SGupta_HW04Question2

Quarto

model = 0 + 1(expend) + 2(salary) + 3(ratio) + 4(takers) + e Linear Regression Model

- : Intercept (baseline total SAT score)
- : Effect of expenditure per student (expend) on total SAT
- : Effect of teacher salary on total SAT
- : Effect of student-teacher ratio on total SAT
- : Effect of the proportion of students taking the test (takers)
- e: Error term capturing variability not explained by the predictors

```r
# Load necessary libraries
library(faraway)

set.seed(42)
data(sat)
head(sat)
```

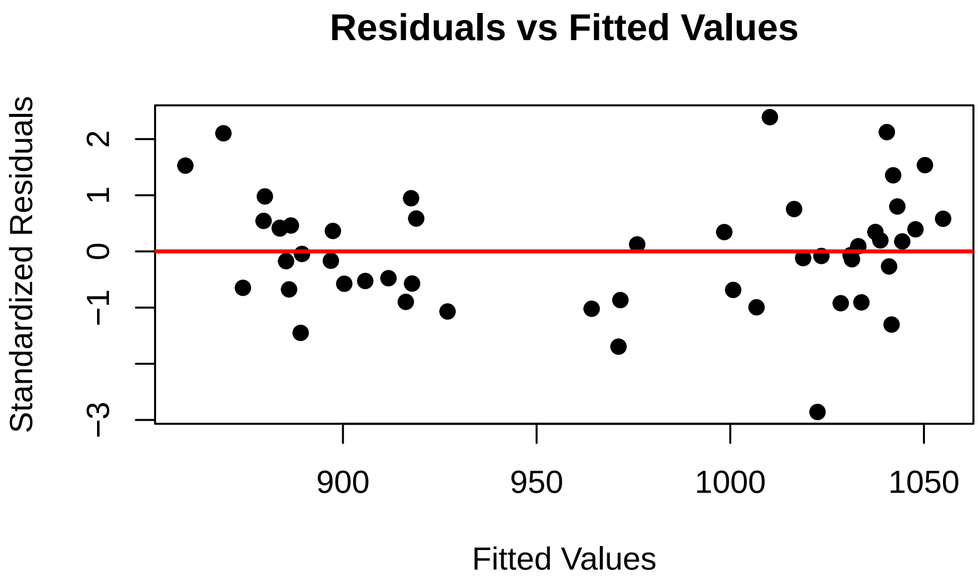
```
           expend ratio salary takers verbal math total
Alabama     4.405  17.2 31.144      8    491  538  1029
Alaska      8.963  17.6 47.951     47    445  489   934
Arizona     4.778  19.3 32.175     27    448  496   944
Arkansas    4.459  17.1 28.934      6    482  523  1005
California  4.992  24.0 41.078     45    417  485   902
Colorado    5.443  18.4 34.571     29    462  518   980
```

```r
# Fit the model
sat_lm <- lm(total ~ expend + salary + ratio + takers, data = sat)

# Print summary
summary(sat_lm)
```

```
Call:
lm(formula = total ~ expend + salary + ratio + takers, data = sat)

Residuals:
    Min      1Q  Median      3Q     Max
-90.531 -20.855  -1.746  15.979  66.571

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
expend         4.4626    10.5465   0.423    0.674
salary         1.6379     2.3872   0.686    0.496
ratio         -3.6242     3.2154  -1.127    0.266
takers        -2.9045     0.2313 -12.559 2.61e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.7 on 45 degrees of freedom
Multiple R-squared:  0.8246,    Adjusted R-squared:  0.809
F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

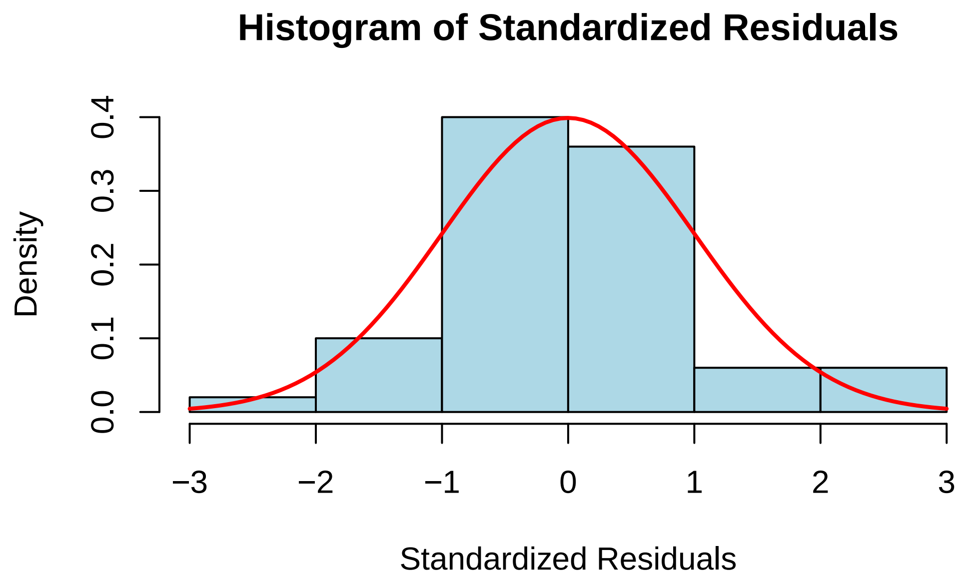
(a) Check the constant variance assumption for the errors.

```r
# (a) Check constant variance
# Standardized residuals and fitted values
res_std <- rstandard(sat_lm)
fitted_vals <- fitted(sat_lm)

# Plot: Standardized Residuals vs Fitted Values
par(mfrow=c(1,1))
plot(fitted_vals, res_std,
     xlab = "Fitted Values",
     ylab = "Standardized Residuals",
     main = "Residuals vs Fitted Values",
     pch = 19)
abline(h = 0, col = "red", lwd = 2)
```

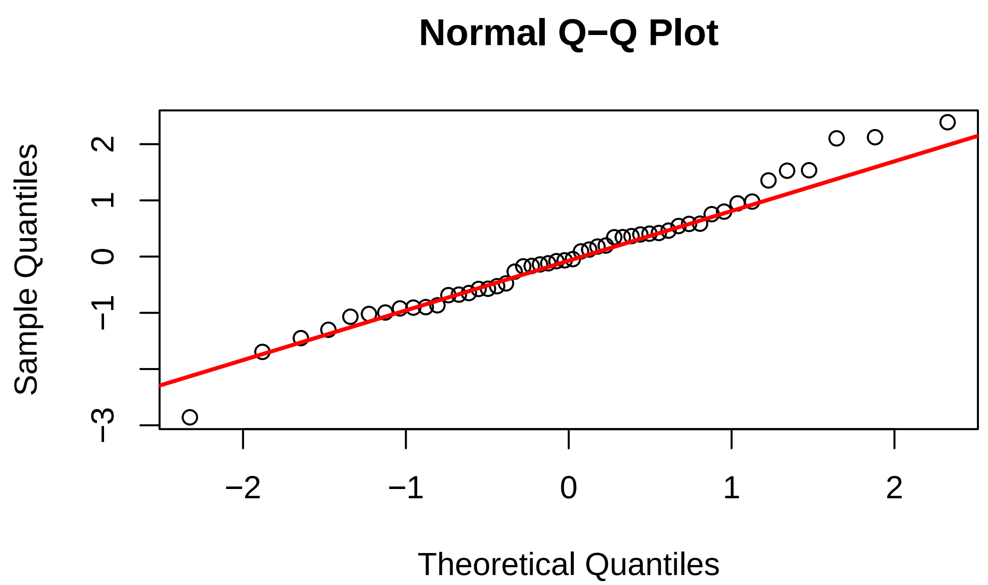


(b) Check the normality assumption.

```
# (b) Check normality
# Histogram of standardized residuals with normal curve
hist(res_std, probability = TRUE,
     main = "Histogram of Standardized Residuals",
     xlab = "Standardized Residuals", col = "lightblue", border = "black")
curve(dnorm(x), add = TRUE, col = "red", lwd = 2)
```




```
qqnorm(res_std, main = "Normal Q-Q Plot")
qqline(res_std, col = "red", lwd = 2)
```

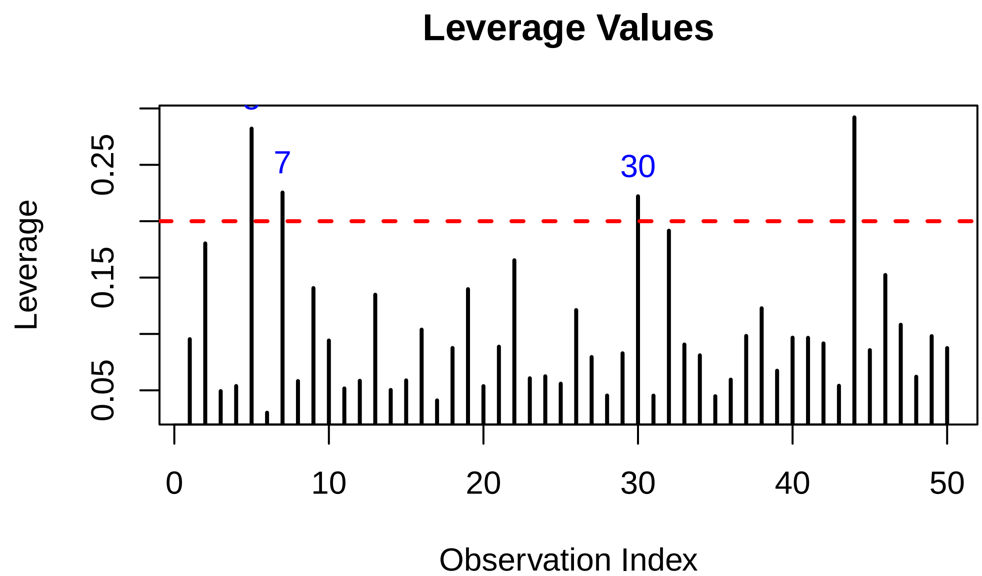


```r
shapiro.test(res_std)
```

```
	Shapiro-Wilk normality test

data:  res_std
W = 0.98021, p-value = 0.5607
```

(c) Check for large leverage points.

```r
# (c) Check for large leverage points
# Calculate leverage values
lev <- hatvalues(sat_lm)

# Plot leverage values
plot(lev, type = "h",
     main = "Leverage Values",
     xlab = "Observation Index",
     ylab = "Leverage", lwd = 2)
abline(h = 2*mean(lev), col = "red", lwd = 2, lty = 2)
text(x = which(lev > 2*mean(lev)),
     y = lev[lev > 2*mean(lev)],
     labels = which(lev > 2*mean(lev)), pos = 3, col = "blue")
```

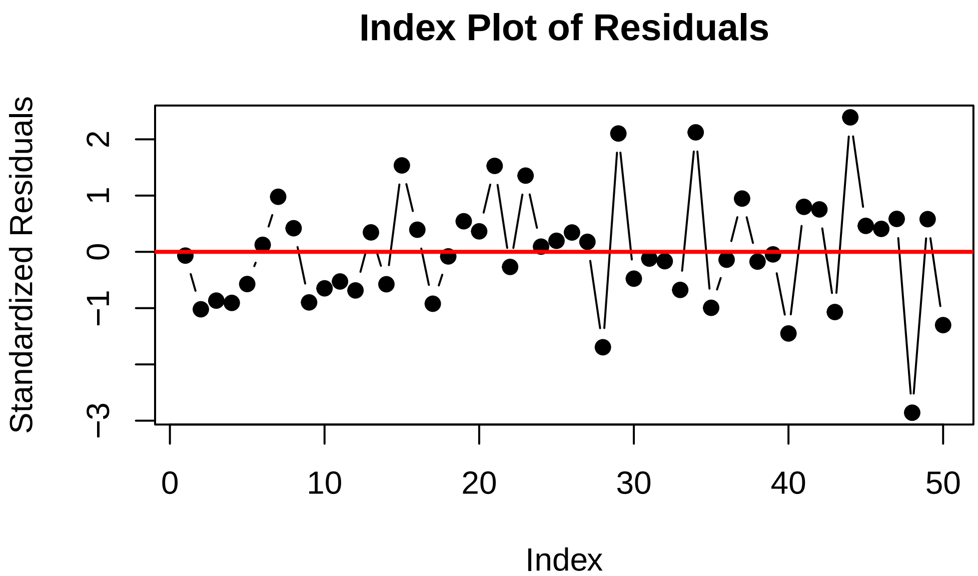


(d) Check for serial correlation in the errors.

```
library(lmtest)
```

```
Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```
# (d) Check for serial correlation
# Index plot of residuals
plot(res_std, type = "b",
     main = "Index Plot of Residuals",
     xlab = "Index",
     ylab = "Standardized Residuals", pch = 19)
abline(h = 0, col = "red", lwd = 2)
```




```
# Load lmtest package for Durbin-Watson test
dwtest(sat_lm)
```

```
	Durbin-Watson test

data:  sat_lm
DW = 2.4525, p-value = 0.9459
alternative hypothesis: true autocorrelation is greater than 0
```

Summary

- **Model Fit:**
  - The estimated regression equation is:
    $\text{total} = 1045.97 + 4.46 \cdot \text{expend} + 1.64 \cdot \text{salary} - 3.62 \cdot \text{ratio} - 2.90 \cdot \text{takers} + \text{e}$
  - Only the coefficient for takers is highly significant ($p < 2.6e\text{-}16$), while the other predictors are not statistically significant.
  - The model explains about 82.5% of the variability in total SAT scores (R^2 = 0.8246).
- **Constant Variance:**
  - The plot of standardized residuals versus fitted values shows no obvious pattern or funneling, which supports the homoscedasticity (constant variance) assumption.

- **Normality:**
  – The histogram of standardized residuals, overlaid with a normal density curve, and the corresponding Q–Q plot both indicate that the residuals are approximately normally distributed.
  – The Shapiro–Wilk test (W = 0.98021, p = 0.5607) confirms that there is no strong evidence against normality.
- **Leverage:**
  – A leverage plot was used to flag observations with unusually high leverage (above twice the average leverage).
  – Observations (indices 5, 7, 30, and 44) are identified as having high leverage, suggesting these cases should be reviewed further for their influence on the model.
- **Serial Correlation:**
  – An index plot of residuals shows no systematic pattern over the order of observations.
  – The Durbin–Watson test yields a statistic of 2.4525 with a p-value of 0.9459, indicating no significant autocorrelation in the residuals.