HW8_1700

Yue Zhang

2024-11-15

#11.89 Null Hypothesis: there’s no relationship between time period score and annual incidents of diabetes ($\beta = 0$) Alternative Hypothesis: there’s a linear relationship between time period score and annual incidents of diabetes ($\beta \neq 0$)

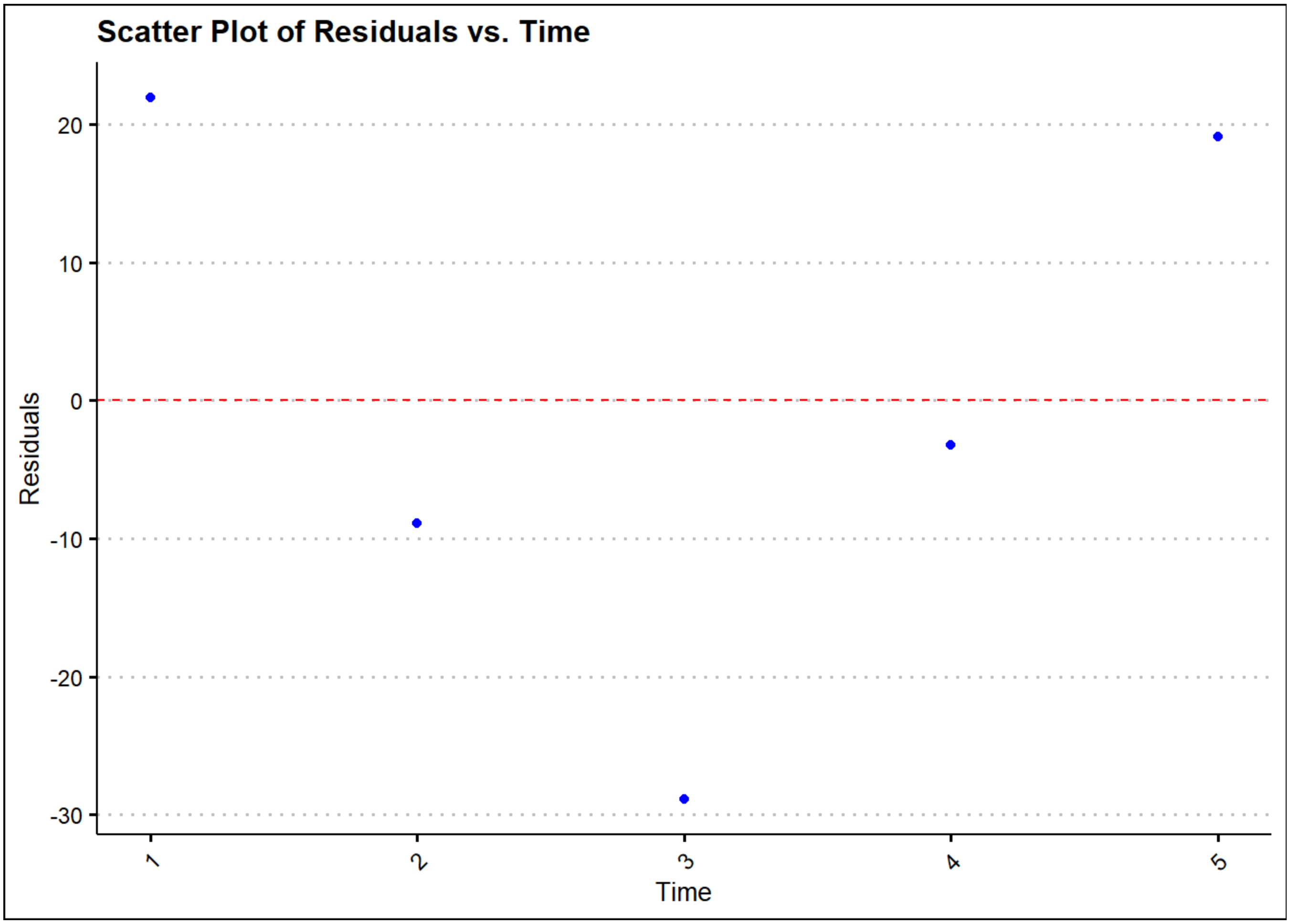
```
diabetes = data.frame(time = 1:5, annual_incidents = c(240.4,
    243.1, 256.7, 315.9, 371.8))

diabetes_model = lm(annual_incidents ~ time, data = diabetes)
summary(diabetes_model)
```

```
##
## Call:
## lm(formula = annual_incidents ~ time, data = diabetes)
##
## Residuals:
##      1      2      3      4      5
##  21.94  -8.92 -28.88  -3.24  19.10
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  184.900     25.478   7.257   0.0054 **
## time          33.560      7.682   4.369   0.0222 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.29 on 3 degrees of freedom
## Multiple R-squared:  0.8642, Adjusted R-squared:  0.8189
## F-statistic: 19.09 on 1 and 3 DF,  p-value: 0.02218
```

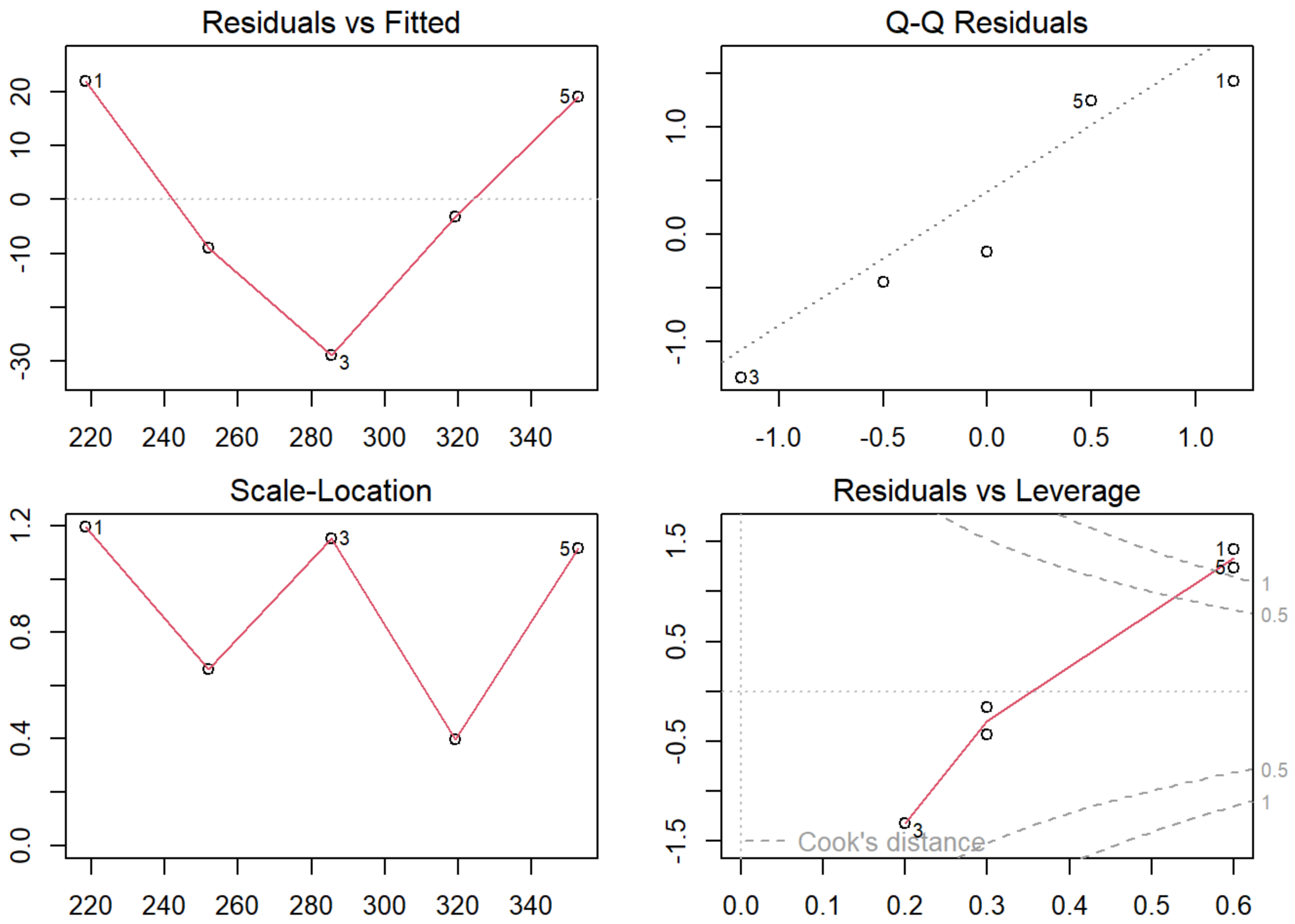
The prediction equation is $y = 184.9 + 33.56 * x$
Since the p-value of the coefficient is 0.022 < 0.05, we can reject the null hypothesis, indicating a statistically significant relationship between time period score and annual incidents of diabetes.

#11.90

```
# Scatter plot of residuals vs. Time
ggplot(diabetes, aes(x = time, y = resid(diabetes_model))) +
    geom_point(color = "blue") + geom_hline(yintercept = 0, color = "red",
    linetype = "dashed") + labs(title = "Scatter Plot of Residuals vs. Time",
    x = "Time", y = "Residuals")
```



```
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(diabetes_model, cex.axis = 1, cex.lab = 1)
```



Scatter plot of residuals vs. Time: The “V” shape of the residuals suggests the heteroscedasticity of the residuals. Also, it indicates a violation of the independence assumption. Residuals vs. Fitted Plot: The curved pattern indicates that it violates the assumption of homoscedasticity as residuals should be randomly scattered around zero. Q-Q Plot: The residuals deviate slightly from the line, suggesting mild departures from normality. Scale-Location Plot: The systematic pattern shows heteroscedasticity. Residuals vs. Leverage Plot: Observations 1, 5, and 3 show higher leverage and Cook’s distance, indicating they may be influential points. These observations could disproportionately impact the model’s fit. Overall, the normality, linearity, independence, and homoscedastic assumptions are all violated meaning linear model is not a good fit.

#11.95

```
weight_change = c(5, 3.8, 5.7, 4.5, 3.3, 6.4, 0.9, 0.6, -0.2,
    3.2, 5.6, 4.3, 6, 7.2, 7.9)
hbg_change = c(-1.5, -2.1, -0.8, 0.7, -1.9, -0.8, 0.4, 0.6, 1.8,
    0.8, 0, 0.5, 0.3, -0.8, -1.9)
spearman_result = cor.test(weight_change, hbg_change, method = "spearman",
    conf.level = 0.95)
print(spearman_result)
```

```
##
##  Spearman's rank correlation rho
##
## data:  weight_change and hbg_change
## S = 853.31, p-value = 0.04508
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.5237721
```

```
# sample ranks for weight and hbg
weight_rank = rank(weight_change)
hbg_rank = rank(hbg_change)

# point estimate of the rank correlation
a = cor.test(weight_rank, hbg_rank)
r_s = a$estimate
n = length(weight_change)

# probits for weight and hbg
weight_hat = qnorm(weight_rank/(n + 1))
hbg_hat = qnorm(hbg_rank/(n + 1))

# Pearson correlation between sample probits
r_h = cor(weight_hat, hbg_hat)

# bias-corrected estimate
rcor_h = r_h * (1 + (1 - r_h^2)/(2 * (n - 4)))

# Fisher's z transform
z_h = 0.5 * log((1 + rcor_h)/(1 - rcor_h))

# 95%CI associated with z_h
z1_h = z_h - 1.96/sqrt(n - 3)
z2_h = z_h + 1.96/sqrt(n - 3)

# 95%CI for r_h
r1_h = (exp(2 * z1_h) - 1)/(exp(2 * z1_h) + 1)
r2_h = (exp(2 * z2_h) - 1)/(exp(2 * z2_h) + 1)

# 95%CI for r_s
r1_s = (6/pi) * asin(r1_h/2)
r2_s = (6/pi) * asin(r2_h/2)

cat("95% Confidence Interval for Spearman's Rank Correlation:",
    r1_s, "to", r2_s, "\n")
```

```
## 95% Confidence Interval for Spearman's Rank Correlation: -0.8179395 to -0.0557686
```

Null hypothesis:there’s no relationship between HgbA1c and weight Alternative hypothesis: there’s a relationship between HgbA1c and weight The two-sided p-value is 0.045 < 0.05, therefore, we can reject the null hypothesis indicating that HgbA1c and weight are related. The 95% CI for the Spearman rank correlation is (-0.818, -0.056)

#Required Additional Problem ##(a)

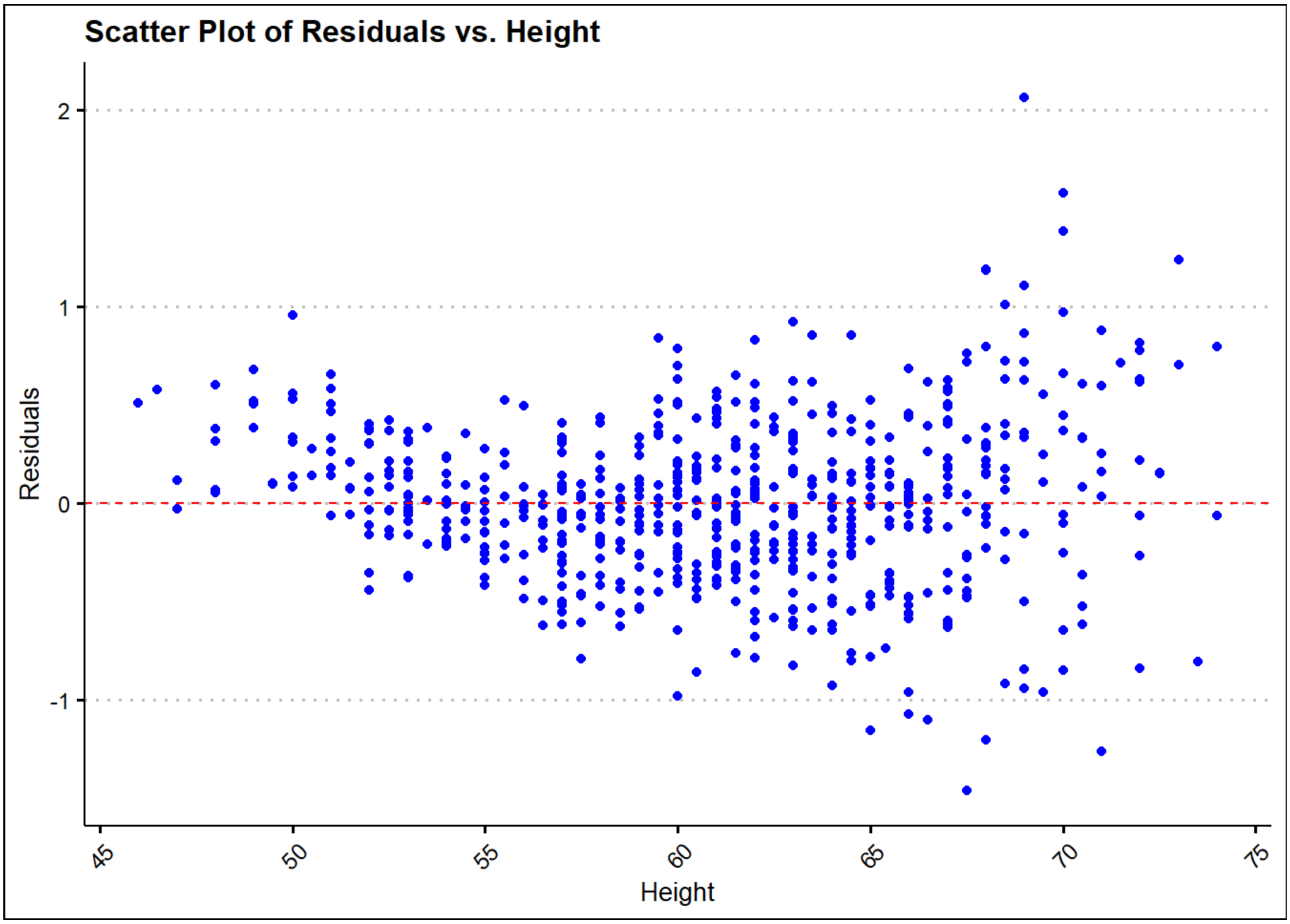
```
fev = read_dta("E:/Biostat/Biostatistics/PHL_1700/Data/Raw/FEV-1.DAT.dta")
# I assign scores for each 5-year category.  Score = 1:
# Ages between 0 and 5 Score = 2: Ages between 6 and 10
# Score = 3: Ages between 11 and 15 Score = 4: Ages between
# 16 and 19
fev$Age_Score = cut(fev$Age, breaks = seq(0, max(fev$Age) + 5,
    by = 5), labels = FALSE, right = TRUE, include.lowest = TRUE)
```

##(b)

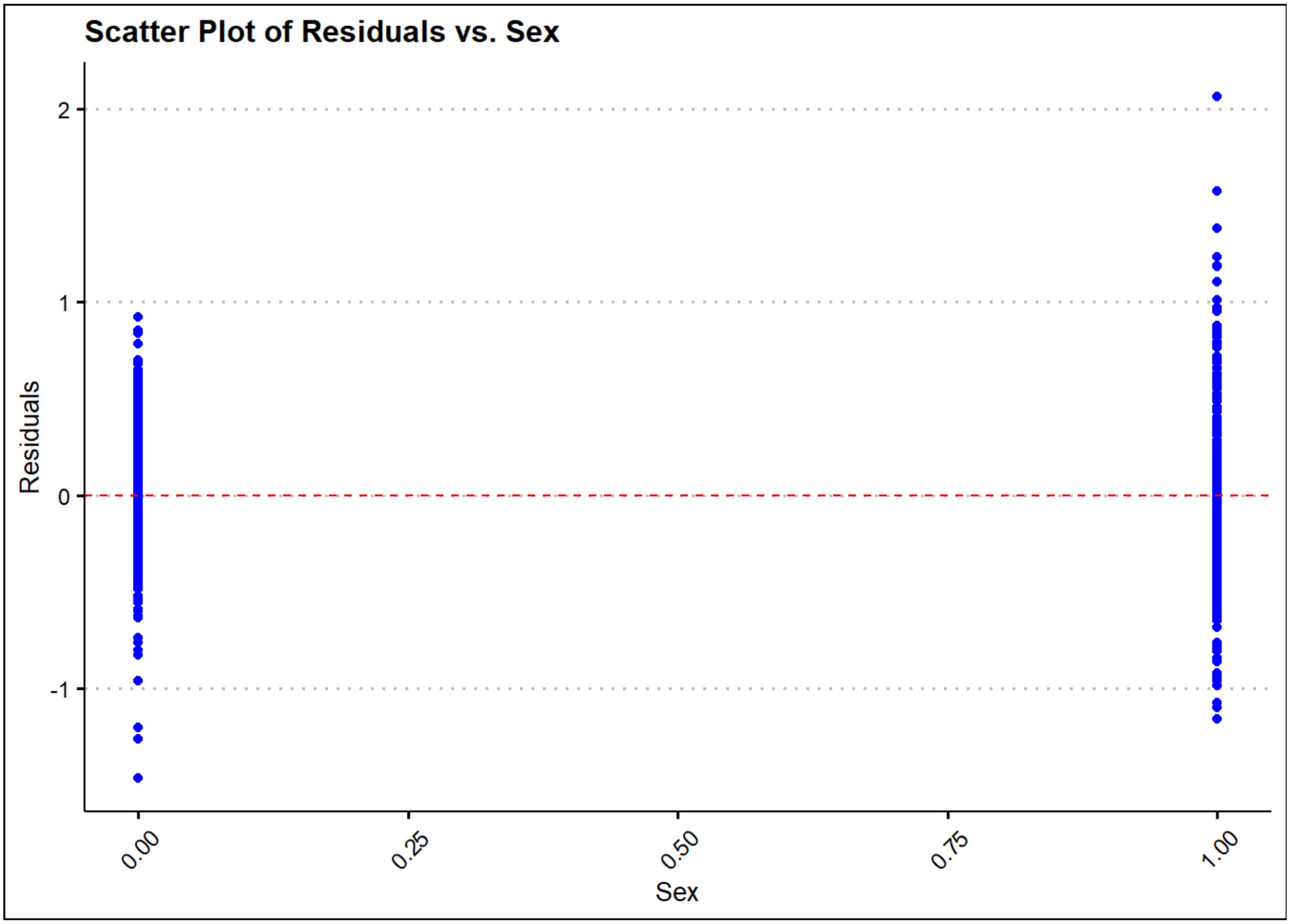
```
# (b)
fev_model = lm(fev ~ Hgt + Sex + Smoke + Age_Score, data = fev)
summary(fev_model)
```

```
##
## Call:
## lm(formula = fev ~ Hgt + Sex + Smoke + Age_Score, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46378 -0.25518 -0.00672  0.24195  2.06485
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.801909   0.203097 -23.643  < 2e-16 ***
## Hgt          0.111802   0.004196  26.645  < 2e-16 ***
## Sex          0.151143   0.033388   4.527 7.12e-06 ***
## Smoke       -0.043888   0.058389  -0.752    0.453
## Age_Score    0.221519   0.036048   6.145 1.39e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4152 on 649 degrees of freedom
## Multiple R-squared:  0.7721, Adjusted R-squared:  0.7707
## F-statistic: 549.8 on 4 and 649 DF,  p-value: < 2.2e-16
```

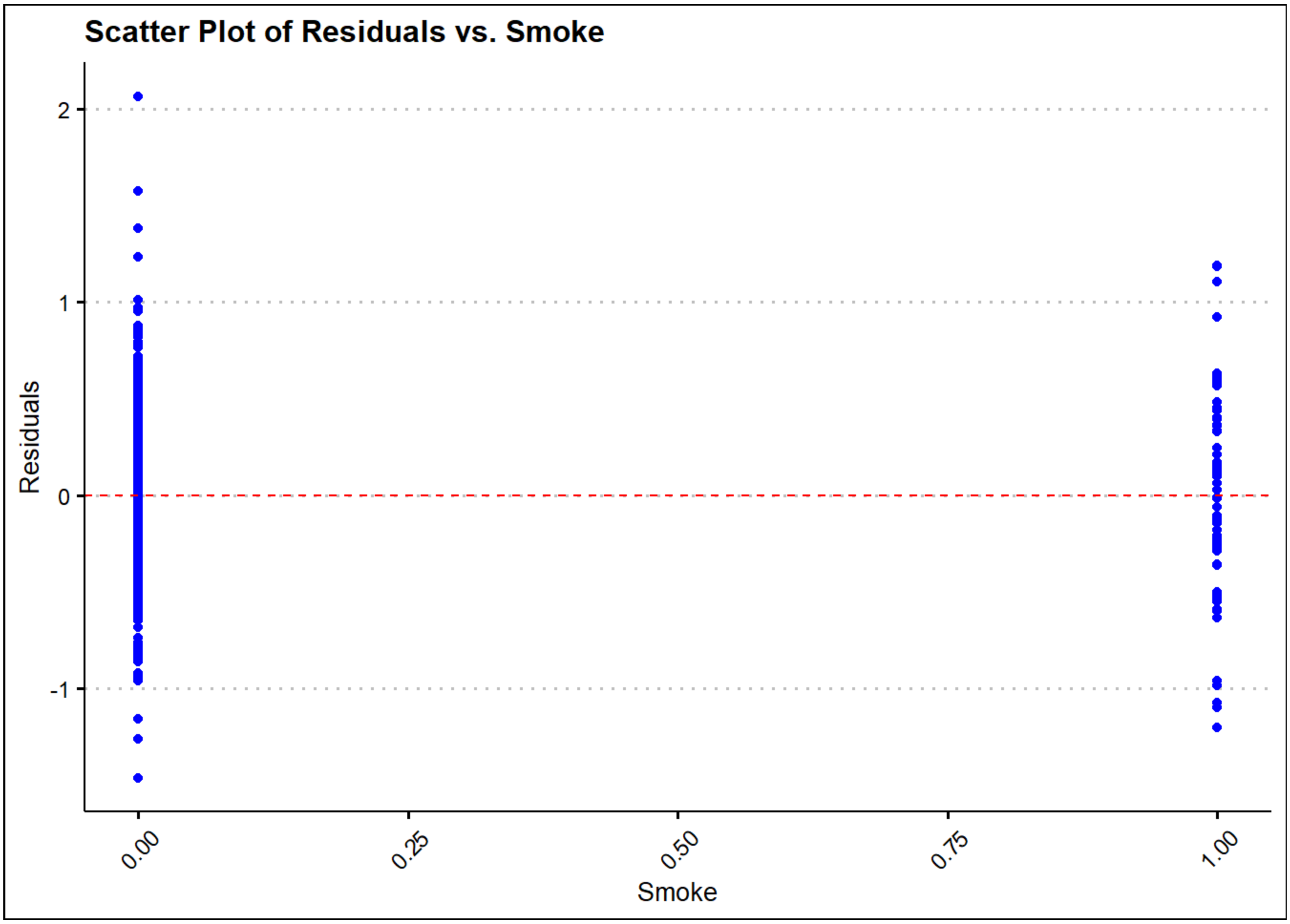
```
# Scatter plot of residuals vs. Hgt
ggplot(fev, aes(x = Hgt, y = resid(fev_model))) + geom_point(color = "blue") +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
    labs(title = "Scatter Plot of Residuals vs. Height", x = "Height",
        y = "Residuals")
```



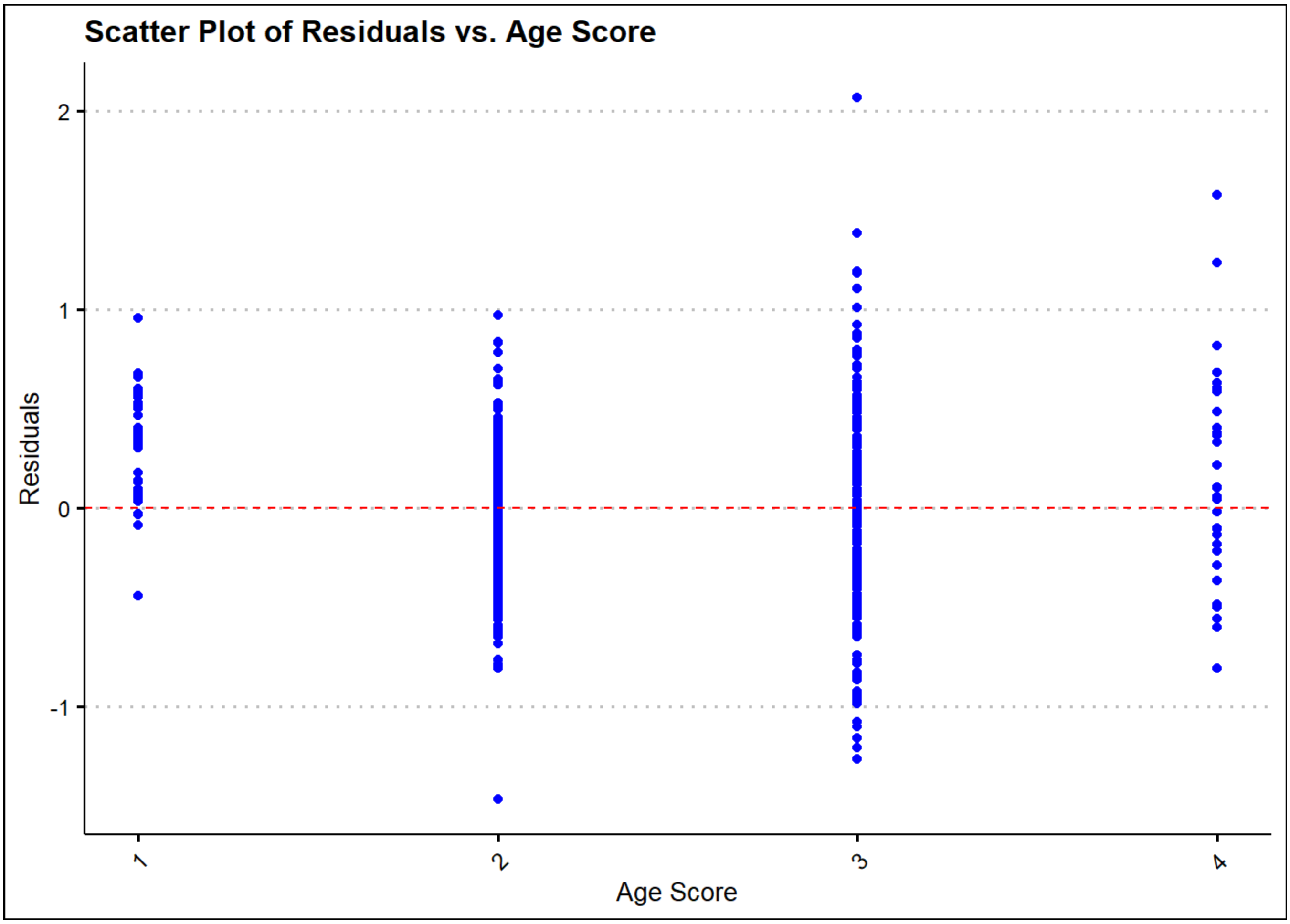
```
# Scatter plot of residuals vs. Sex
ggplot(fev, aes(x = Sex, y = resid(fev_model))) + geom_point(color = "blue") +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
    labs(title = "Scatter Plot of Residuals vs. Sex", x = "Sex",
        y = "Residuals")
```



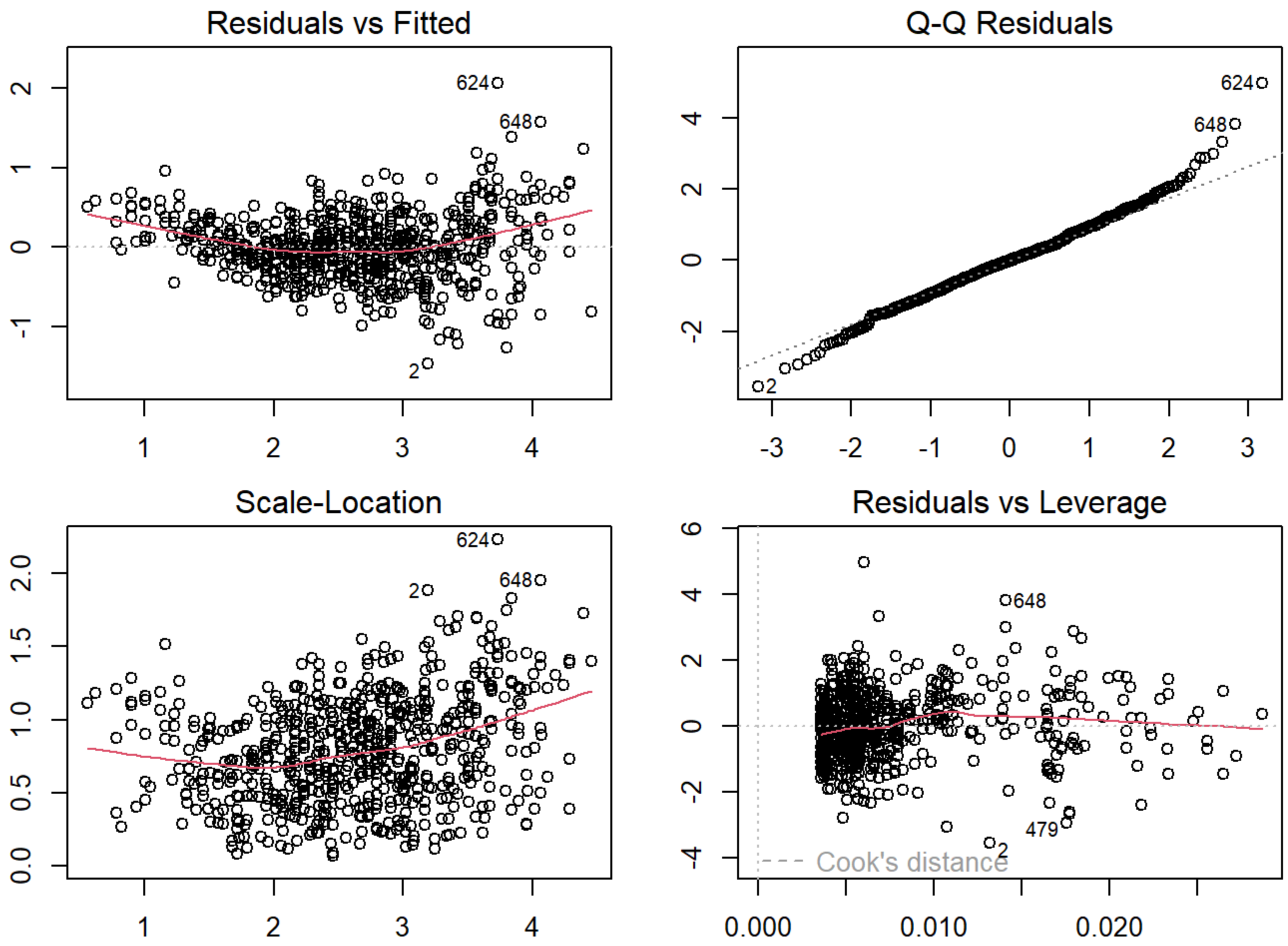
```
# Scatter plot of residuals vs. Smoke
ggplot(fev, aes(x = Smoke, y = resid(fev_model))) + geom_point(color = "blue") +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
    labs(title = "Scatter Plot of Residuals vs. Smoke", x = "Smoke",
        y = "Residuals")
```



```
# Scatter plot of residuals vs. Age Score
ggplot(fev, aes(x = Age_Score, y = resid(fev_model))) + geom_point(color = "blue") +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
    labs(title = "Scatter Plot of Residuals vs. Age Score", x = "Age Score",
        y = "Residuals")
```



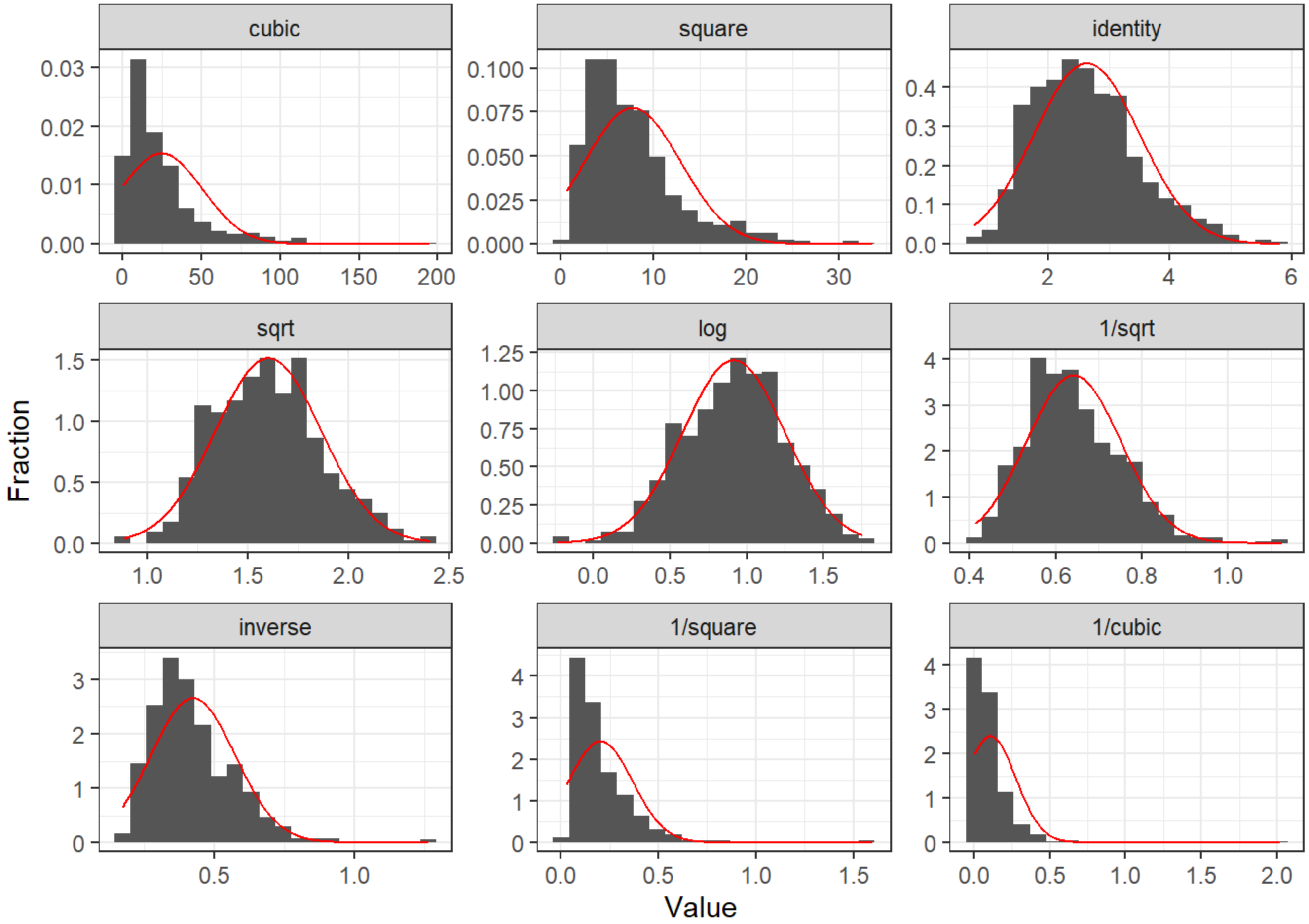
```
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(fev_model, cex.axis = 1, cex.lab = 1)
```



Scatter plot of residuals vs. Independent Variables: In the scatter plot of residuals vs. height, there seems to be more variability at higher heights, suggesting the linear model might not be capturing the relationship perfectly. Also, the residuals increase as height increases, indicating a potential heteroscedasticity and dependence issue. Residuals vs. Fitted Plot: The curved pattern indicates that it violates the assumption of homoscedasticity and linearity as residuals should be randomly scattered around zero. Q-Q Plot: The residuals deviate slightly from the line, suggesting mild departures from normality. Scale-Location Plot: The systematic pattern shows heteroscedasticity. Residuals vs. Leverage Plot: Observations 648, 2, and 479 show higher leverage and Cook’s distance, indicating they may be influential points. These observations could disproportionately impact the model’s fit. Overall, the linearity, independence, and homoscedastic assumptions are all violated meaning linear model is not a good fit.

##(c)

```
gladder(fev$fev)
```



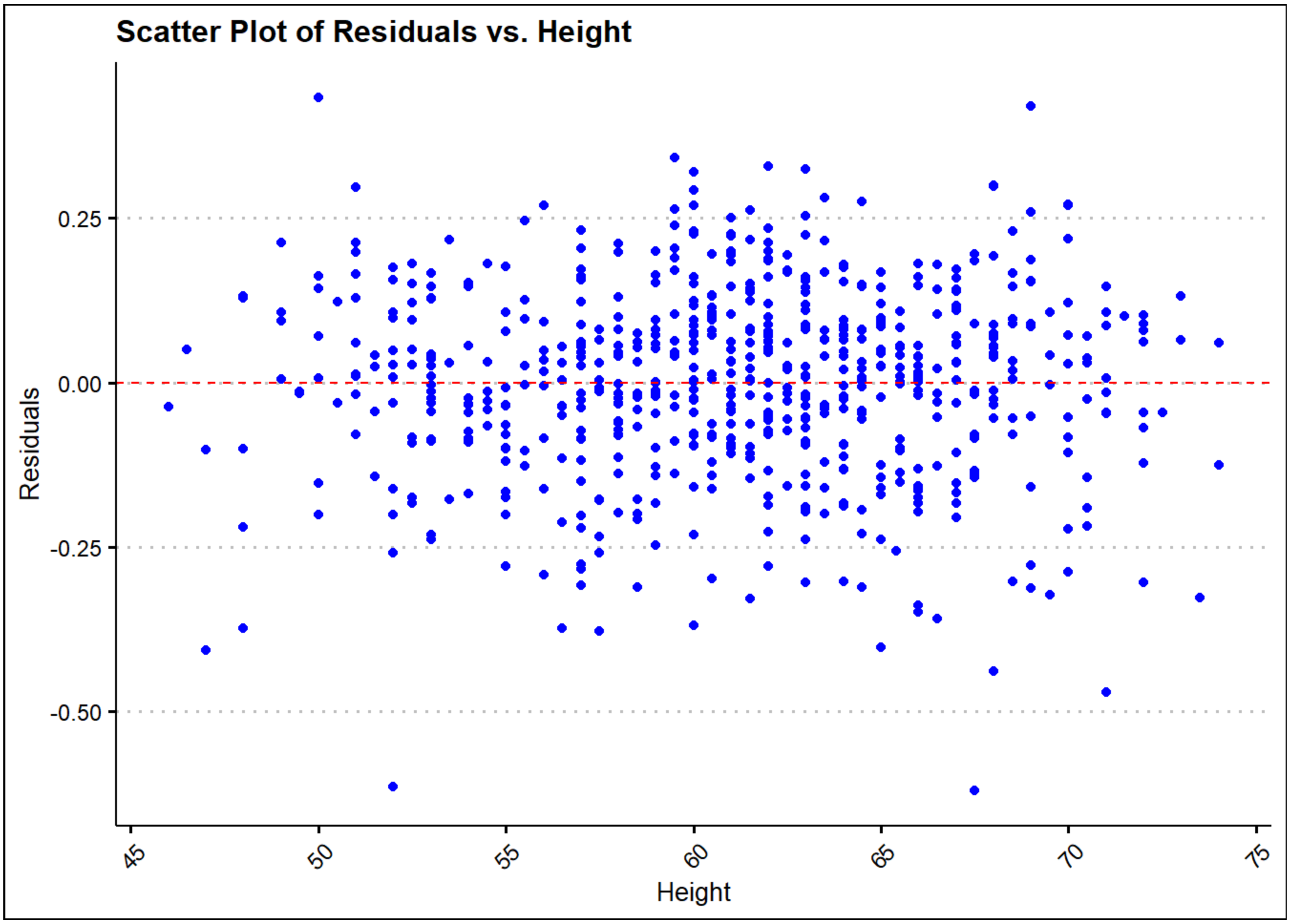
Based on the plots, the log transformation would be the best choice. The histogram for the log-transformed variable is roughly symmetric, and it approximates a normal distribution well. While other transformation would be skewed and not normally distributed.

##(d)

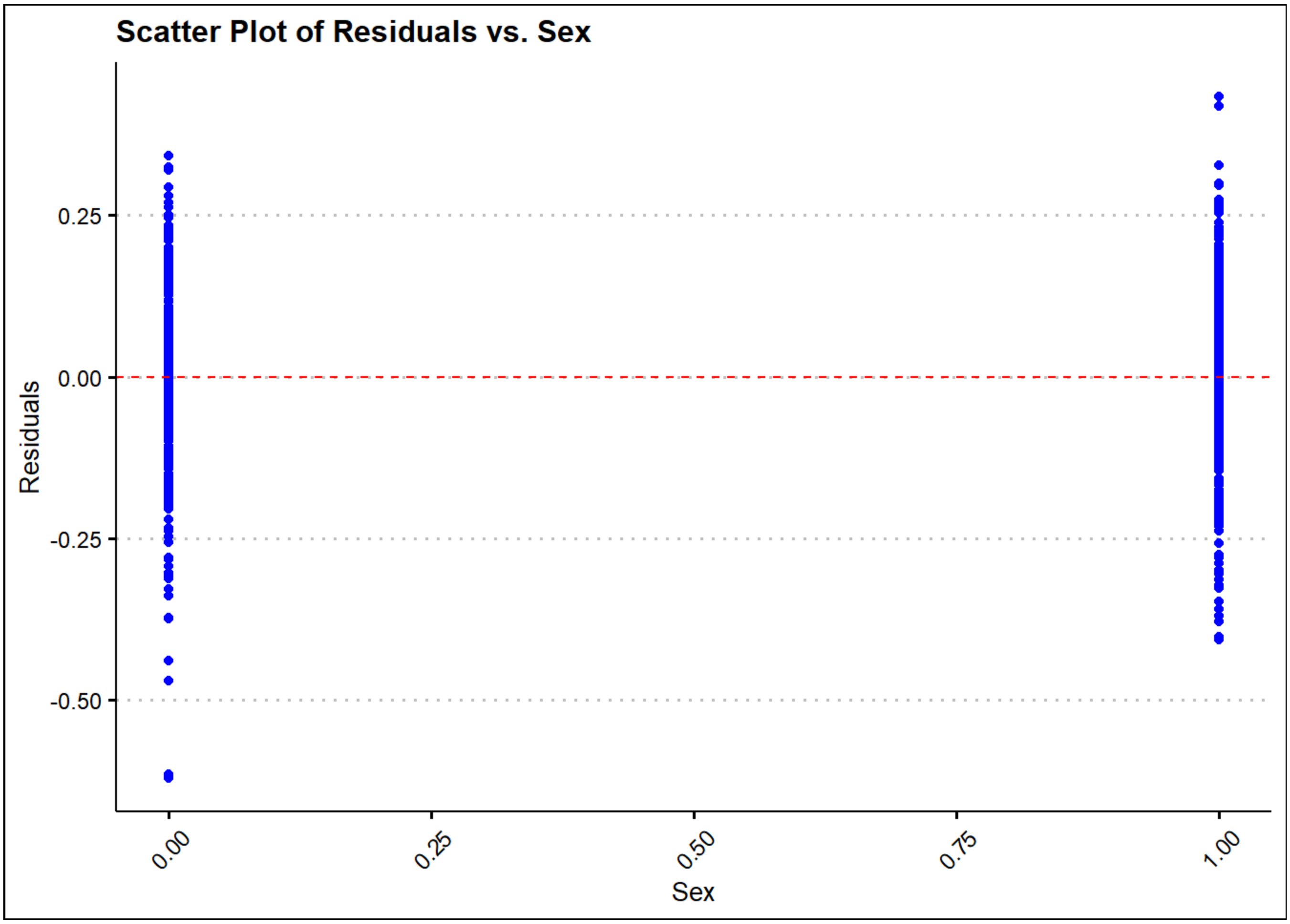
```
fev_model2 = lm(log(fev) ~ Hgt + Sex + Smoke + Age_Score, data = fev)
summary(fev_model2)
```

```
##
## Call:
## lm(formula = log(fev) ~ Hgt + Sex + Smoke + Age_Score, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62043 -0.08530  0.00721  0.09508  0.43349
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.072952   0.071806 -28.869  < 2e-16 ***
## Hgt          0.045696   0.001484  30.802  < 2e-16 ***
## Sex          0.026948   0.011804   2.283   0.0228 *
## Smoke       -0.029779   0.020644  -1.443   0.1496
## Age_Score    0.076786   0.012745   6.025 2.84e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1468 on 649 degrees of freedom
## Multiple R-squared:  0.8072, Adjusted R-squared:  0.806
## F-statistic: 679.3 on 4 and 649 DF,  p-value: < 2.2e-16
```

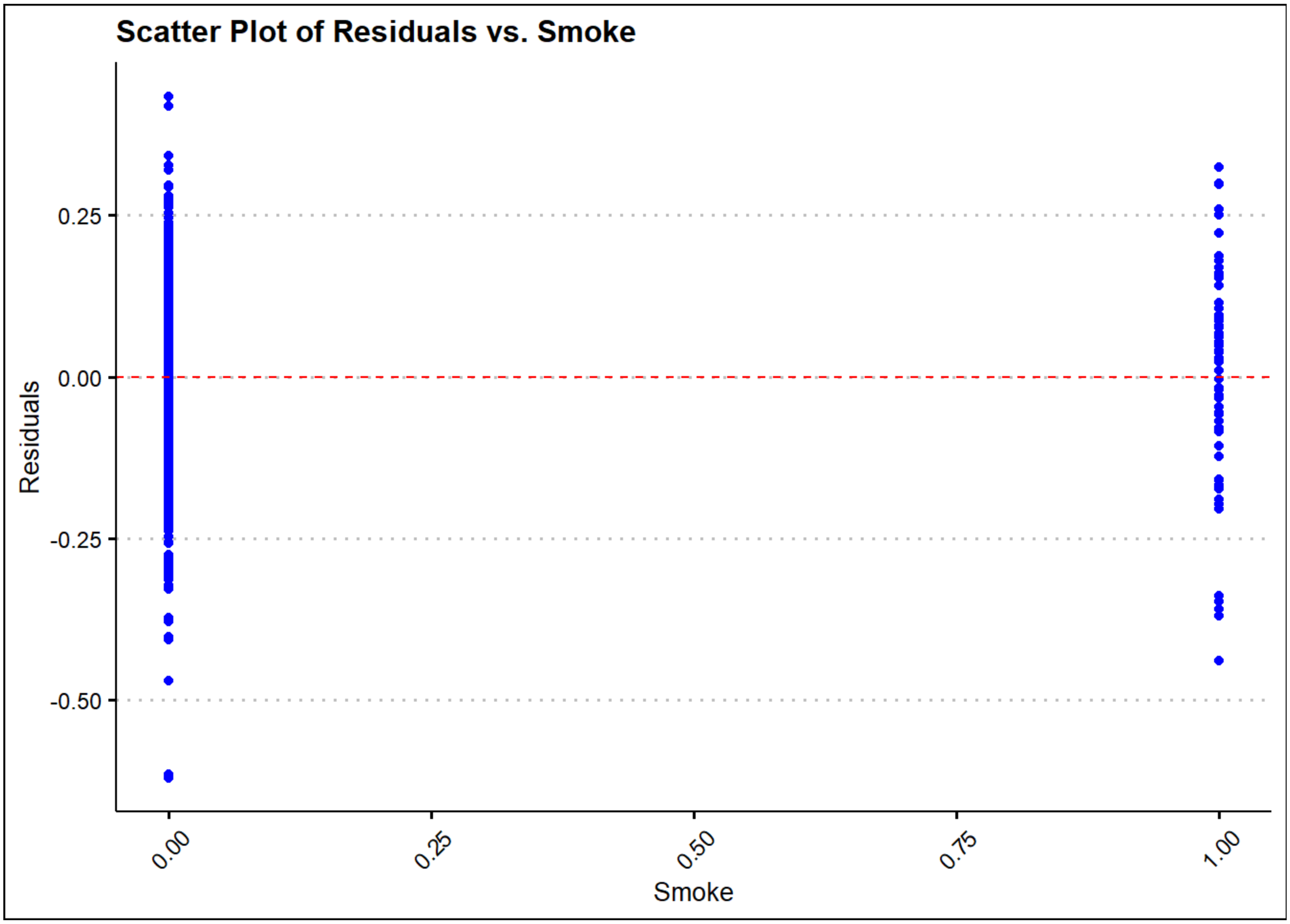
```
# Scatter plot of residuals vs. Hgt
ggplot(fev, aes(x = Hgt, y = resid(fev_model2))) + geom_point(color = "blue") +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
    labs(title = "Scatter Plot of Residuals vs. Height", x = "Height",
        y = "Residuals")
```



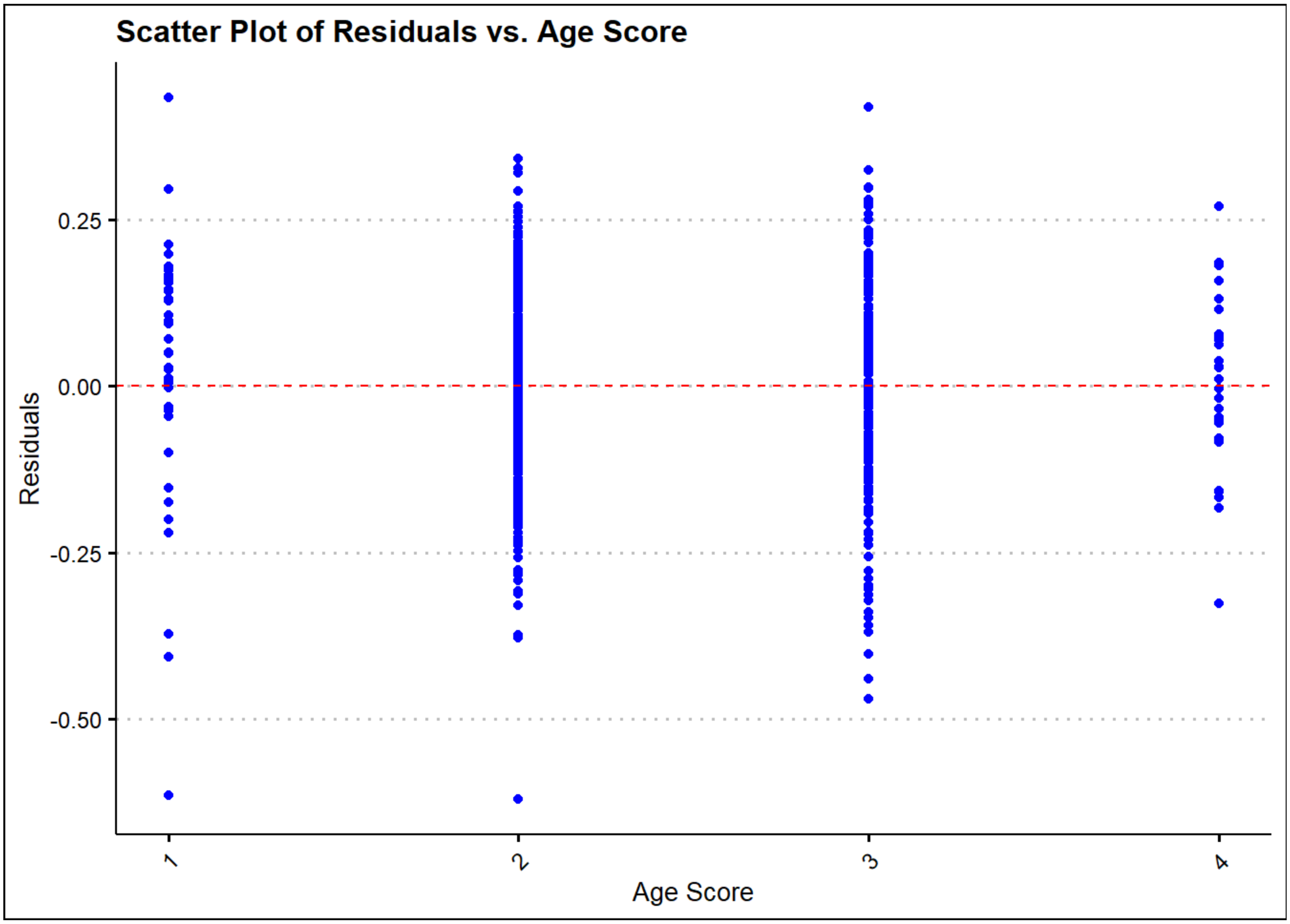
```
# Scatter plot of residuals vs. Sex
ggplot(fev, aes(x = Sex, y = resid(fev_model2))) + geom_point(color = "blue") +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
    labs(title = "Scatter Plot of Residuals vs. Sex", x = "Sex",
        y = "Residuals")
```



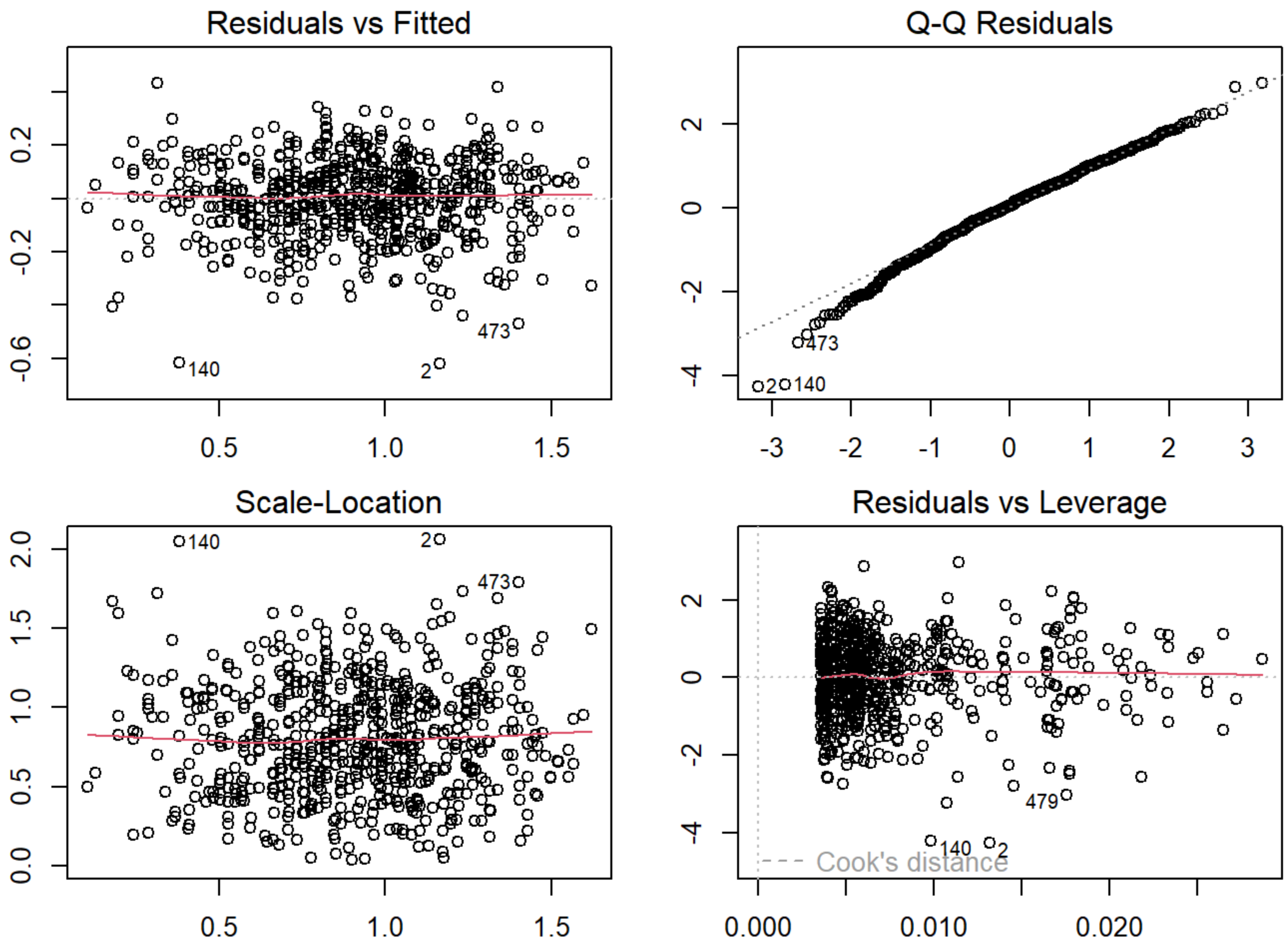
```
# Scatter plot of residuals vs. Smoke
ggplot(fev, aes(x = Smoke, y = resid(fev_model2))) + geom_point(color = "blue") +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
    labs(title = "Scatter Plot of Residuals vs. Smoke", x = "Smoke",
        y = "Residuals")
```



```
# Scatter plot of residuals vs. Age Score
ggplot(fev, aes(x = Age_Score, y = resid(fev_model2))) + geom_point(color = "blue") +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
    labs(title = "Scatter Plot of Residuals vs. Age Score", x = "Age Score",
        y = "Residuals")
```



```
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(fev_model2, cex.axis = 1, cex.lab = 1)
```



Scatter plot of residuals vs. Independent Variables: There’s no indication of dependence between residuals and independent variables. The residuals are randomly scattered around zero, suggesting the assumptions of homoscedasticity and linearity are satisfied. Residuals vs. Fitted Plot: There’s no indication of non-linearity and heteroscedasticity. Q-Q Plot: The residuals deviate slightly from the line at the lower tail, suggesting minor departures from normality. Scale-Location Plot: The systematic pattern shows homoscedasticity. Residuals vs. Leverage Plot: Observations 140, 2, and 479 show higher leverage and Cook’s distance, indicating they may be influential points. These observations could disproportionately impact the model’s fit. Overall, all assumptions are satisfied, indicating the log-transformation model is much better than the linear model.

##(e)

```
summary(fev_model2)
```

```
##
## Call:
## lm(formula = log(fev) ~ Hgt + Sex + Smoke + Age_Score, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62043 -0.08530  0.00721  0.09508  0.43349
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.072952   0.071806 -28.869  < 2e-16 ***
## Hgt          0.045696   0.001484  30.802  < 2e-16 ***
## Sex          0.026948   0.011804   2.283   0.0228 *
## Smoke       -0.029779   0.020644  -1.443   0.1496
## Age_Score    0.076786   0.012745   6.025 2.84e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1468 on 649 degrees of freedom
## Multiple R-squared:  0.8072, Adjusted R-squared:  0.806
## F-statistic: 679.3 on 4 and 649 DF,  p-value: < 2.2e-16
```

###(i) The p-value is less than 0.05, indicating the model is statistically significant.

###(ii) The adjusted R-squared is 0.806, indicating that approximately 80.6% of the variation is explained by the model.

###(iii)

```
Smoke_CIL = -0.029779 - 1.96 * 0.020644
Smoke_CIU = -0.029779 + 1.96 * 0.020644
cat("95% Confidence Interval for smoking:", Smoke_CIL, "to",
    Smoke_CIU, "\n")
```

```
## 95% Confidence Interval for smoking: -0.07024124 to 0.01068324
```

The p-value for smoking is 0.1496 > 0.05, therefore, the effect of smoking is not statistically significant. The 95% CI is (-0.0702, 0.0107)

###(iv)

```
exp(0.045696)
```

```
## [1] 1.046756
```

The p-value of height is less than 0.05, indicating the coefficient for height is significant. For one inch increases in height, the expected FEV would increase by 4.68%.

###(v)

```
exp(0.076786)
```

```
## [1] 1.079811
```

Null hypothesis: the categories of age don’t have a significant effect on log(fev) Alternative hypothesis: the categories of age have a significant effect on log(fev) The p-value of age score is less than 0.05 and the t value is 6.025. Thus, we can reject the null hypothesis, meaning that the categories of age are significant. For every 5 years increase in age, the expected FEV would increase by 7.98%.

##(f)

```
new_data = data.frame(
  Hgt = 60,
  Sex = 1,
  Smoke = 0,
  Age_Score = 4

)

predict_value = predict(fev_model2, newdata = new_data)
print(predict_value)
```

```
##        1
## 1.002876
```

```
exp(predict_value)
```

```
##        1
## 2.726112
```

The log(FEV) is 1.002876 and the predicted FEV is 2.726112.