

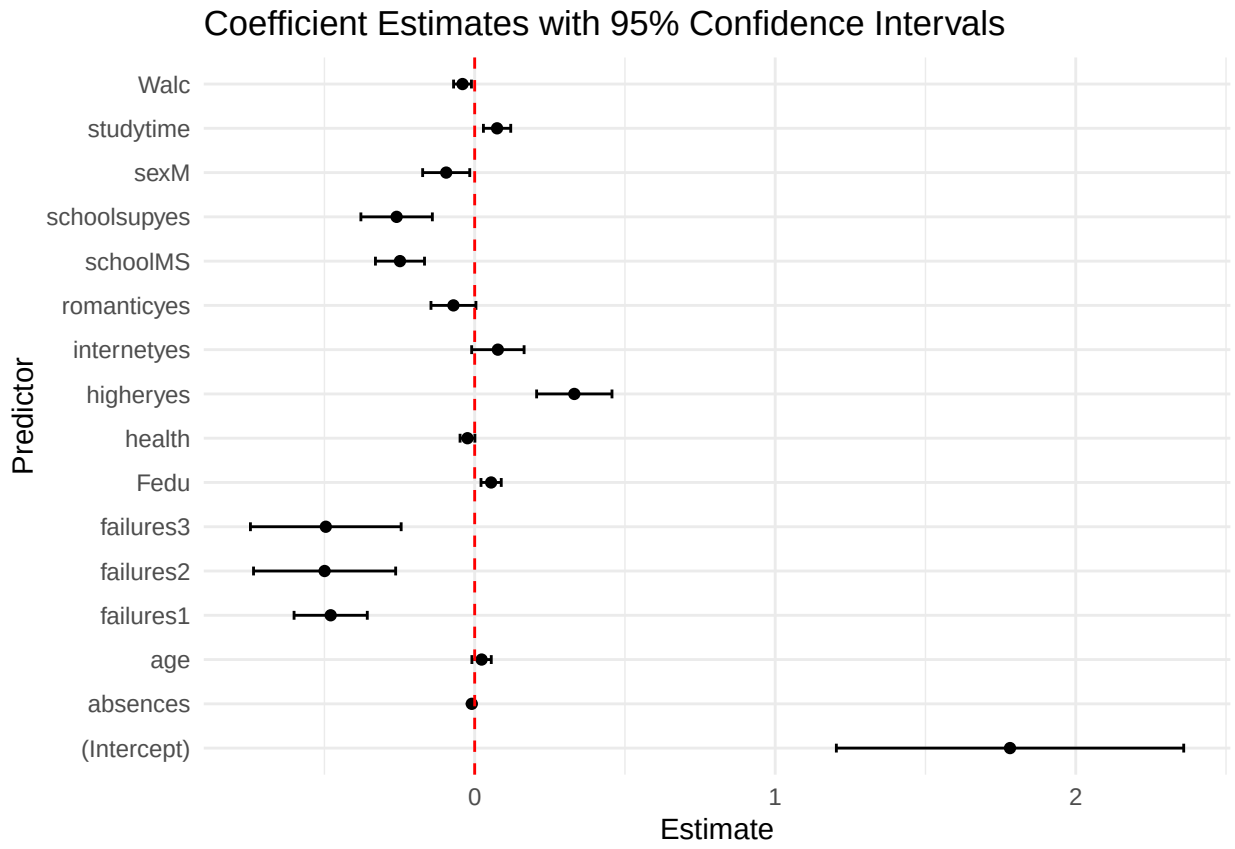
```
##
## Call:
## lm(formula = GPA ~ failures + school + higher + studytime + schoolsup +
##     Fedu + sex + absences + health + internet + romantic + age +
##     Walc, data = df)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05162 -0.28281 -0.02298  0.28100  1.44148
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.781332   0.294301   6.053 2.44e-09 ***
## failures1    -0.479162   0.062036  -7.724 4.43e-14 ***
## failures2    -0.499412   0.120409  -4.148 3.82e-05 ***
## failures3    -0.495413   0.127704  -3.879 0.000116 ***
## schoolMS     -0.248584   0.041566  -5.980 3.72e-09 ***
## higheryes     0.331468   0.063840   5.192 2.80e-07 ***
## studytime     0.074467   0.023127   3.220 0.001348 **
## schoolsupyes -0.259786   0.060547  -4.291 2.06e-05 ***
## Fedu          0.054687   0.017147   3.189 0.001496 **
## sexM         -0.094793   0.039899  -2.376 0.017806 *
## absences     -0.009670   0.004108  -2.354 0.018867 *
## health       -0.023983   0.012573  -1.907 0.056922 .
## internetyes   0.077152   0.044400   1.738 0.082759 .
## romanticyes  -0.070679   0.038102  -1.855 0.064059 .
## age           0.022948   0.016390   1.400 0.161969
## Walc         -0.040525   0.015144  -2.676 0.007646 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4534 on 633 degrees of freedom
## Multiple R-squared:  0.3748, Adjusted R-squared:   0.36
## F-statistic:  25.3 on 15 and 633 DF,  p-value: < 2.2e-16
```

```
# Confindence intevals of Beta Coefficients
tidy_model <- tidy(stepAIC_model, conf.int = TRUE)
tidy_model
```

```
## # A tibble: 16 x 7
##    term         estimate std.error statistic  p.value conf.low conf.high
##    <chr>           <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
##  1 (Intercept)   1.78      0.294        6.05 2.44e- 9  1.20     2.36
##  2 failures1    -0.479     0.0620      -7.72 4.43e-14 -0.601   -0.357
##  3 failures2    -0.499     0.120       -4.15 3.82e- 5 -0.736   -0.263
##  4 failures3    -0.495     0.128       -3.88 1.16e- 4 -0.746   -0.245
##  5 schoolMS     -0.249     0.0416      -5.98 3.72e- 9 -0.330   -0.167
##  6 higheryes     0.331     0.0638       5.19 2.80e- 7  0.206    0.457
##  7 studytime     0.0745    0.0231       3.22 1.35e- 3  0.0291   0.120
##  8 schoolsupyes -0.260     0.0605      -4.29 2.06e- 5 -0.379   -0.141
##  9 Fedu          0.0547    0.0171       3.19 1.50e- 3  0.0210   0.0884
## 10 sexM         -0.0948    0.0399      -2.38 1.78e- 2 -0.173   -0.0164
## 11 absences     -0.00967   0.00411     -2.35 1.89e- 2 -0.0177  -0.00160
## 12 health       -0.0240    0.0126      -1.91 5.69e- 2 -0.0487   0.000708
## 13 internetyes   0.0772    0.0444       1.74 8.28e- 2 -0.0100   0.164
## 14 romanticyes  -0.0707    0.0381      -1.86 6.41e- 2 -0.146    0.00414
## 15 age           0.0229    0.0164       1.40 1.62e- 1 -0.00924  0.0551
## 16 Walc         -0.0405    0.0151      -2.68 7.65e- 3 -0.0703  -0.0108
```

```r
ggplot(tidy_model, aes(x = estimate, y = term)) +
  geom_point() +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high), height = 0.2) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  theme_minimal() +
  labs(
    title = "Coefficient Estimates with 95% Confidence Intervals",
    x = "Estimate",
    y = "Predictor"
  )
```



Exclude internet, romantic, health, and age because they have a possibility of actually being insignificant at explaining the variability in a student's GPA.

```r
ols_model <- lm(GPA ~ Walc + studytime + sex + schoolsup + school + higher + Fedu + failures + absences
summary(ols_model)
```

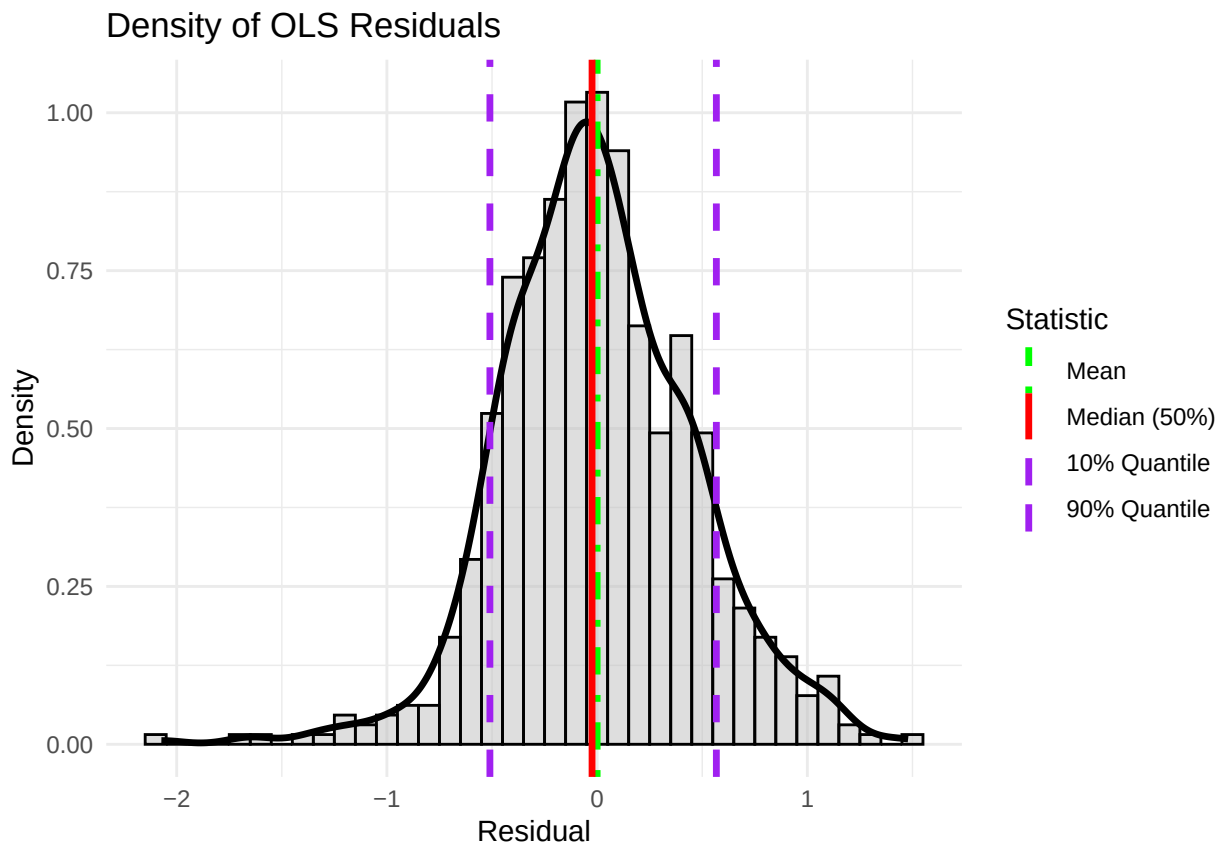
```
##
## Call:
## lm(formula = GPA ~ Walc + studytime + sex + schoolsup + school +
##     higher + Fedu + failures + absences, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07294 -0.29032 -0.02411  0.27814  1.47930
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.111341   0.095033  22.217  < 2e-16 ***
## Walc         -0.039916   0.015125  -2.639 0.008516 **
## studytime     0.077486   0.023138   3.349 0.000859 ***
## sexM         -0.097551   0.039593  -2.464 0.014008 *
## schoolsupyes -0.272407   0.059853  -4.551 6.39e-06 ***
## schoolMS     -0.261297   0.040770  -6.409 2.85e-10 ***
## higheryes     0.322992   0.063305   5.102 4.44e-07 ***
## Fedu          0.057423   0.017086   3.361 0.000823 ***
## failures1    -0.467440   0.061229  -7.634 8.34e-14 ***
## failures2    -0.502074   0.118379  -4.241 2.55e-05 ***
## failures3    -0.483442   0.126844  -3.811 0.000152 ***
## absences     -0.009105   0.004100  -2.221 0.026714 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4562 on 637 degrees of freedom
## Multiple R-squared:  0.3631, Adjusted R-squared:  0.3521
## F-statistic: 33.01 on 11 and 637 DF,  p-value: < 2.2e-16
```

Assumptions Check

Normality check

```r
res <- resid(ols_model)

res_quantiles <- quantile(res, probs = c(0.10, 0.50, 0.90), na.rm = TRUE)
res_mean <- mean(res, na.rm = TRUE)

vline_data <- data.frame(
  xintercept = c(res_mean, res_quantiles["50%"], res_quantiles["10%"], res_quantiles["90%"]),
  Label = factor(c("Mean", "Median (50%)", "10% Quantile", "90% Quantile"),
                 levels = c("Mean","Median (50%)","10% Quantile","90% Quantile"))
)

# Build a small data frame for ggplot
res_df <- data.frame(residuals = res)

ggplot(res_df, aes(x = residuals)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 0.1,
                 fill = "gray", color = "black", alpha = 0.5) +
  geom_density(linewidth = 1.2) +
  geom_vline(data = vline_data,
             aes(xintercept = xintercept, color = Label, linetype = Label),
             linewidth = 1.2) +
  scale_color_manual(values = c("Mean"="green","Median (50%)"="red",
                                "10% Quantile"="purple","90% Quantile"="purple")) +
  scale_linetype_manual(values = c("Mean"="dotdash","Median (50%)"="solid",
                                   "10% Quantile"="dashed","90% Quantile"="dashed")) +
  labs(title = "Density of OLS Residuals",
       x = "Residual", y = "Density", color = "Statistic", linetype = "Statistic") +
  theme_minimal()
```




```
mean(res_df$residuals)
```

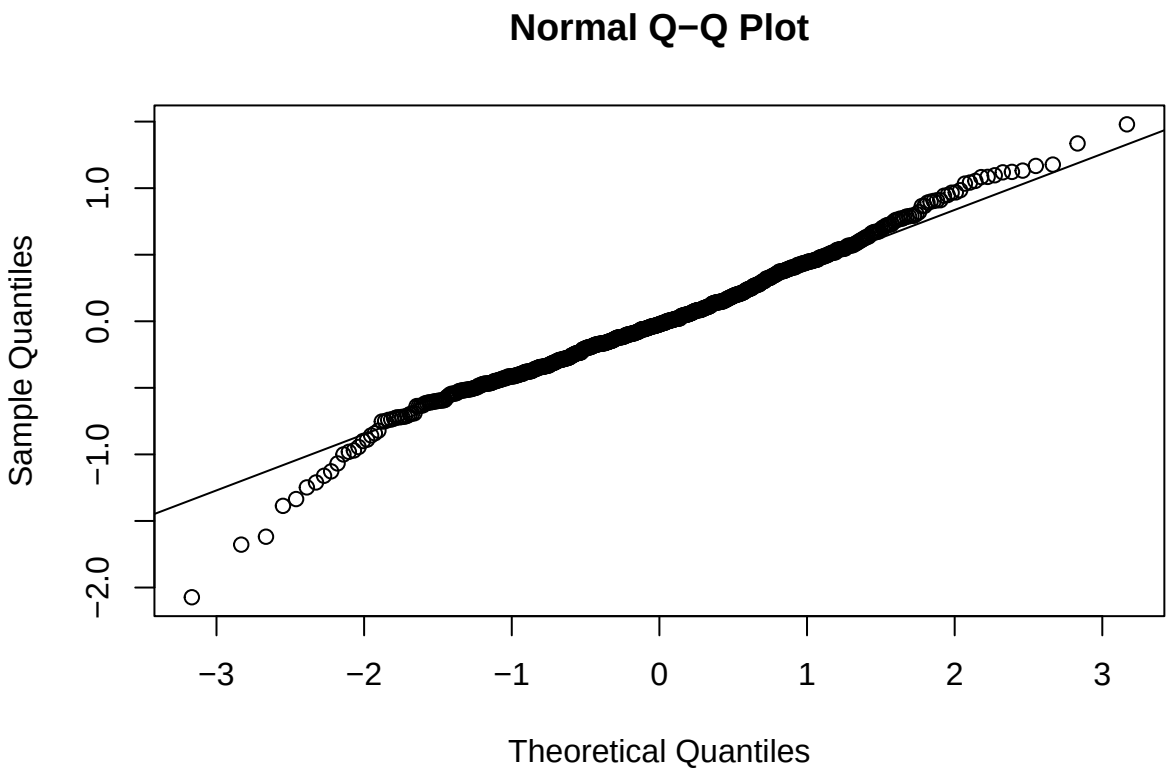
```
## [1] -1.237809e-17
```

```
median(res_df$residuals)
```

```
## [1] -0.0241052
```

OLS Model residuals' mean > median. (Right-Skewed).

```
qqnorm(resid(ols_model))
qqline(resid(ols_model))
```




```
# Shapiro-Wilk normality test
shapiro.test(resid(ols_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(ols_model)
## W = 0.98732, p-value = 2.067e-05
```

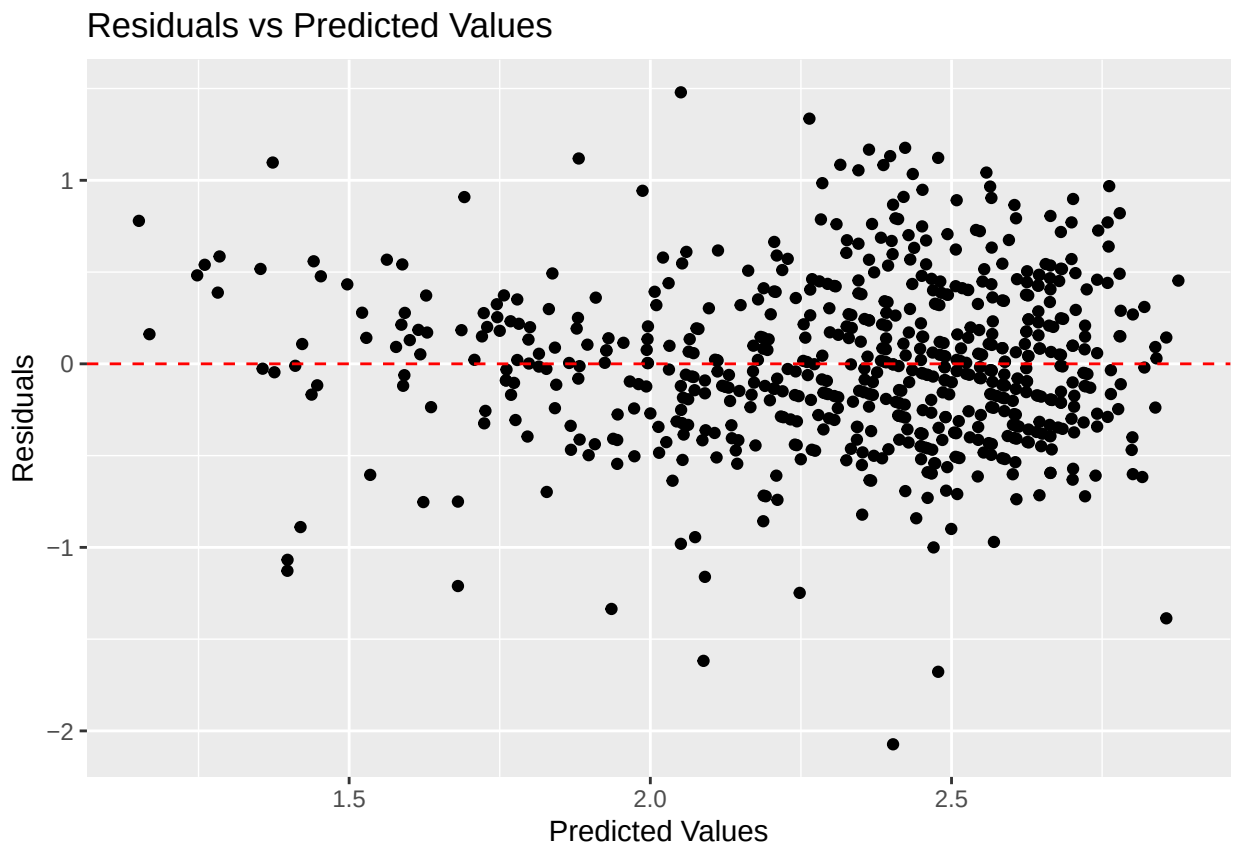
```
# Ho: Sample data comes from a normal distribution
```

(P-value = 2.067e-05) $<$ ($\alpha = 0.05$) $\Rightarrow$ Reject the null because the Shaprio-Wilk's test confirms a statistically significant difference from normality. The density visual suggests a right-skewed distribution.

Constant Variance Check

Fitted vs Residual

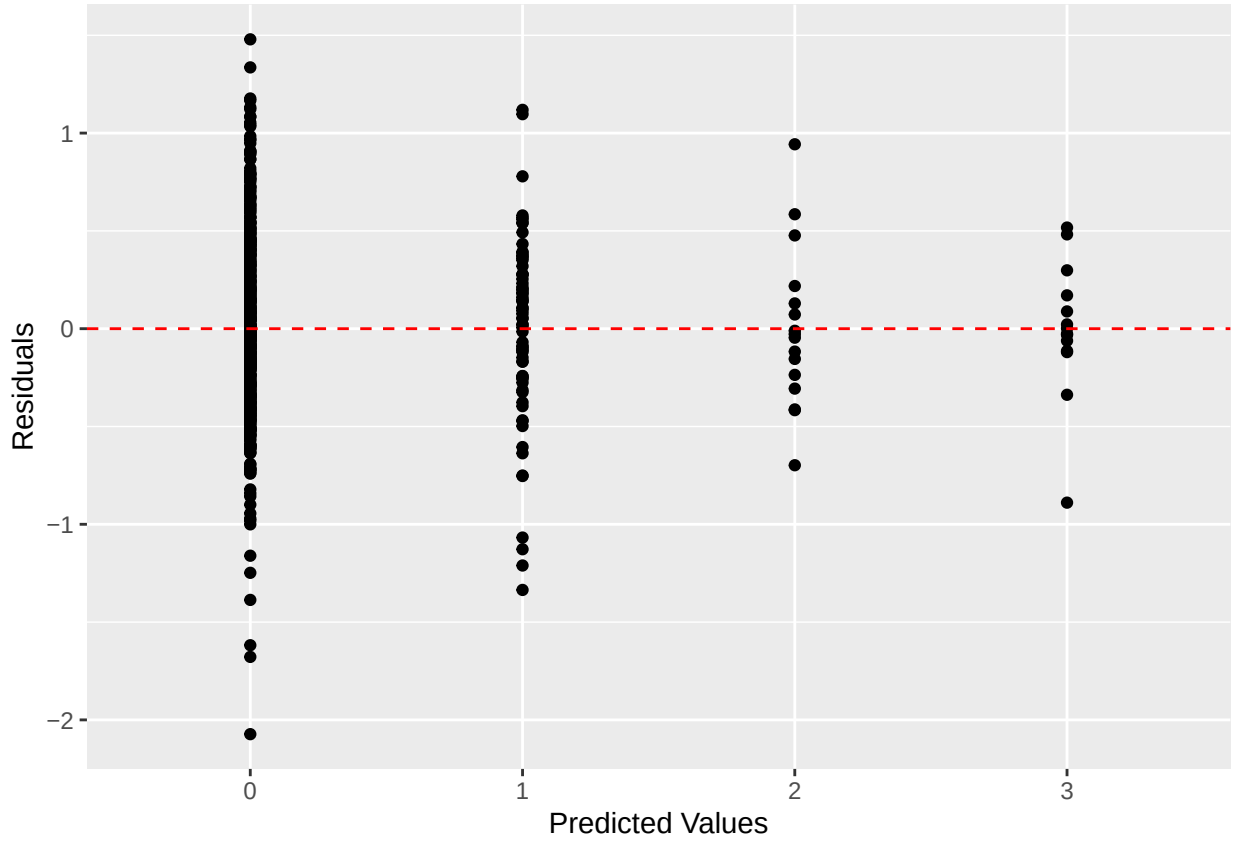
```
ggplot(data = ols_model, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 1) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted Values", y = "Residuals") +
  ggtitle("Residuals vs Predicted Values")
```



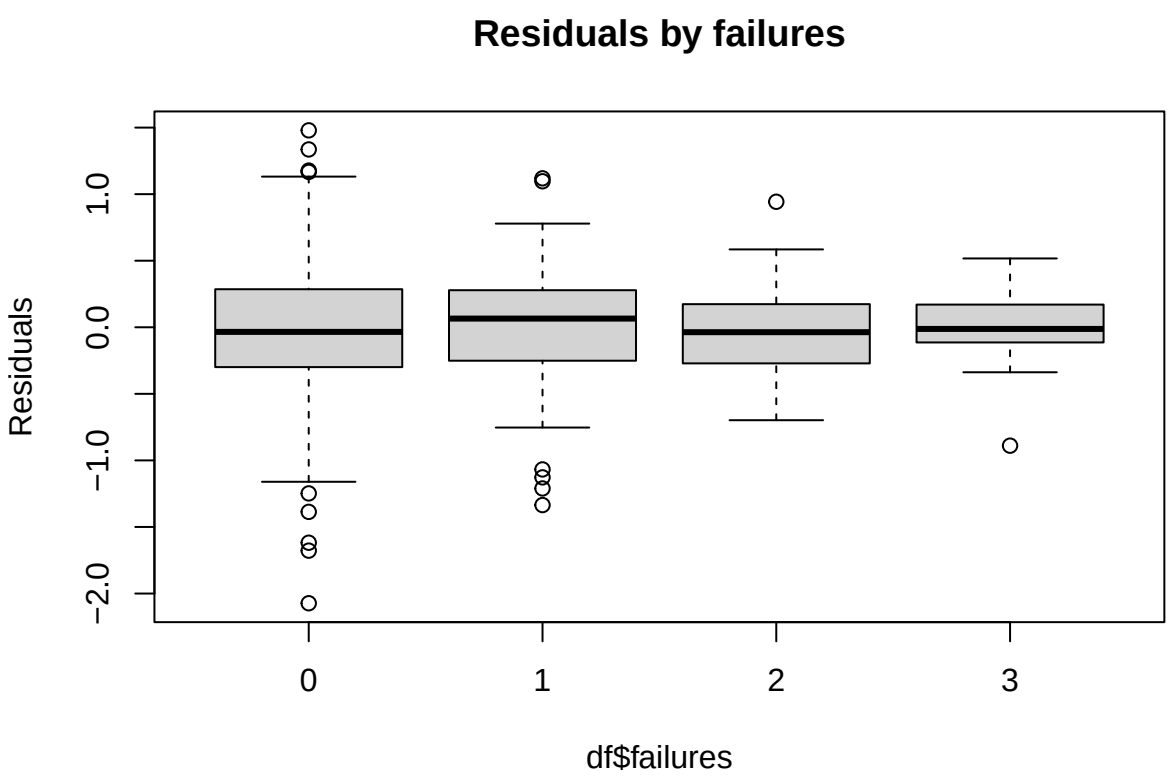
Variance(spread) of residuals does not appear constant - clustering around (2.5, 0).

Failures vs Residuals

```
ggplot(data = ols_model, aes(x = failures, y = .resid)) +
  geom_point(alpha = 1) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted Values", y = "Residuals")
```




```
boxplot(resid(ols_model) ~ df$failures, main = "Residuals by failures", ylab = "Residuals")
```




```
# Levene's Test
# Used to check whether goups (levels) have equal variance.
# H0: the residual variance of 2(or more) groups are equal (homogeneity of variance) vs Ha: at least on
leveneTest(resid(ols_model) ~ model.frame(ols_model)$failures)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  0.9651 0.4087
##       645
```

Levene's test reports ($p - value = 0.4087$) > (α=0.05). This suggests that there is not enough evidence to reject the Null. Therefore, we do not detect variance difference across the different failures. There is a possibility that variance is constant.

```
prop.table(table(df$failures))
```

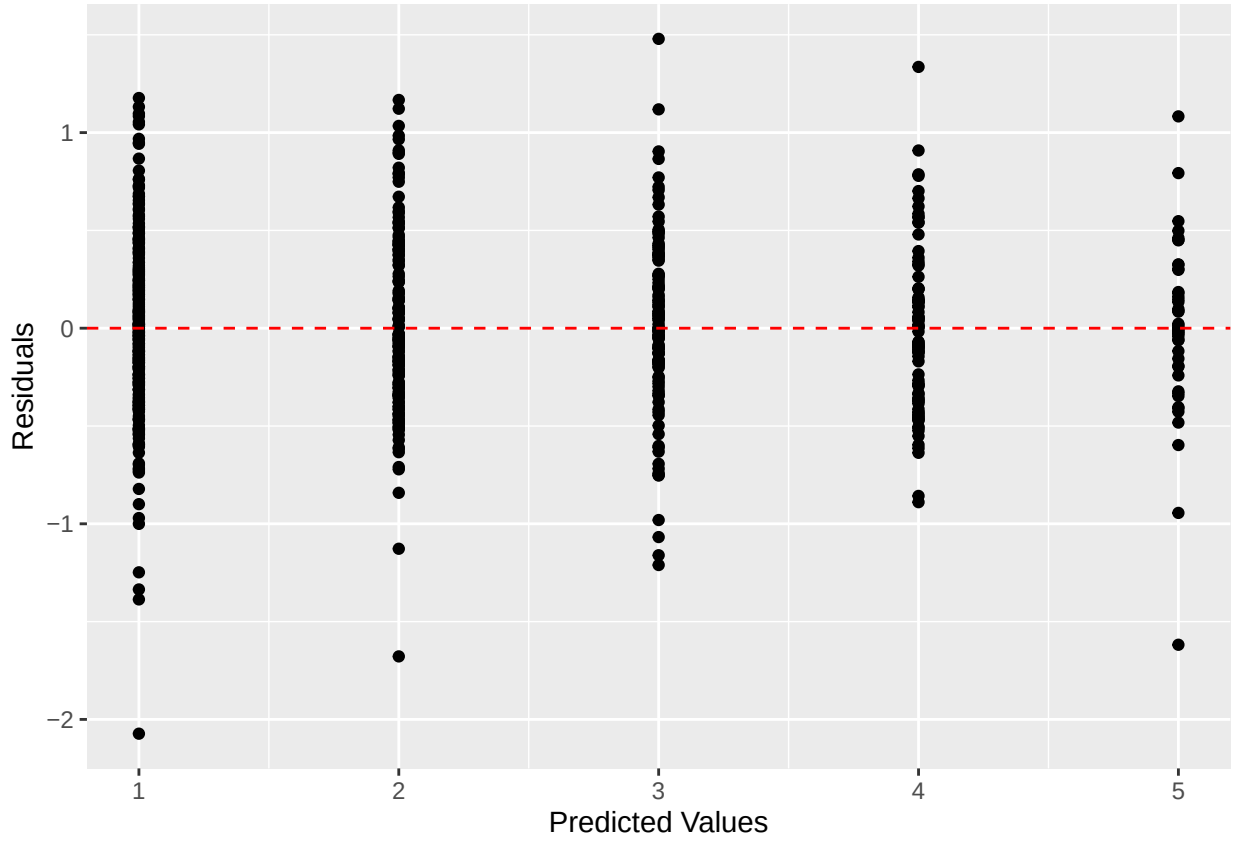
```
##
##          0          1          2          3
## 0.84591680 0.10785824 0.02465331 0.02157165
```

Levene's test may not be robust to this imbalanced behavior. There's no evidence of variance difference, but the power behind this claim may be limited due to small groups.

Walc vs Residuals

```
ggplot(data = ols_model, aes(x = Walc, y = .resid)) +
  geom_point(alpha = 1) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
```

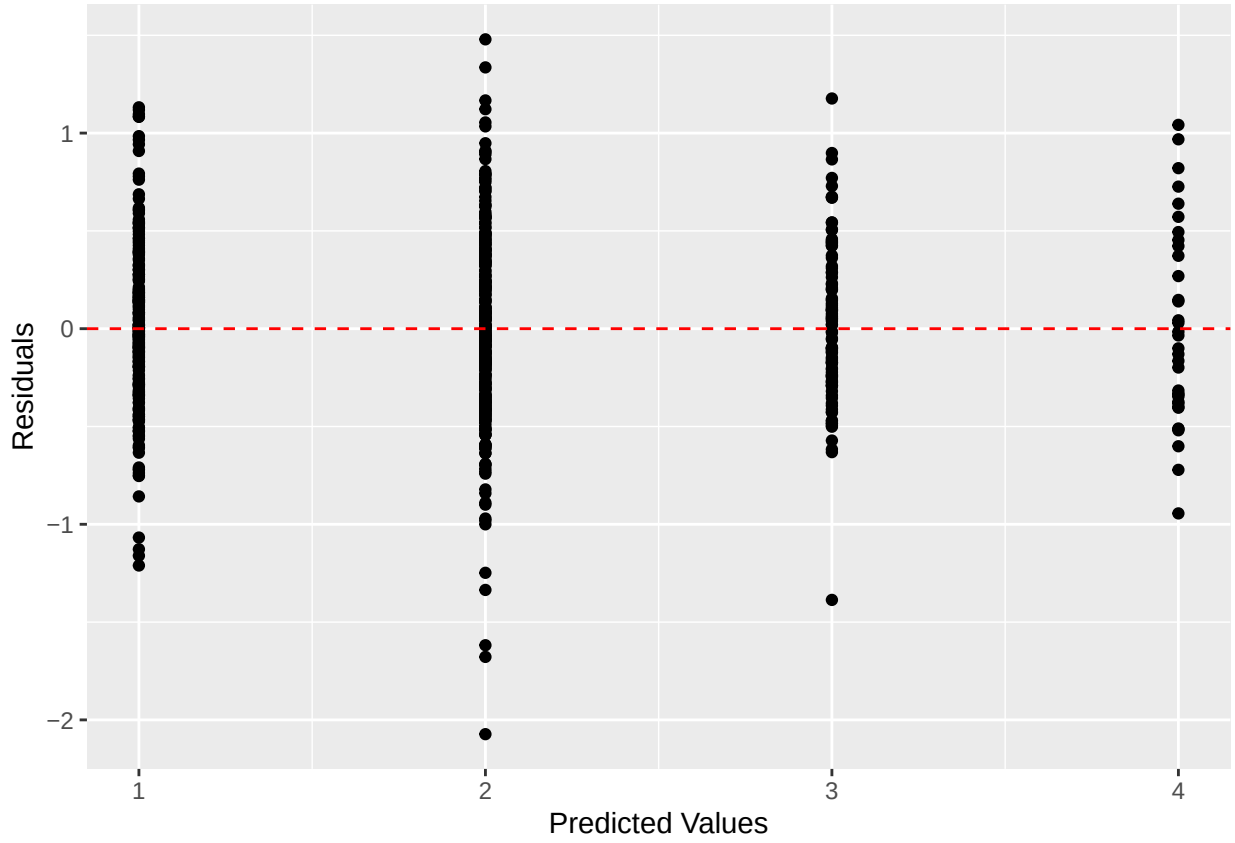
```
  labs(x = "Predicted Values", y = "Residuals")
```



There is a slight funnel pattern (from left to right).

Studytime vs Residuals

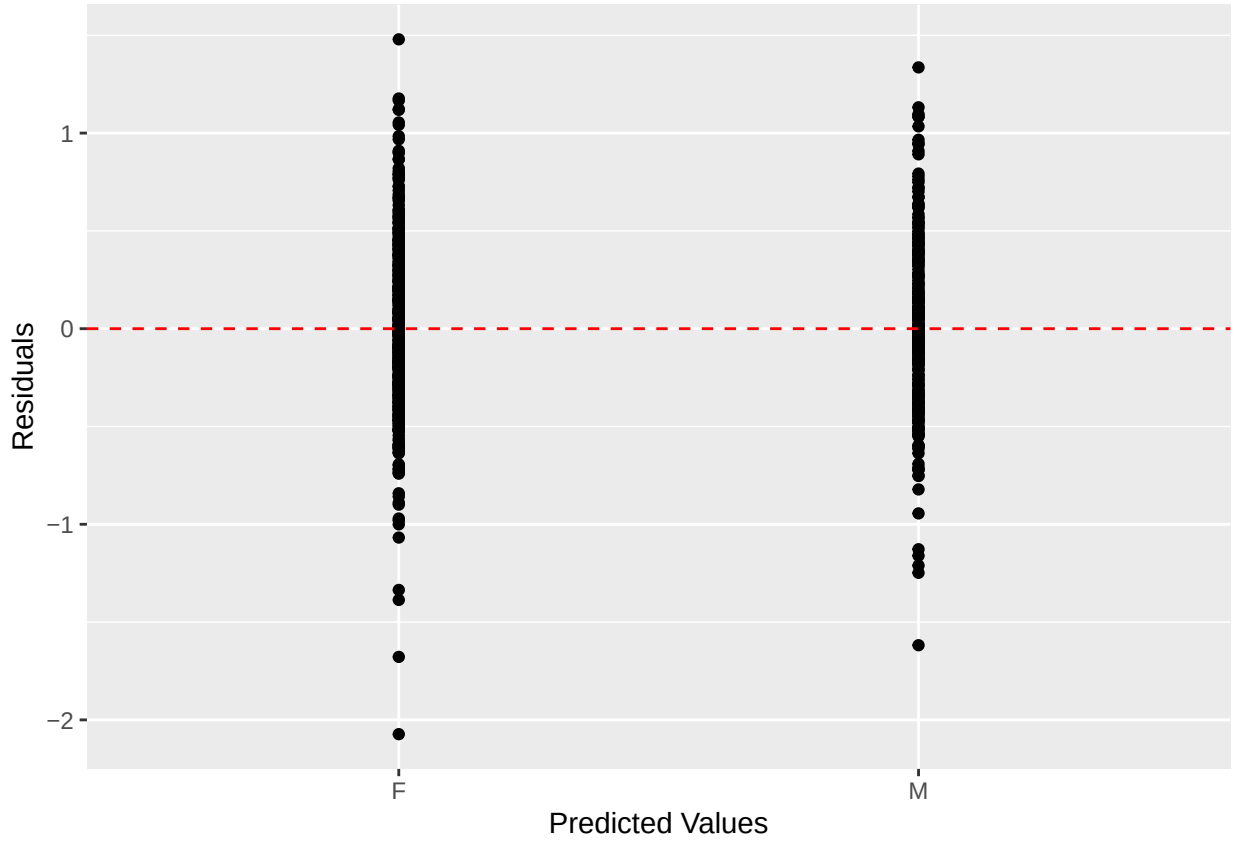
```
ggplot(data = ols_model, aes(x = studytime, y = .resid)) +
  geom_point(alpha = 1) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted Values", y = "Residuals")
```



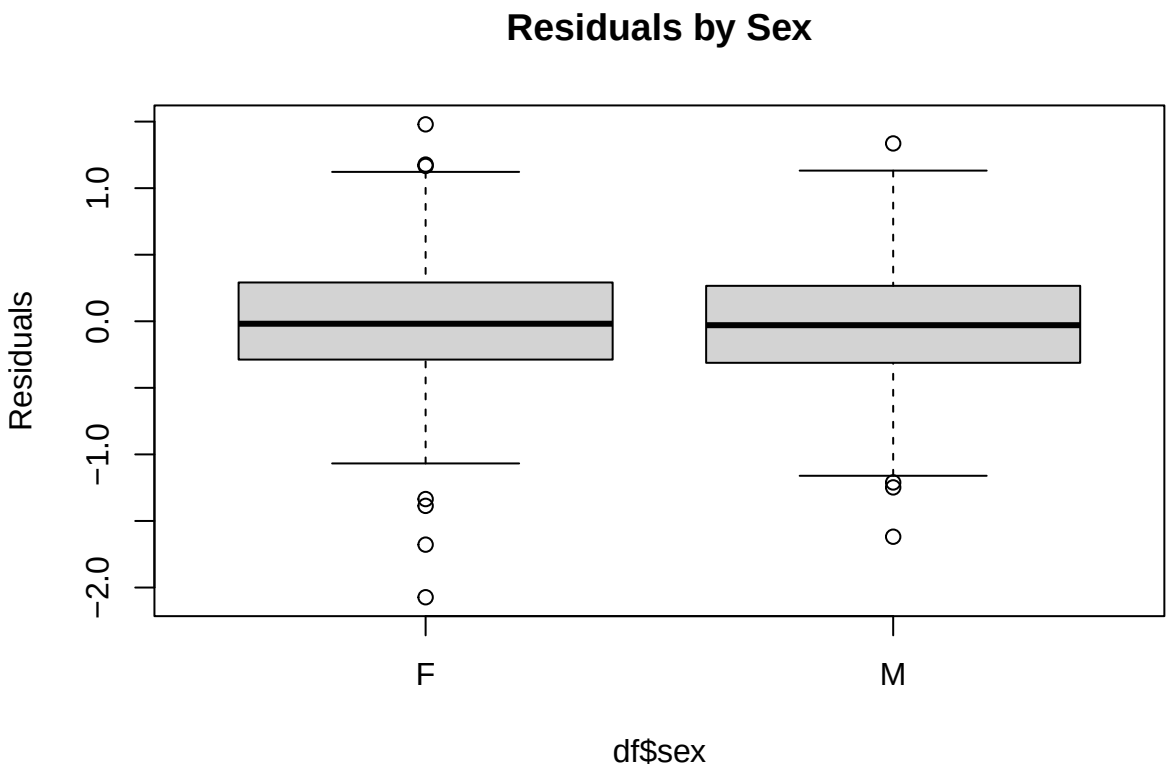
Variance does not appear constant across the plot.

Sex vs Residuals

```
ggplot(data = ols_model, aes(x = sex, y = .resid)) +
  geom_point(alpha = 1) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted Values", y = "Residuals")
```




```
boxplot(resid(ols_model) ~ df$sex, main = "Residuals by Sex", ylab = "Residuals")
```




```
# Levene's Test
leveneTest(resid(ols_model) ~ model.frame(ols_model)$sex)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.4273 0.5136
##       647
```

Levene's test reports ($p - value = 0.5136$) > (α=0.05). This suggests that there is not enough evidence to reject the Null. Therefore, we do not detect variance difference between the two different sexes. There is a possibility that variance is constant.

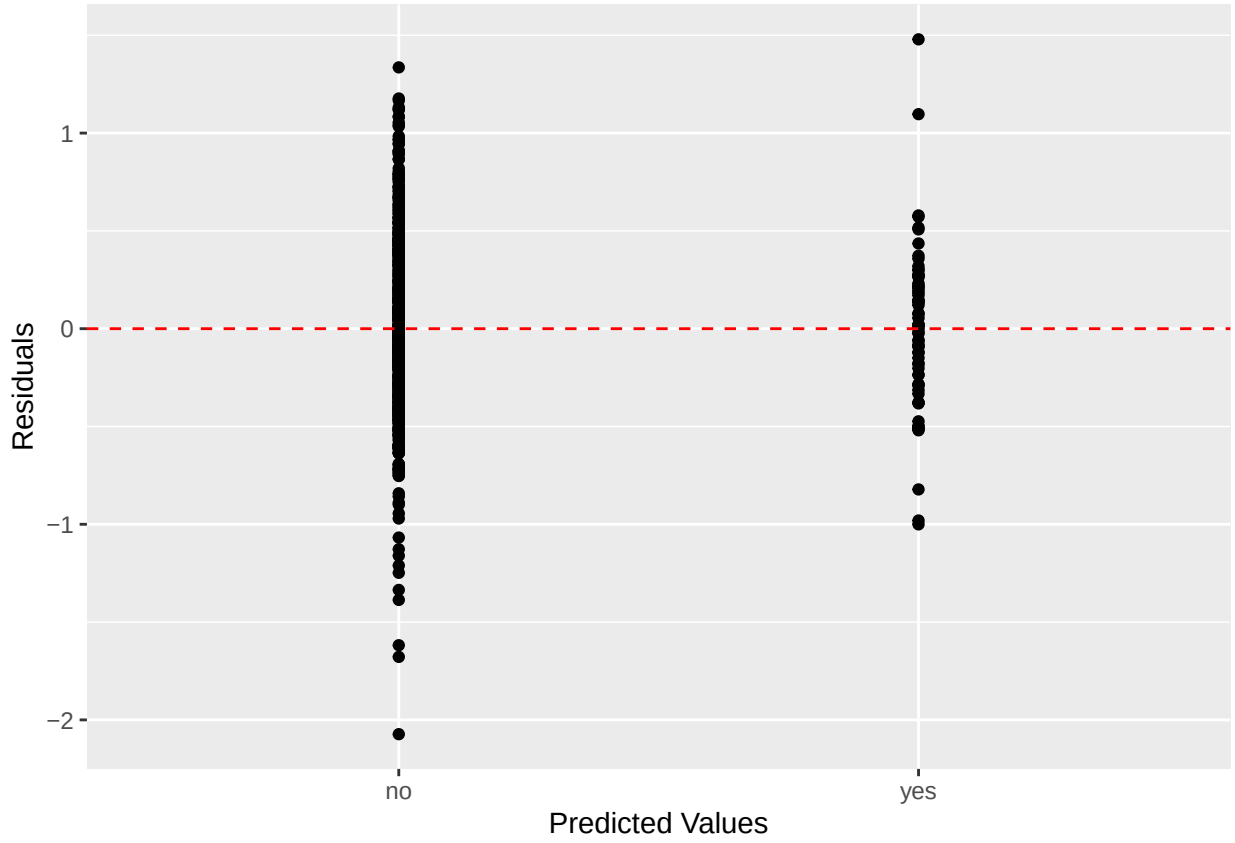
```
prop.table(table(df$sex))
```

```
##
##         F         M
## 0.5901387 0.4098613
```

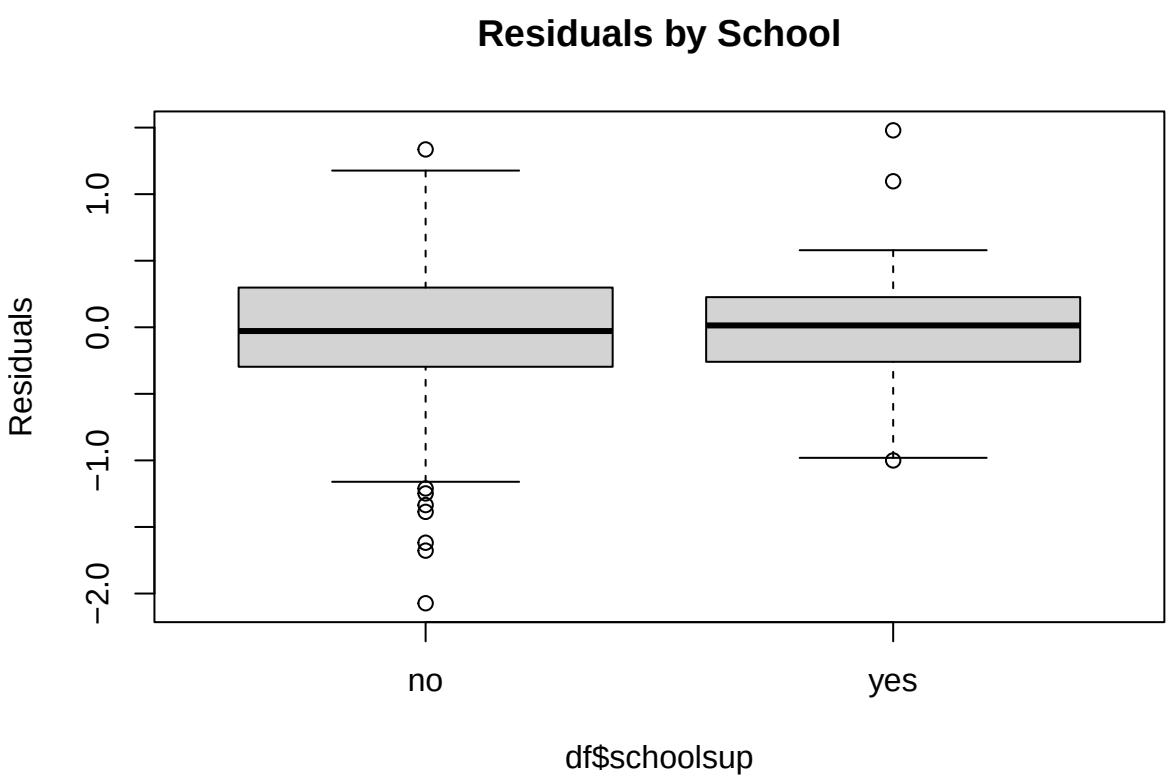
Due to this imbalanced behavior, Levene's power may be limited.

Schoolsup vs Residauls

```
ggplot(data = ols_model, aes(x = schoolsup, y = .resid)) +
  geom_point(alpha = 1) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted Values", y = "Residuals")
```




```
boxplot(resid(ols_model) ~ df$schoolsup, main = "Residuals by School", ylab = "Residuals")
```




```
# Levene's Test
leveneTest(resid(ols_model) ~ model.frame(ols_model)$schoolsup)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  1.6437 0.2003
##       647
```

Levene's test reports ($p - value = 0.2003$) > (α=0.05). This suggests that there is not enough evidence to reject the Null. Therefore, we do not detect variance difference between the groups in school support. There is a possibility that variance is constant.

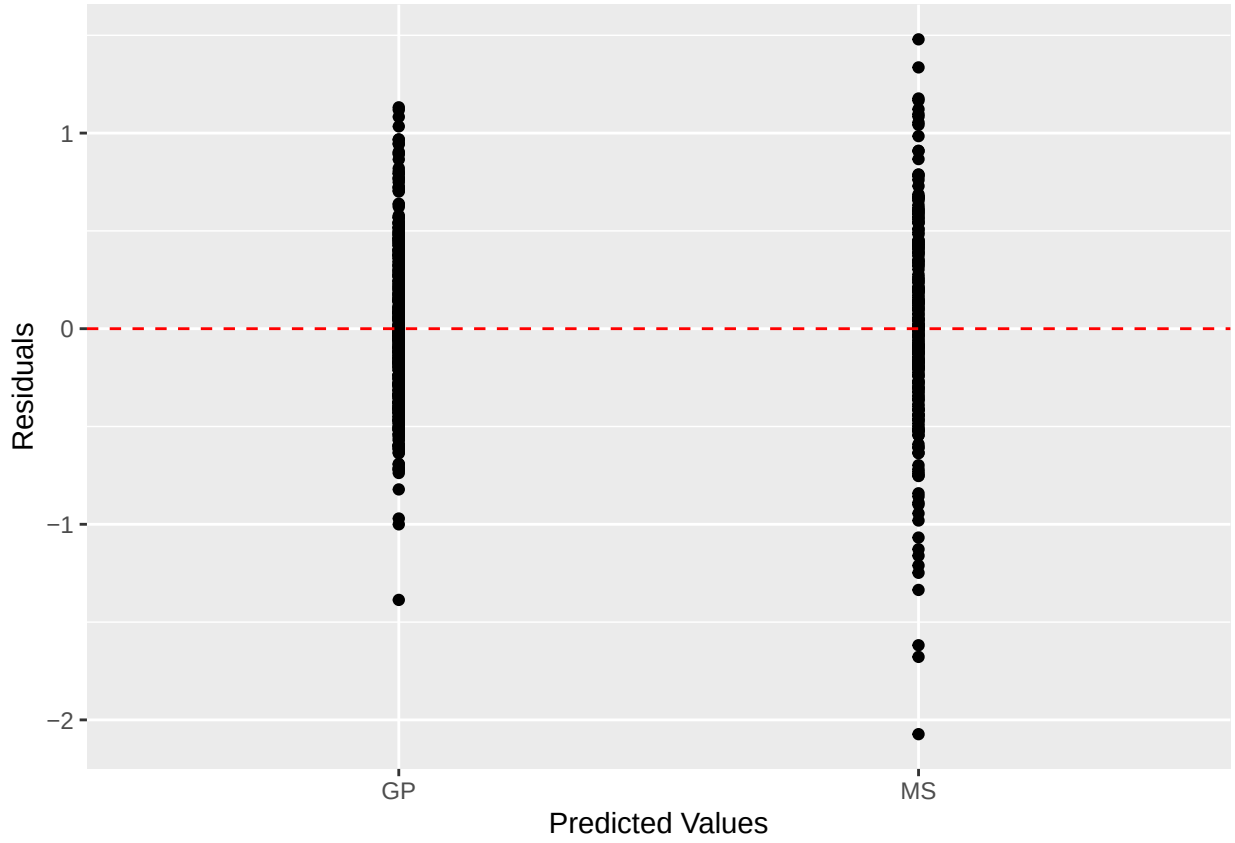
```
prop.table(table(df$schoolsup))
```

```
##
##        no       yes
## 0.8952234 0.1047766
```

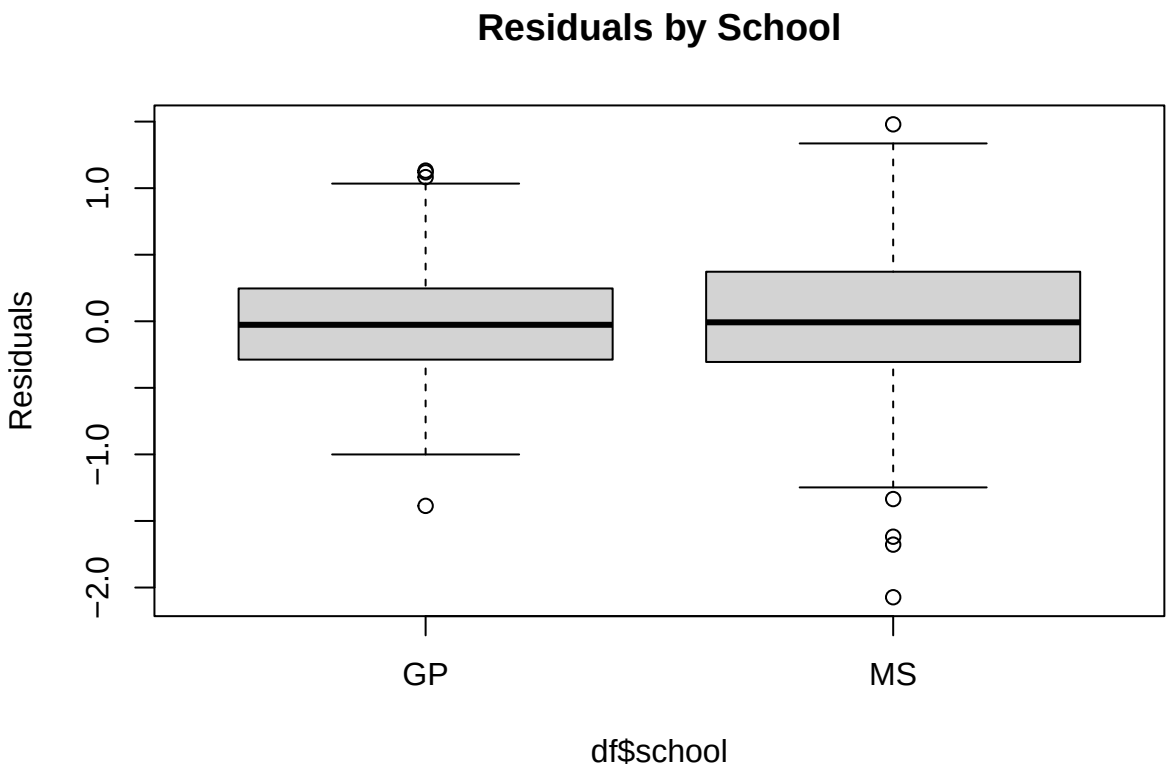
Due to this very unbalance behavior, Levene's power may be limited.

School vs Residauls

```
ggplot(data = ols_model, aes(x = school, y = .resid)) +
  geom_point(alpha = 1) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted Values", y = "Residuals")
```




```
boxplot(resid(ols_model) ~ df$school, main = "Residuals by School", ylab = "Residuals")
```



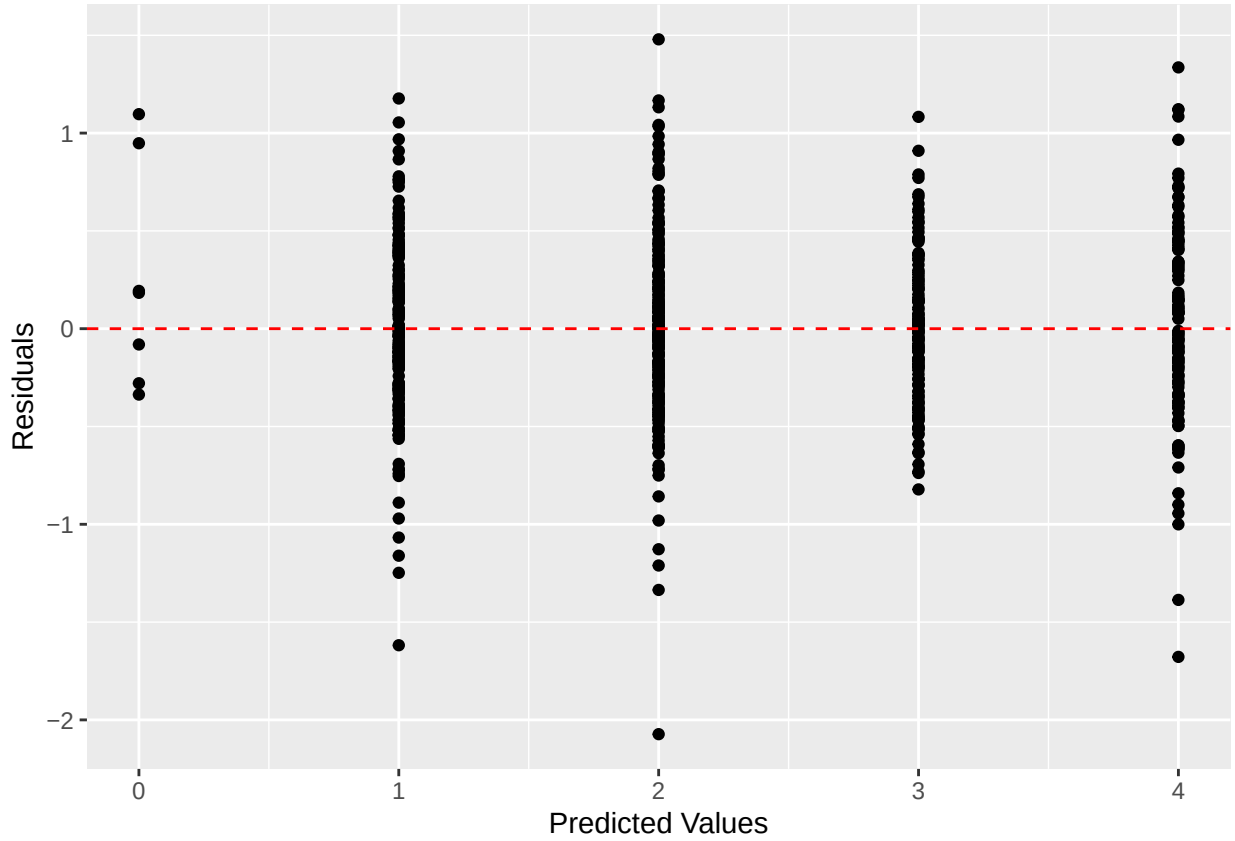

```
# Levene's Test
leveneTest(resid(ols_model) ~ model.frame(ols_model)$school)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   1  19.997 9.16e-06 ***
##       647
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Levene's test suggest that the residual variance is not equal across the two different school groups - this matches with the plots. Therefore, we conclude that the constant variance assumption is not satisfied.

Fedu vs Residuals

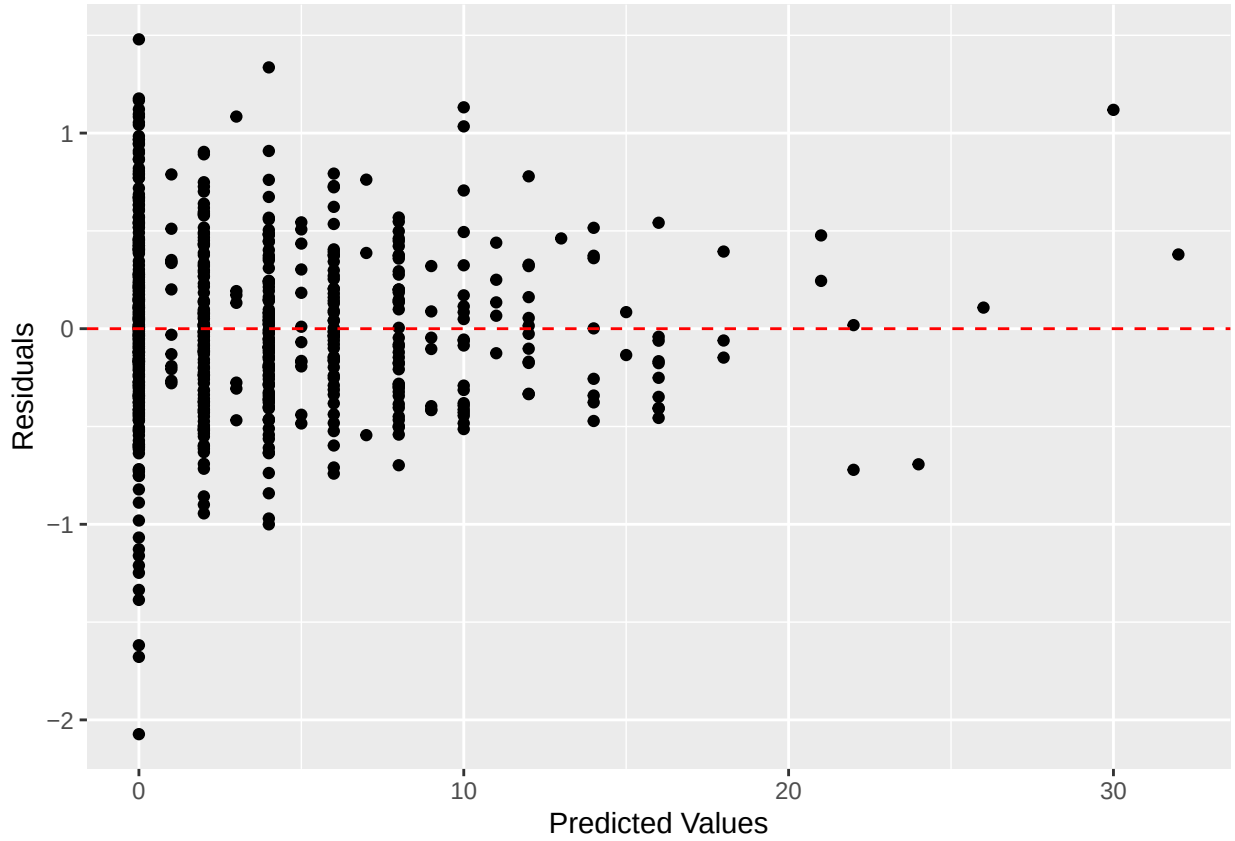
```
ggplot(data = ols_model, aes(x = Fedu, y = .resid)) +
  geom_point(alpha = 1) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted Values", y = "Residuals")
```



Variance does not appear constant across all levels of father's education.

Absences vs Residuals

```
ggplot(data = ols_model, aes(x = absences, y = .resid)) +
  geom_point(alpha = 1) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted Values", y = "Residuals")
```



There appears to be a funnel-like pattern. Constant variance assumption is not satisfied.

Fit an OLS Model without Zero GPA Students

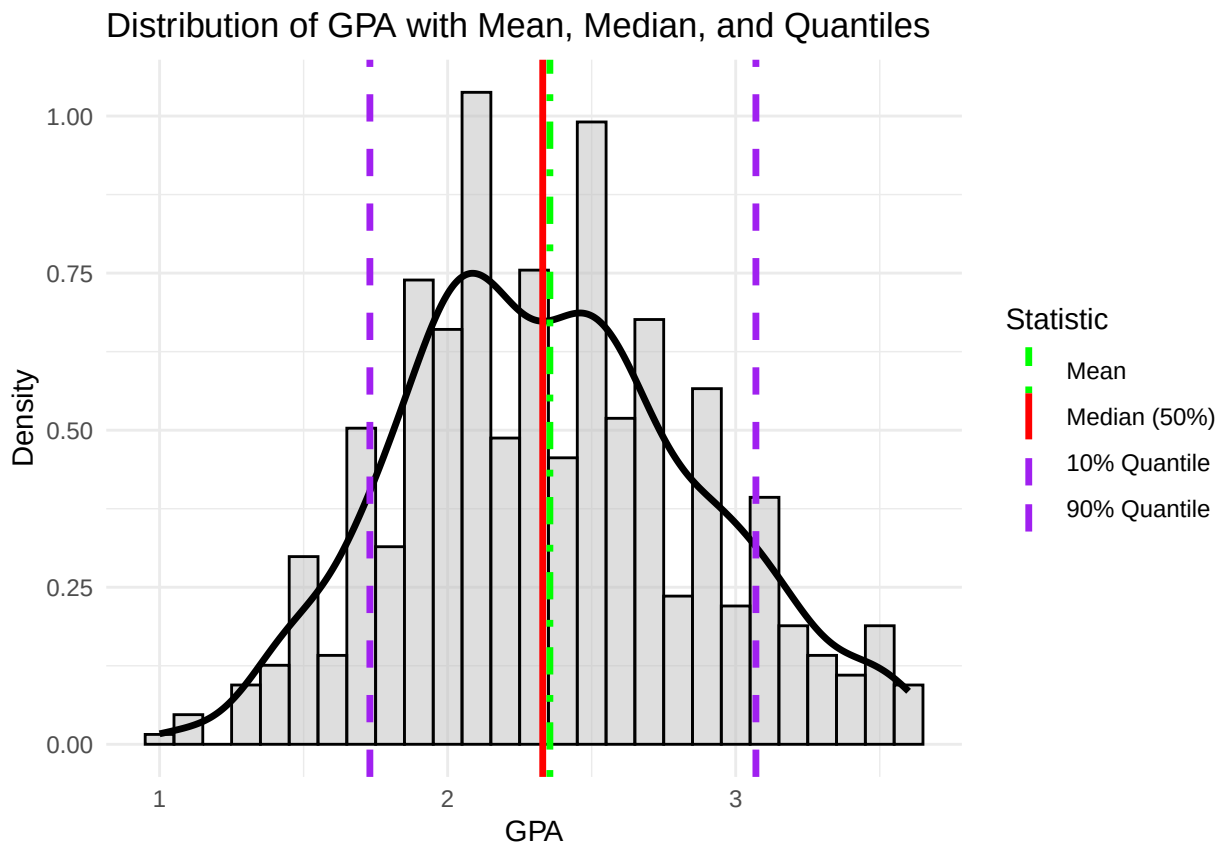
```
# Compute IQR boundaries
Q1 <- quantile(df$GPA, 0.25, na.rm = TRUE)
Q3 <- quantile(df$GPA, 0.75, na.rm = TRUE)
IQR_value <- Q3 - Q1

# Thresholds (1.5 * IQR rule)
lower_bound <- Q1 - 1.5 * IQR_value
upper_bound <- Q3 + 1.5 * IQR_value

# Remove outliers (removing students with GPA=0)
df_no_outliers <- df[df$GPA >= lower_bound & df$GPA <= upper_bound, ]

summary(df_no_outliers$GPA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.330   2.355   2.670   3.600
```

```
n_removed <- nrow(df) - nrow(df_no_outliers)
cat("Number of outliers removed:", n_removed)
```

```
## Number of outliers removed: 13
```

```r
# Compute quantiles and mean
gpa_quantiles <- quantile(df_no_outliers$GPA, probs = c(0.1, 0.5, 0.9))
gpa_mean <- mean(df_no_outliers$GPA)

# Create a data frame for vertical lines
vline_data <- data.frame(
  xintercept = c(
    gpa_quantiles["10%"], gpa_quantiles["50%"], gpa_quantiles["90%"], gpa_mean
  ),
  Label = factor(c(
    "10% Quantile", "Median (50%)", "90% Quantile", "Mean"
  ),
  levels = c("Mean", "Median (50%)", "10% Quantile", "90% Quantile"))
)

# Plot
ggplot(df_no_outliers, aes(x = GPA)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.1, fill = "gray", color = "black", alpha = 0.5) +
  geom_density(color = "black", size = 1.2) +

  # Add vlines with mapped aesthetics for legend
  geom_vline(data = vline_data, aes(xintercept = xintercept, color = Label, linetype = Label), size = 1

  scale_color_manual(values = c(
    "Mean" = "green",
    "Median (50%)" = "red",
    "10% Quantile" = "purple",
    "90% Quantile" = "purple"
  )) +

  scale_linetype_manual(values = c(
    "Mean" = "dotdash",
    "Median (50%)" = "solid",
    "10% Quantile" = "dashed",
    "90% Quantile" = "dashed"
  )) +

  labs(
    title = "Distribution of GPA with Mean, Median, and Quantiles",
    x = "GPA",
    y = "Density",
    color = "Statistic",
    linetype = "Statistic"
  ) +
  theme_minimal()
```



Fit OLS Model 2

```
# Initialize models
null_model2 <- lm(GPA ~ 1, data = df_no_outliers)
Full_model2 <- lm(GPA ~ ., data = df_no_outliers)
#summary(Full_model2)
```

```
# Variable selection
stepAIC_model2 <- stepAIC(null_model2, scope = list(lower = null_model2, upper = Full_model2), directio
summary(stepAIC_model2)
```

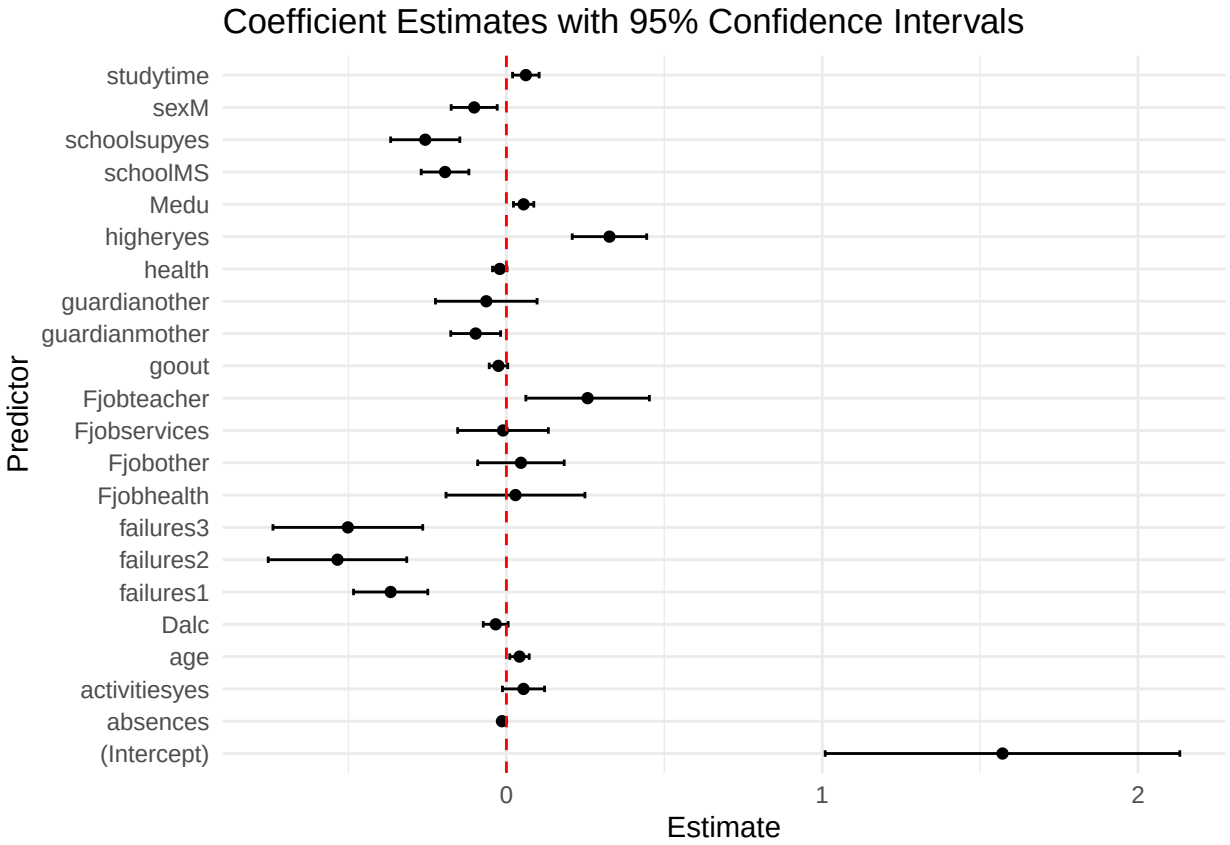
```
##
## Call:
## lm(formula = GPA ~ failures + higher + school + studytime + schoolsup +
##     absences + Medu + sex + Fjob + age + guardian + health +
##     activities + goout + Dalc, data = df_no_outliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34216 -0.26626 -0.03543  0.26206  1.33659
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.570842   0.285839   5.496 5.71e-08 ***
## failures1      -0.366506   0.059856  -6.123 1.64e-09 ***
## failures2      -0.534801   0.111662  -4.789 2.10e-06 ***
```

```
## failures3      -0.502030   0.120723  -4.159 3.66e-05 ***
## higheryes       0.326335   0.060006   5.438 7.77e-08 ***
## schoolMS       -0.194260   0.038428  -5.055 5.68e-07 ***
## studytime       0.061597   0.021402   2.878 0.004140 **
## schoolsupyes   -0.257006   0.055755  -4.610 4.91e-06 ***
## absences       -0.013976   0.003833  -3.647 0.000288 ***
## Medu            0.054544   0.016401   3.326 0.000935 ***
## sexM           -0.101955   0.037064  -2.751 0.006121 **
## Fjobhealth      0.028760   0.111974   0.257 0.797383
## Fjobother       0.046053   0.069765   0.660 0.509428
## Fjobservices   -0.010686   0.073129  -0.146 0.883865
## Fjobteacher     0.257186   0.099566   2.583 0.010022 *
## age             0.041473   0.015654   2.649 0.008274 **
## guardianmother -0.097269   0.040161  -2.422 0.015726 *
## guardianother  -0.063723   0.081899  -0.778 0.436828
## health         -0.020922   0.011627  -1.800 0.072424 .
## activitiesyes   0.054142   0.033939   1.595 0.111168
## goout          -0.025101   0.014870  -1.688 0.091920 .
## Dalc           -0.033890   0.020063  -1.689 0.091694 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4143 on 614 degrees of freedom
## Multiple R-squared:  0.3803, Adjusted R-squared:  0.3591
## F-statistic: 17.94 on 21 and 614 DF,  p-value: < 2.2e-16
```

```
# Confindence intervals of Beta Coefficients
tidy_model2<- tidy(stepAIC_model2, conf.int = TRUE)
tidy_model2
```

```
## # A tibble: 22 x 7
##    term         estimate std.error statistic       p.value conf.low conf.high
##    <chr>           <dbl>     <dbl>     <dbl>         <dbl>    <dbl>     <dbl>
##  1 (Intercept)    1.57     0.286        5.50 0.0000000571    1.01     2.13
##  2 failures1     -0.367    0.0599      -6.12 0.00000000164  -0.484   -0.249
##  3 failures2     -0.535    0.112       -4.79 0.00000210     -0.754   -0.316
##  4 failures3     -0.502    0.121       -4.16 0.0000366      -0.739   -0.265
##  5 higheryes      0.326    0.0600       5.44 0.0000000777    0.208    0.444
##  6 schoolMS      -0.194    0.0384      -5.06 0.000000568    -0.270   -0.119
##  7 studytime      0.0616   0.0214       2.88 0.00414         0.0196   0.104
##  8 schoolsupyes  -0.257    0.0558      -4.61 0.00000491     -0.367   -0.148
##  9 absences      -0.0140   0.00383     -3.65 0.000288       -0.0215  -0.00645
## 10 Medu           0.0545   0.0164       3.33 0.000935        0.0223   0.0868
## # i 12 more rows
```

```
ggplot(tidy_model2, aes(x = estimate, y = term)) +
  geom_point() +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high), height = 0.2) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  theme_minimal() +
  labs(
    title = "Coefficient Estimates with 95% Confidence Intervals",
    x = "Estimate",
    y = "Predictor"
  )
```



Exclude health, higher education, guardian, gout, Fjob, Dalc, and activities because they have a possibility of actually being insignificant to explaining the variability in a student's GPA.

```
ols_model2 <- lm(GPA ~ studytime + sex + schoolsup + school + Medu + higher + failures + age + absences
summary(ols_model2)
```

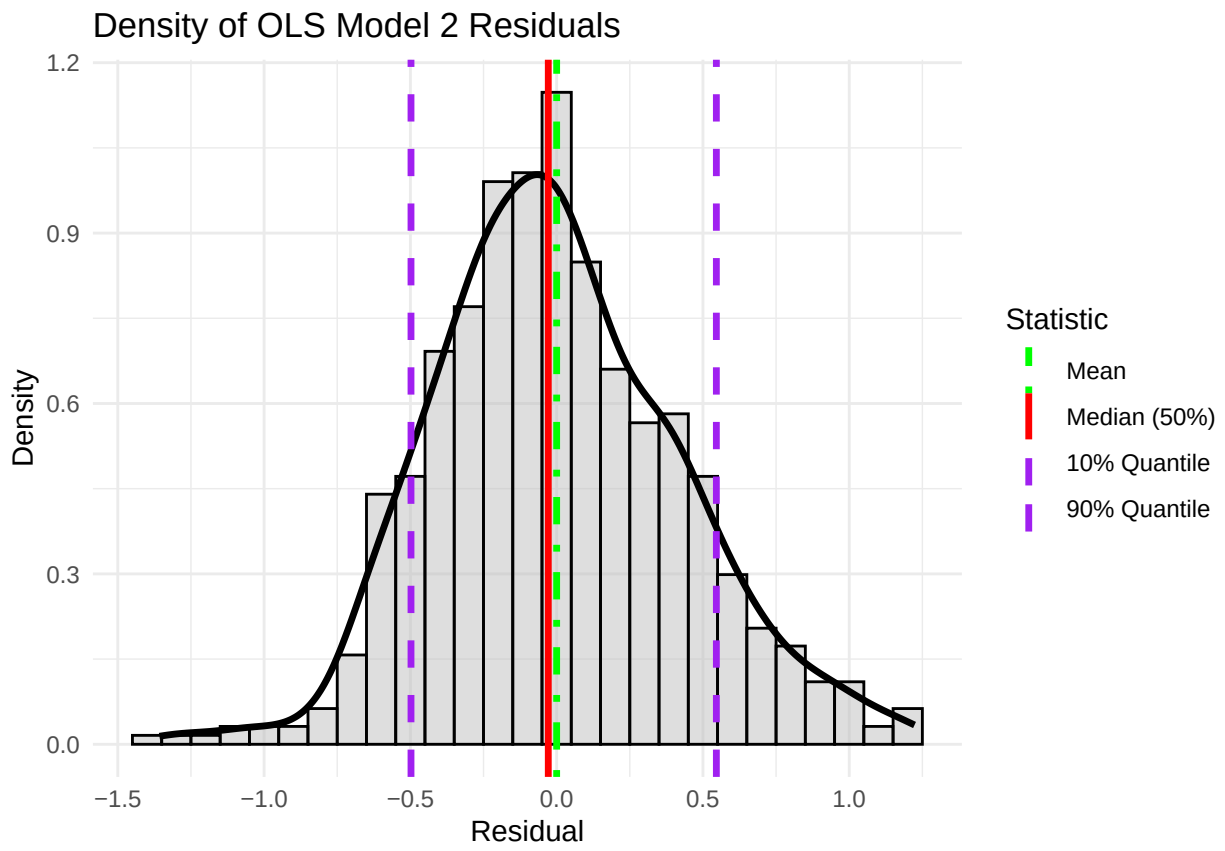
```
##
## Call:
## lm(formula = GPA ~ studytime + sex + schoolsup + school + Medu +
##     higher + failures + age + absences, data = df_no_outliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35780 -0.27565 -0.02865  0.27476  1.22320
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.484472   0.271493   5.468 6.60e-08 ***
## studytime     0.066520   0.021453   3.101 0.002018 **
## sexM         -0.119105   0.035684  -3.338 0.000895 ***
## schoolsupyes -0.247932   0.056131  -4.417 1.18e-05 ***
## schoolMS     -0.206539   0.038247  -5.400 9.48e-08 ***
## Medu          0.063784   0.015849   4.025 6.41e-05 ***
## higheryes     0.326670   0.060580   5.392 9.88e-08 ***
## failures1    -0.374517   0.060144  -6.227 8.73e-10 ***
## failures2    -0.563045   0.111630  -5.044 5.99e-07 ***
## failures3    -0.502477   0.121969  -4.120 4.30e-05 ***
```

```
## age           0.033056   0.015186   2.177 0.029868 *
## absences     -0.015763   0.003815  -4.132 4.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4212 on 624 degrees of freedom
## Multiple R-squared:  0.3491, Adjusted R-squared:  0.3376
## F-statistic: 30.43 on 11 and 624 DF,  p-value: < 2.2e-16
```

Assumptions Check

Normality check

```
res2 <- resid(ols_model2)

res2_quantiles <- quantile(res2, probs = c(0.10, 0.50, 0.90), na.rm = TRUE)
res2_mean <- mean(res2, na.rm = TRUE)

vline_data2 <- data.frame(
  xintercept = c(res2_mean, res2_quantiles["50%"], res2_quantiles["10%"], res2_quantiles["90%"]),
  Label = factor(c("Mean", "Median (50%)", "10% Quantile", "90% Quantile"),
                 levels = c("Mean","Median (50%)","10% Quantile","90% Quantile"))
)

# Build a small data frame for ggplot
res2_df <- data.frame(residuals = res2)

ggplot(res2_df, aes(x = residuals)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 0.1,
                 fill = "gray", color = "black", alpha = 0.5) +
  geom_density(linewidth = 1.2) +
  geom_vline(data = vline_data2,
             aes(xintercept = xintercept, color = Label, linetype = Label),
             linewidth = 1.2) +
  scale_color_manual(values = c("Mean"="green","Median (50%)"="red",
                                "10% Quantile"="purple","90% Quantile"="purple")) +
  scale_linetype_manual(values = c("Mean"="dotdash","Median (50%)"="solid",
                                   "10% Quantile"="dashed","90% Quantile"="dashed")) +
  labs(title = "Density of OLS Model 2 Residuals",
       x = "Residual", y = "Density", color = "Statistic", linetype = "Statistic") +
  theme_minimal()
```




```
mean(res2_df$residuals)
```

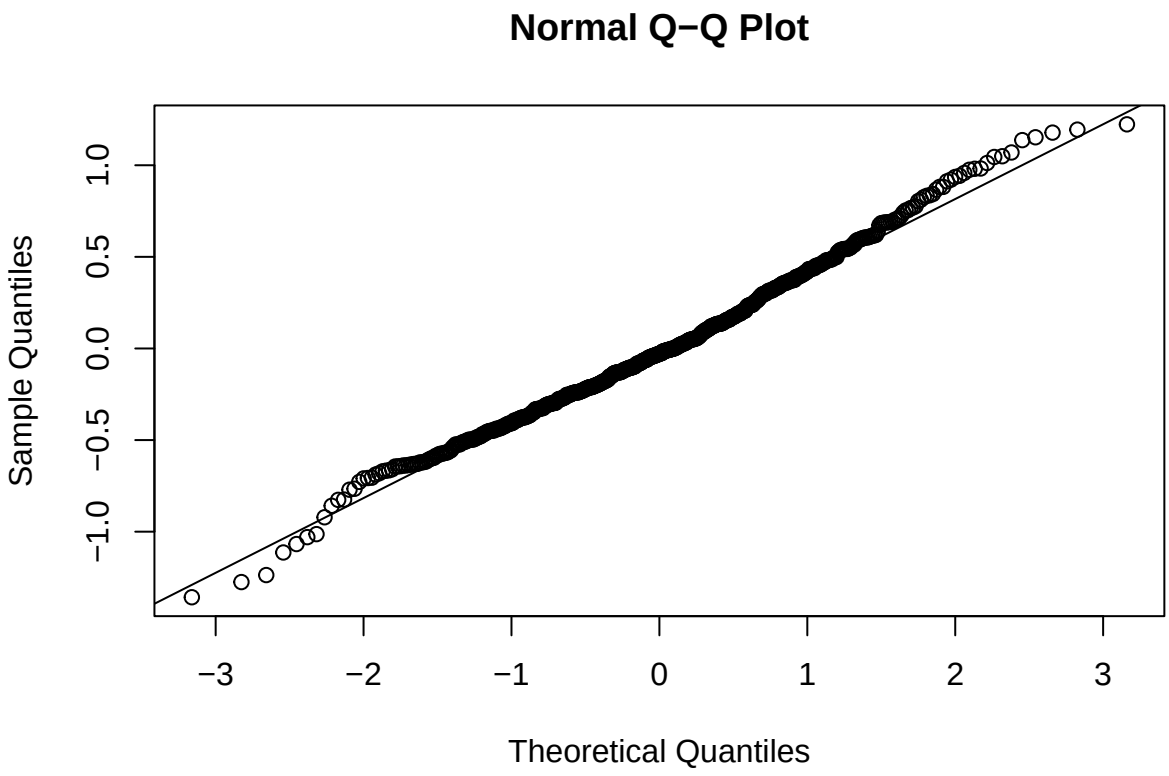
```
## [1] 8.714062e-18
```

```
median(res2_df$residuals)
```

```
## [1] -0.02864612
```

The OLS Model 2 residuals' mean > median. (Right-Skewed).

```
qqnorm(resid(ols_model2))
qqline(resid(ols_model2))
```




```
# Shapiro-Wilk normality test
shapiro.test(resid(ols_model2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(ols_model2)
## W = 0.99231, p-value = 0.002285
```

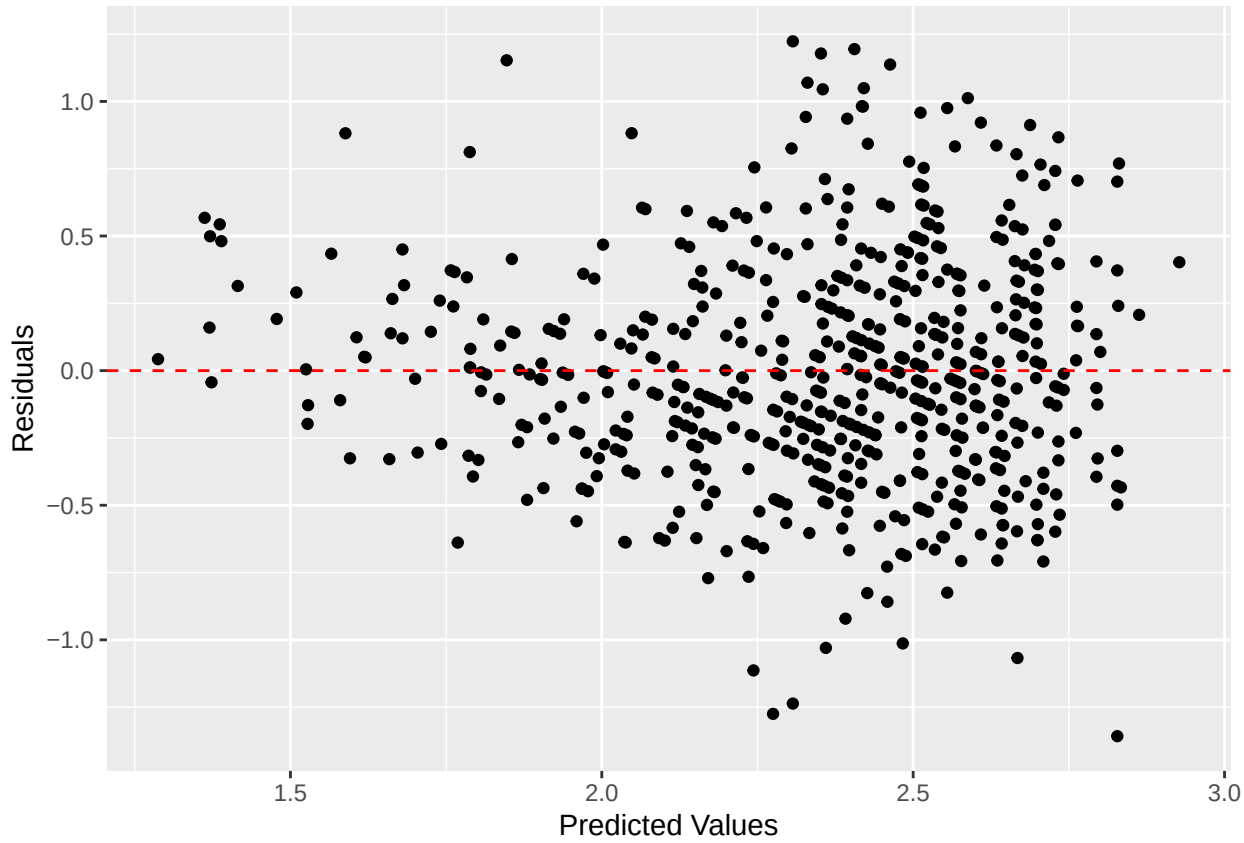
```
# Ho: Sample data comes from a normal distribution
```

(P-value $= 0.002285$) $<$ ($\alpha = 0.05$) $\Rightarrow$. Reject the null because the Shaprio-Wilk's test confirms a statistically significant difference from normality. The density visual suggests a right-skewed distribution.

Constant Variance Check

Fitted vs Residual

```
ggplot(data = ols_model2, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 1) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted Values", y = "Residuals")
```



Variance(spread) of residuals does not appear constant.