Residual Analysis & WLS

Eduardo Martinez

November, 2021

Contents

This was a homework assignment for Applied Statistics & Regression, a graduate course that I completed at Illinois Institute of Technology. Many problems come from the following source:

- Textbook: Chatterjee, S., & Hadi, A.S. (2012). Regression Analysis by Example. 5th Edition.

```
library(tidyverse)
```

Problem 1: Exercise 4.8 (b)

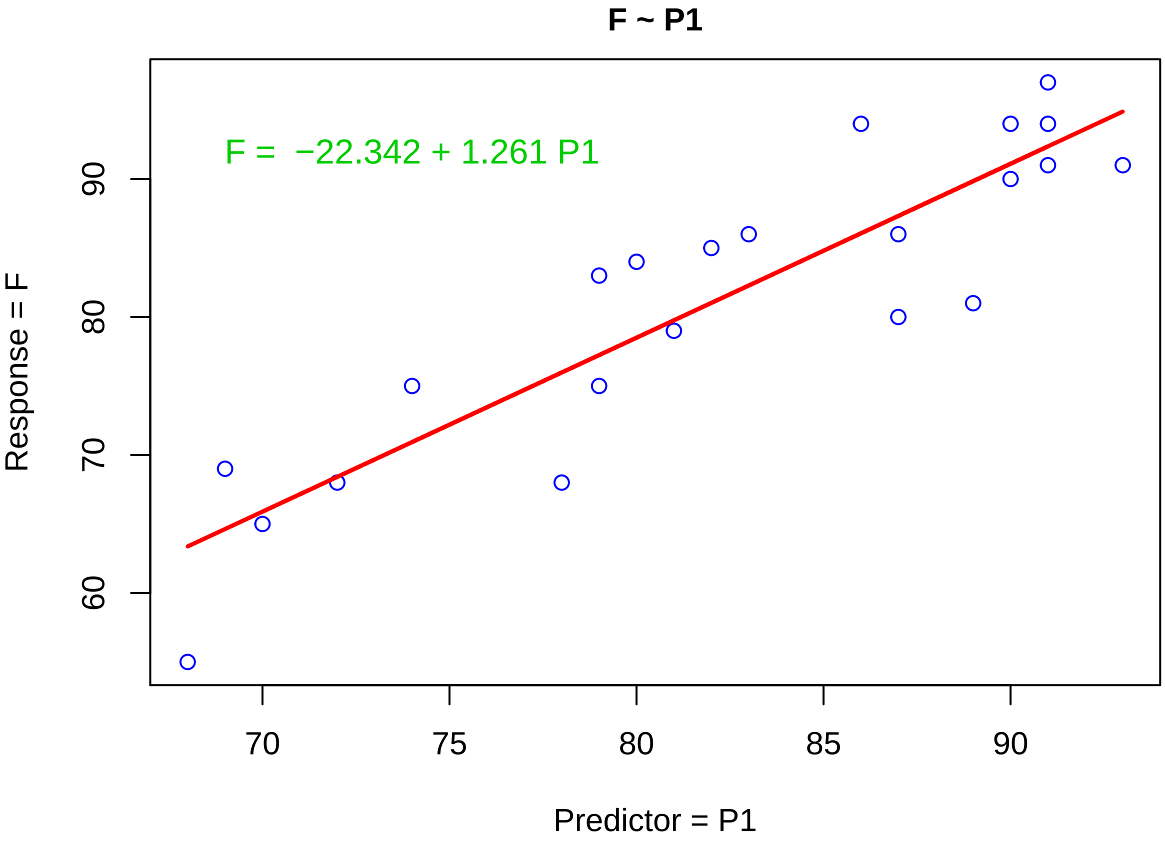
Instructions Consider again the Examination Data used in Exercise 3.3 and given in Table 3.10:

Part (b) What model would you use to predict the final score F?

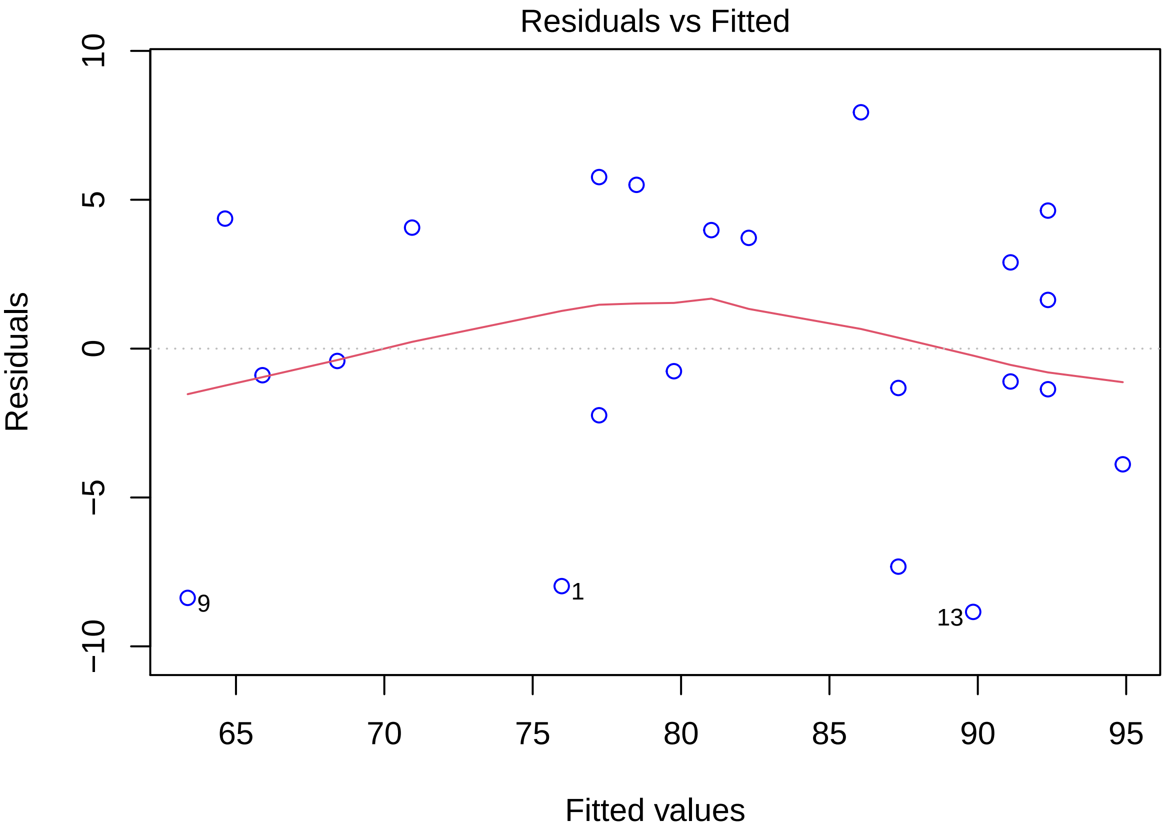
```
ExamData <- read_tsv("Table3.10.txt")
head(ExamData, 5)
```

```
# A tibble: 5 x 3
      F    P1    P2
  <dbl> <dbl> <dbl>
1    68    78    73
2    75    74    76
3    85    82    79
4    94    90    96
5    86    87    90
```

```
Model1fit <- lm(`F` ~ P1, data = ExamData)
Model1 <- summary(Model1fit)
```

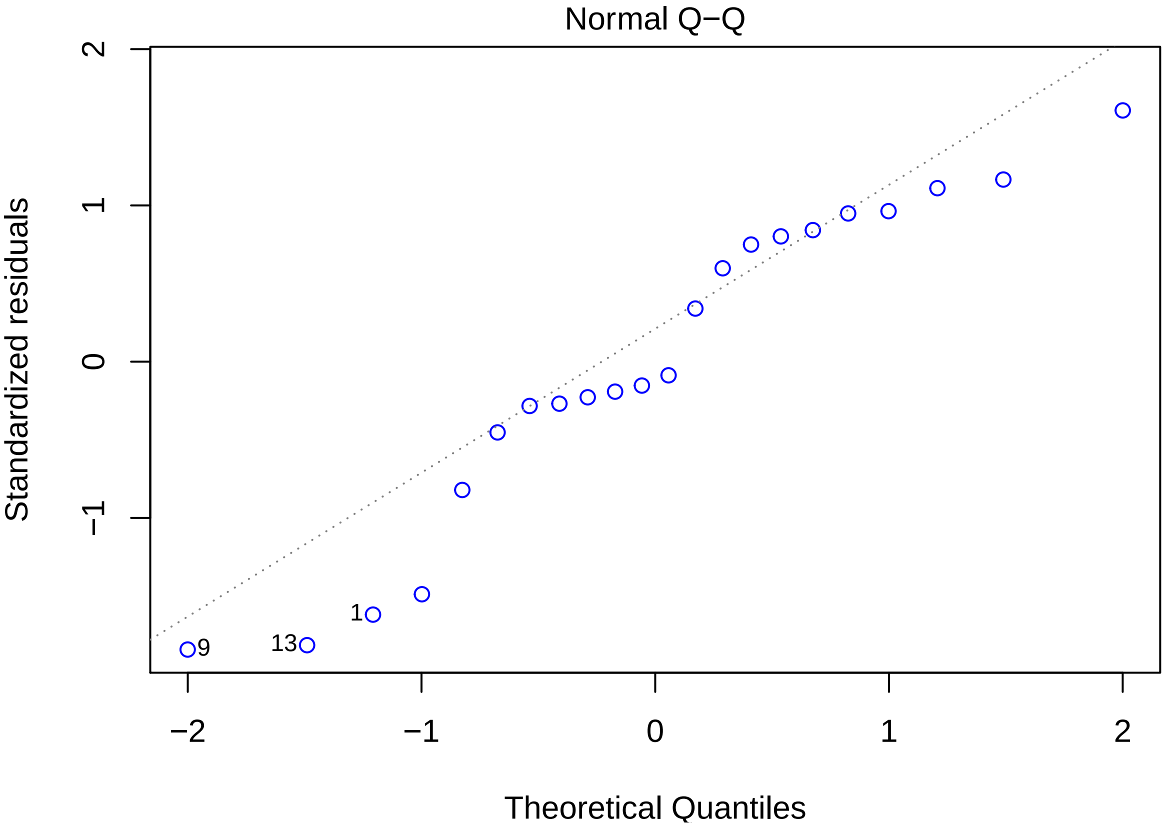


```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(Model1fit, which = 1, col = "blue")
```



Observations: The 1st, 9th, and 13th observations have noticeablly higher residual error, suggesting that they may be unusual observations. Further, all of these 3 fitted values overestimate the actual value of F; for instance, observation 1 (P1 = 78) predicts that F = 76 when it really equal 68. This means that there is not equal variance in the residual errors.

```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(Model1fit, which = 2, col = "blue")
```

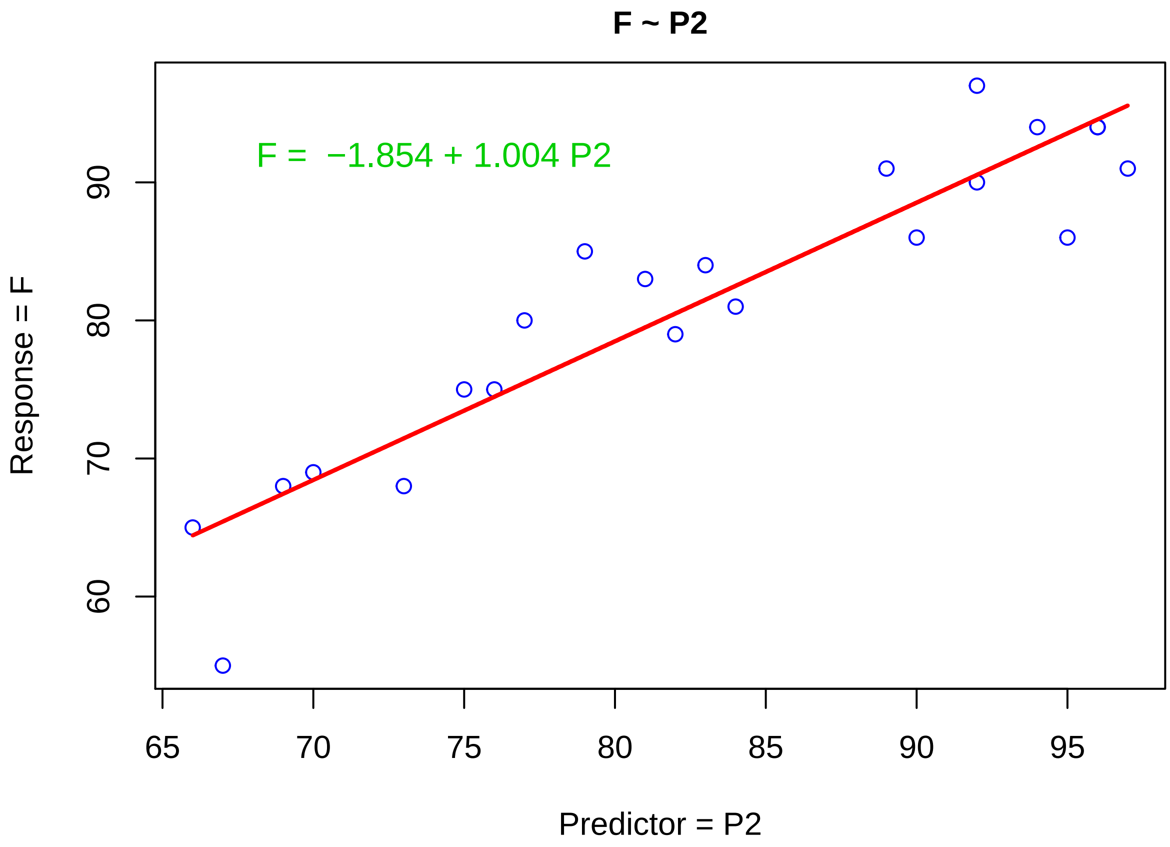


Observations: The Normal Q-Q plot reinforces my previous observations. Most points fall below the Normal line, so the distribution is skewed to the left. Hence, the residuals do not appear normally distributed.

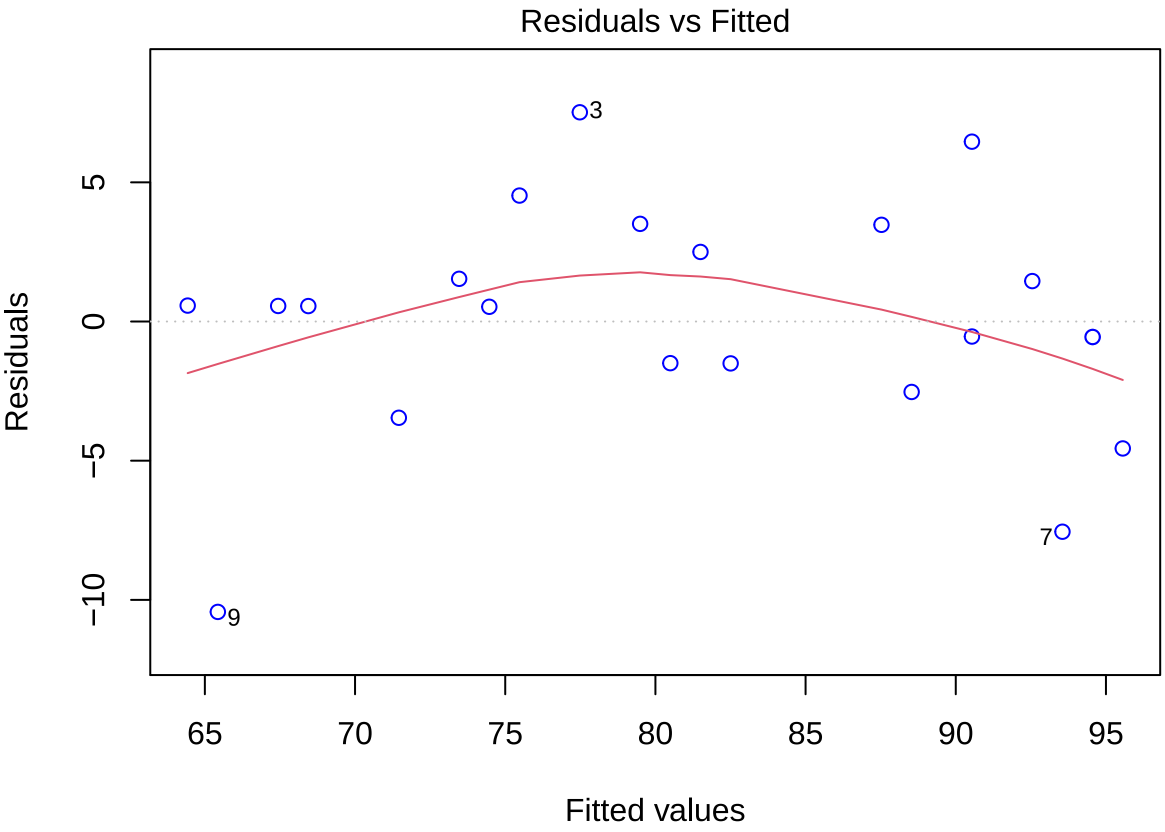
Next, I will check if P2 does a better job at demonstrating the residual assumptions.

```
Model2fit <- lm(F ~ P2, data = ExamData)
Model2 <- summary(Model2fit)
```

```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(x = ExamData$P2, y = ExamData$F, col = "blue",
     main = "F ~ P2", xlab = "Predictor = P2", ylab = "Response = F", cex.main = 1)
lines(x = ExamData$P2, y = Model2fit$fitted.values, col = "red", lwd = 2)
text(74, 92, col = "green3", cex = 1.1,
     paste("F = ", round(coefficients(Model2fit)[[1]],3), "+", round(coefficients(Model2fit)[[2]],3), "
```

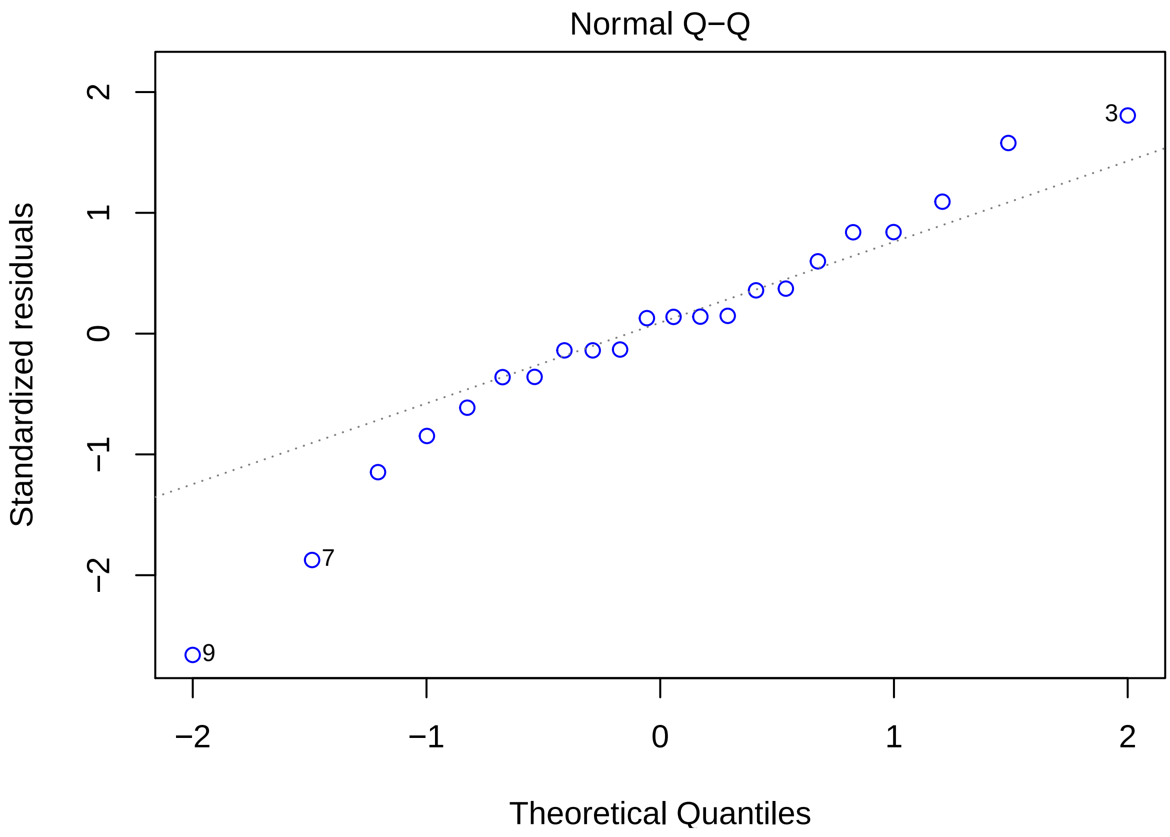


```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(Model2fit, which = 1, col = "blue")
```



Observations: The 3rd, 7th, and 9th observations have noticeablly higher residual error, suggesting that they may be unusual observations and are potential outliers. In addition, there seems to be a pattern such that significant residual error is initially below zero, it increases above zero, and ends below zero (bell shape). This means that the fitted values do not have equal variance.

```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(Model2fit, which = 2, col = "blue")
```

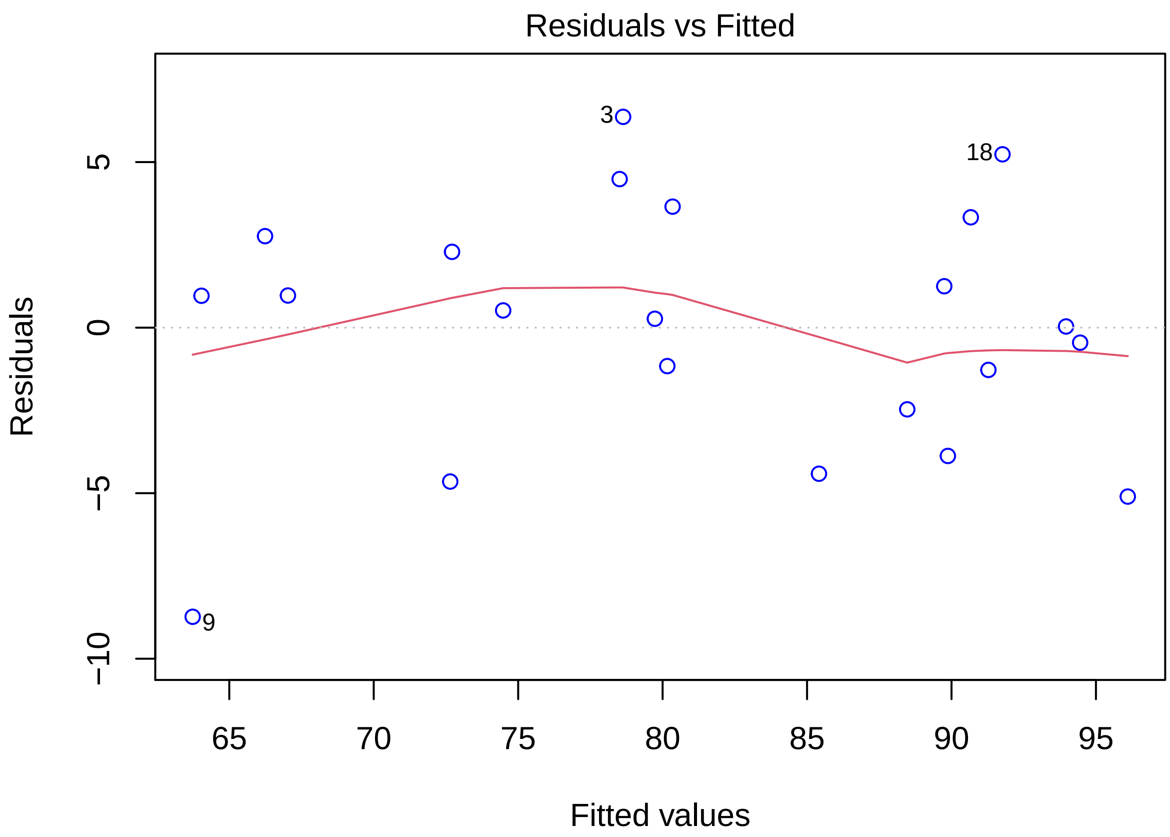


Observations: The points are not consistently close the Normal Q-Q line. The distribution is heavy-tailed as points curve away from the line at each end; interestingly, they tails curve in opposite directions. As a result, the residuals do not appear normally distributed.

When P1 and P2 are used as the only predictors of F, they do not appear to pass the error assumptions for linear regression. That is, the residuals do not Normally distributed or equal variance. Therefore, I will assess if the combined model follows the error assumptions.

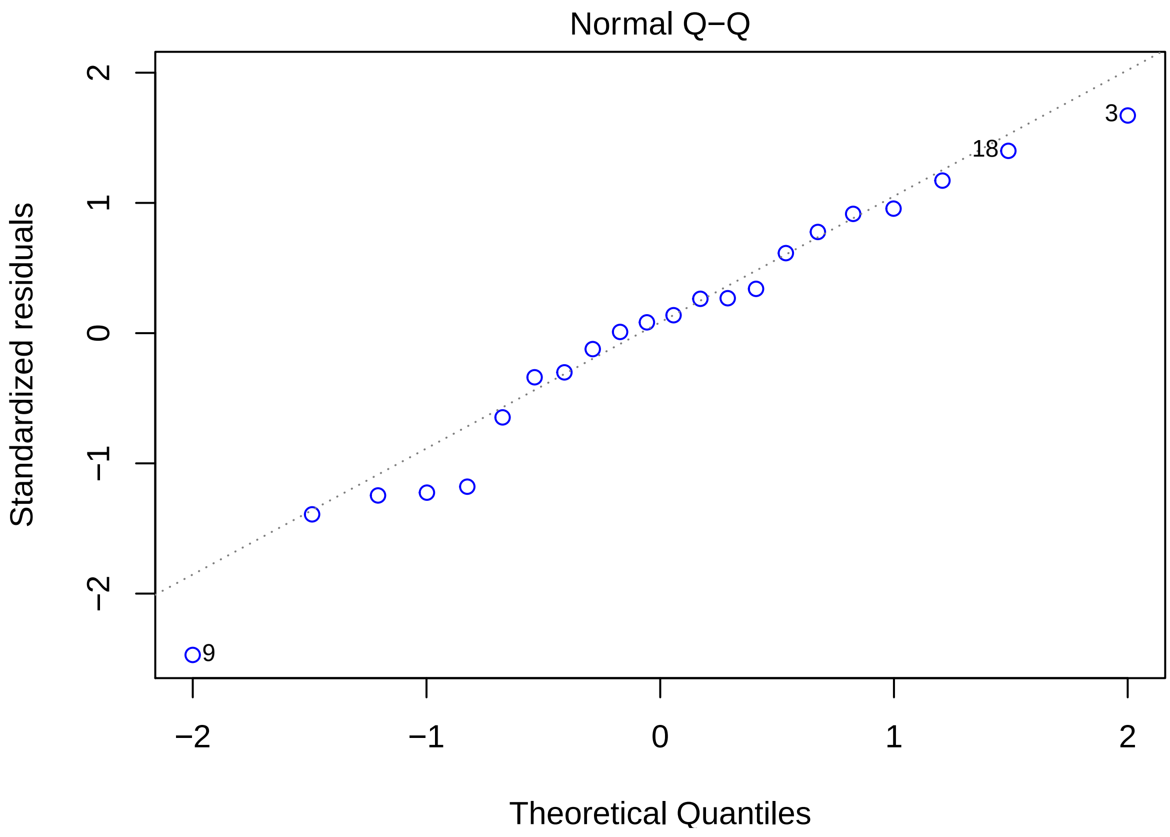
```
FullModelFit <- lm(`F` ~ P1 + P2, data = ExamData)
FullModel <- summary(FullModelFit)
```

```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(FullModelFit, which = 1, col = c("blue"))
```



Observations: Compared to the previous two models, more fitted values have residual error close to zero. Although observations 3, 9, and 18 are potential outliers, the residuals appear to have equal variances overall.

```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(FullModelFit, which = 2, col = "blue")
```



Observations: Aside from observation 9, the distribution seems to be normally distributed as most points fall close to the Normal Q-Q line.

Conclusion: Based on the residual plots, I would use the Full Model that uses both P1 and P2 to predict F. Model 1 and Model 2 noticeably violate the error assumptions, while the Full Model does not._____

Problem 2

Instructions: Given data in Table 6.2, the response variable is n_t, representing the number of surviving bacteria (in hundreds) after being exposed to X-ray for t intervals. The predictor variable is t.

Part 1 First regress_ n_t on time t, plot residuals against the fitted values $\hat{n}_t$. Conclude if the relationship between the mean response and the the predictor is linear.

Part 2 Use data transformation on the response variable variable, i.e., regress $\log(N_t)$ on t.

- What is the regression line equation?
- Plot the residuals against the fitted values, and conclude if the violation of the "L" assumption still exists.

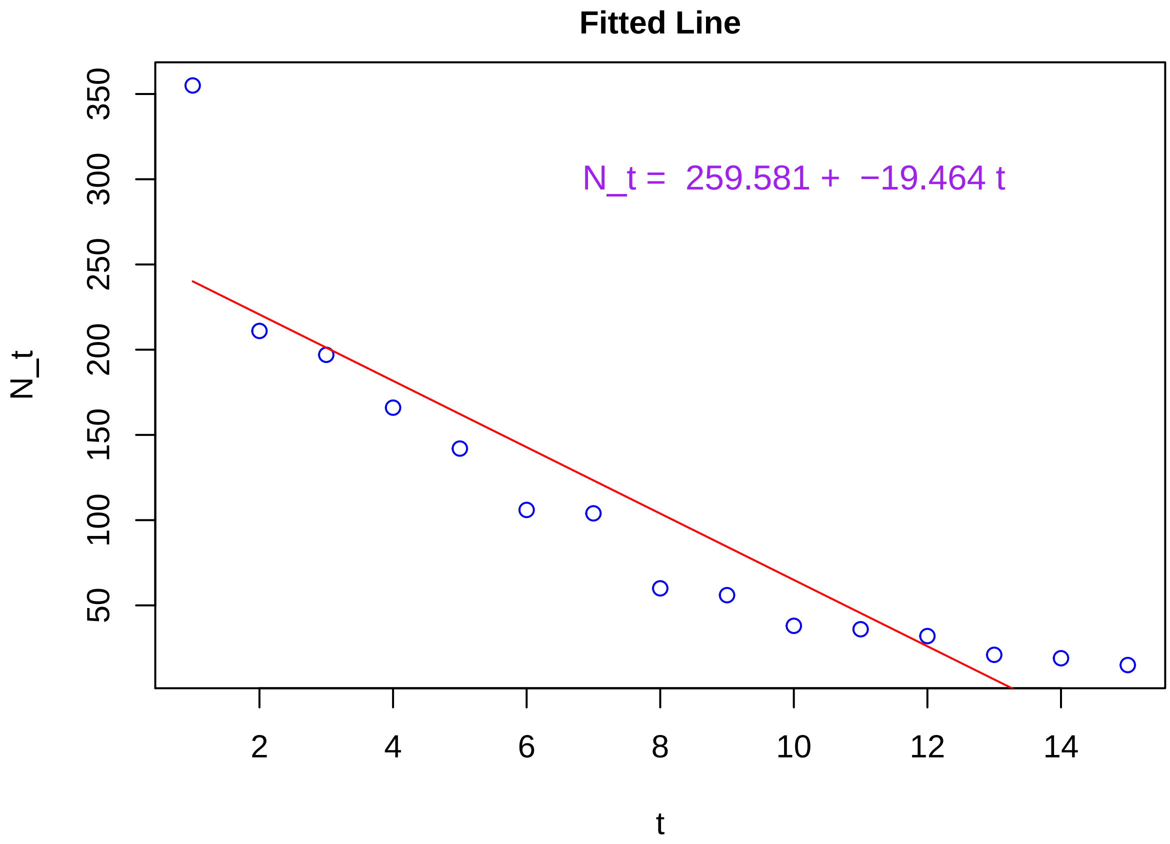
```
Table6.2 <- read_tsv("Table6.2.txt")
head(Table6.2, 5)
```

```
# A tibble: 5 x 2
      t   N_t
  <dbl> <dbl>
1     1   355
2     2   211
3     3   197
4     4   166
5     5   142
```

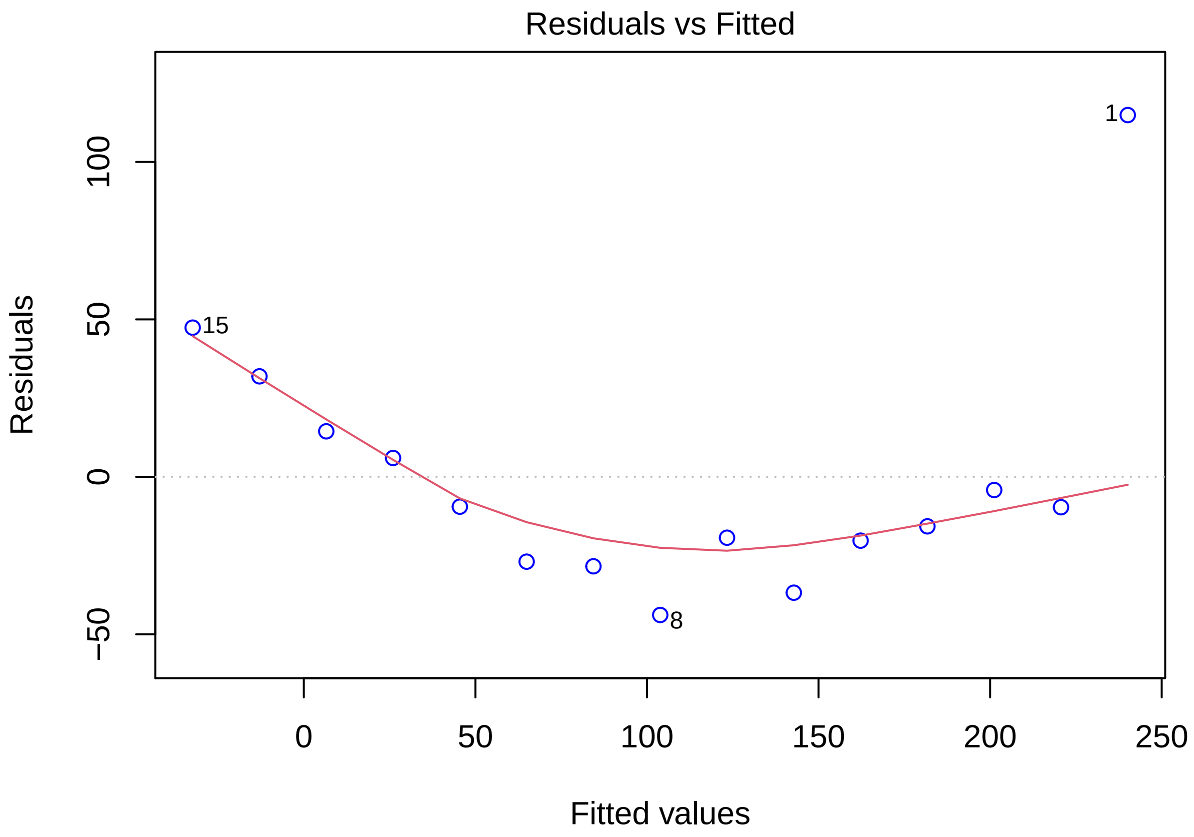
Part 1

```
Model1Fit <- lm(N_t ~ t, data = Table6.2)
Model1 <- summary(Model1Fit)
Fits <- Model1Fit$fitted.values
```

```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(x = Table6.2$t, y = Table6.2$N_t, col = "blue",
     main = "Fitted Line", xlab = "t", ylab = "N_t", cex.main = 1)
lines(x = Table6.2$t, y = Fits, col = "red", lwd = 1)
text(10, 300, cex = 1.1, col = "purple",
     paste("N_t = ", round(coefficients(Model1Fit)[[1]],3), "+ ", round(coefficients(Model1Fit)[[2]],3)
```



```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(Model1Fit, which = 1, col = "blue")
```



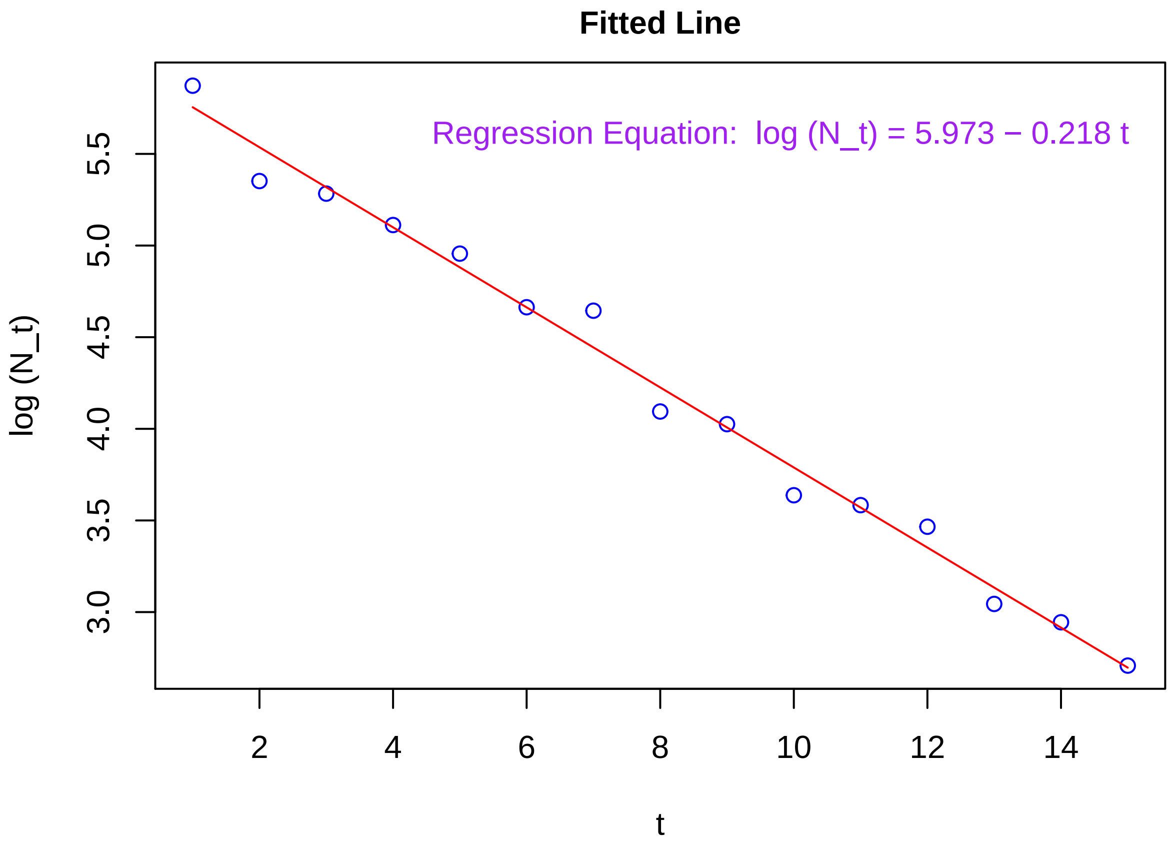
Conclusion: It is quite obvious that the relationship between N_t and t is not linear. Initially, the actual data points decrease rapidly; then, they start to flatten when t = 8. Moreover, the residual plot does not exhibit equal variance as the residual errors have a backward hill shape. These plots suggests an exponential relationship.

Part 2

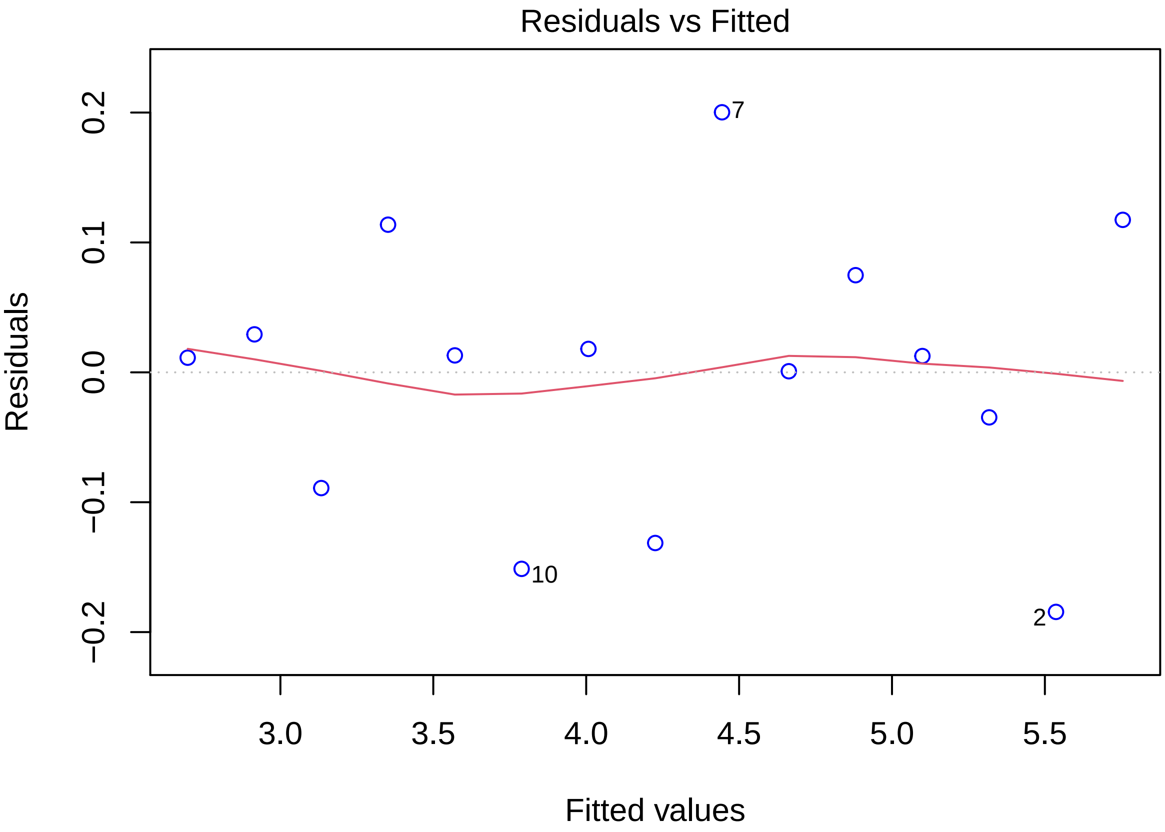
```
y <- log(Table6.2$N_t)
x <- Table6.2$t
```

```
LogModelFit <- lm(y ~ x)
LogModel <- summary(LogModelFit)
LogFits <- LogModelFit$fitted.values
```

```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(x, y, col = "blue",
     main = "Fitted Line", xlab = "t", ylab = "log (N_t)", cex.main = 1)
lines(x, y = LogFits, col = "red", lwd = 1)
text(9.8, 5.6, cex = 1, col = "purple",
     paste("Regression Equation:  log (N_t) =",
           round(coefficients(LogModelFit)[[1]],3), "-", abs(round(coefficients(LogModelFit)[[2]],3)),
```



```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(LogModelFit, which = 1, col = "blue")
```



Conclusion: The violation of the "L" assumption no longer exist. The data points are consistently very close to the regression line; the residuals have equal variance and their means are approximately zero.

Problem 3

Instructions Given the data Table 6.6, the response variable Y, is the number of injury incidents, and the predictor variable N is the proportion of flights.

Part 1 First regress Y on N, plot residuals against the fitted values $\hat{Y}$. Conclude if the error is heteroscedastic, i.e., the "E" assumption is violated.

Part 2 Use data transformation on the response variable variable, i.e., regress $\sqrt{Y}$ on N. The rational behind this transformation is that the occurrence of accidents, Y, tends to follow the Poisson probability distribution, and the variance of $\sqrt{Y}$ is approximately equal to 0.25, see Table 6.5

What is the regression line equation?
Plot the residuals against the fitted values, and conclude if there is still evidence of heteroscedasticity.

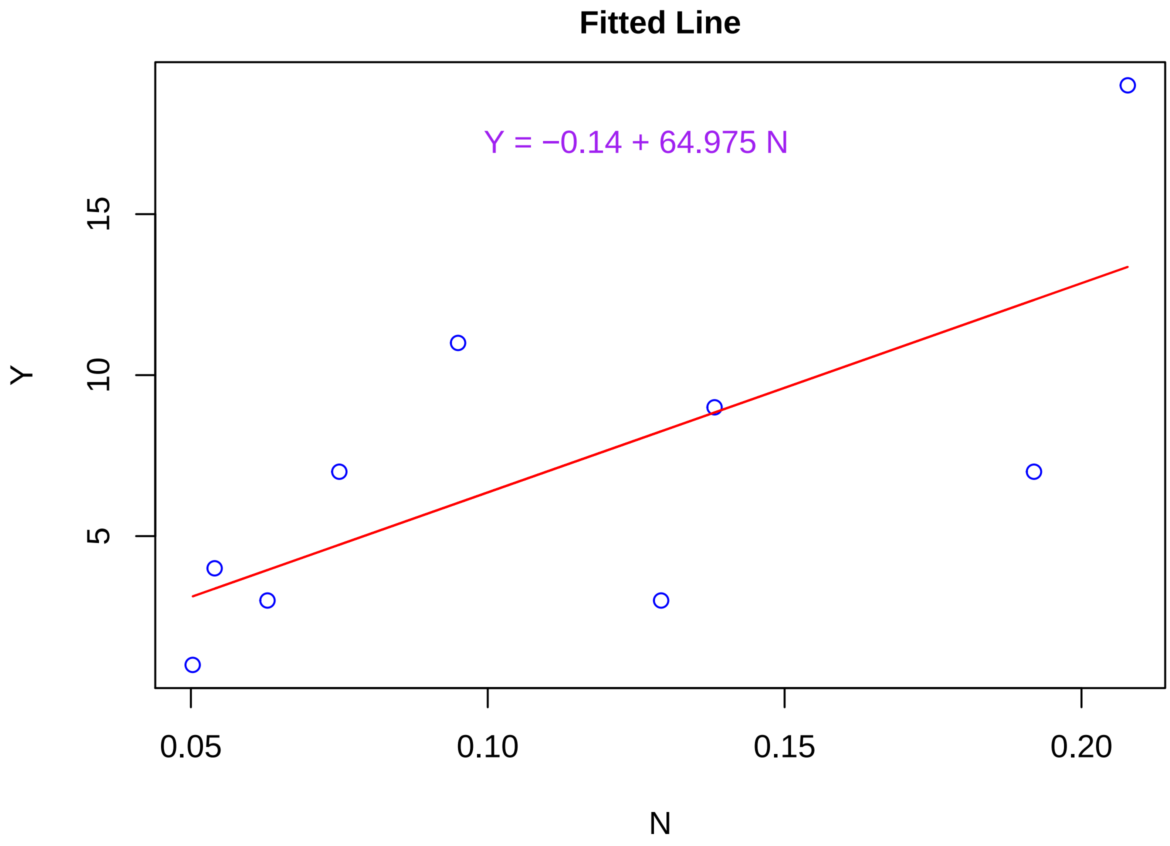
```
Table6.6 <- read_tsv("Table6.6.txt")
```

```
# A tibble: 5 x 2
      Y     N
  <dbl> <dbl>
1    11 0.095
2     7 0.192
3     7 0.075
4    19 0.208
5     9 0.138
```

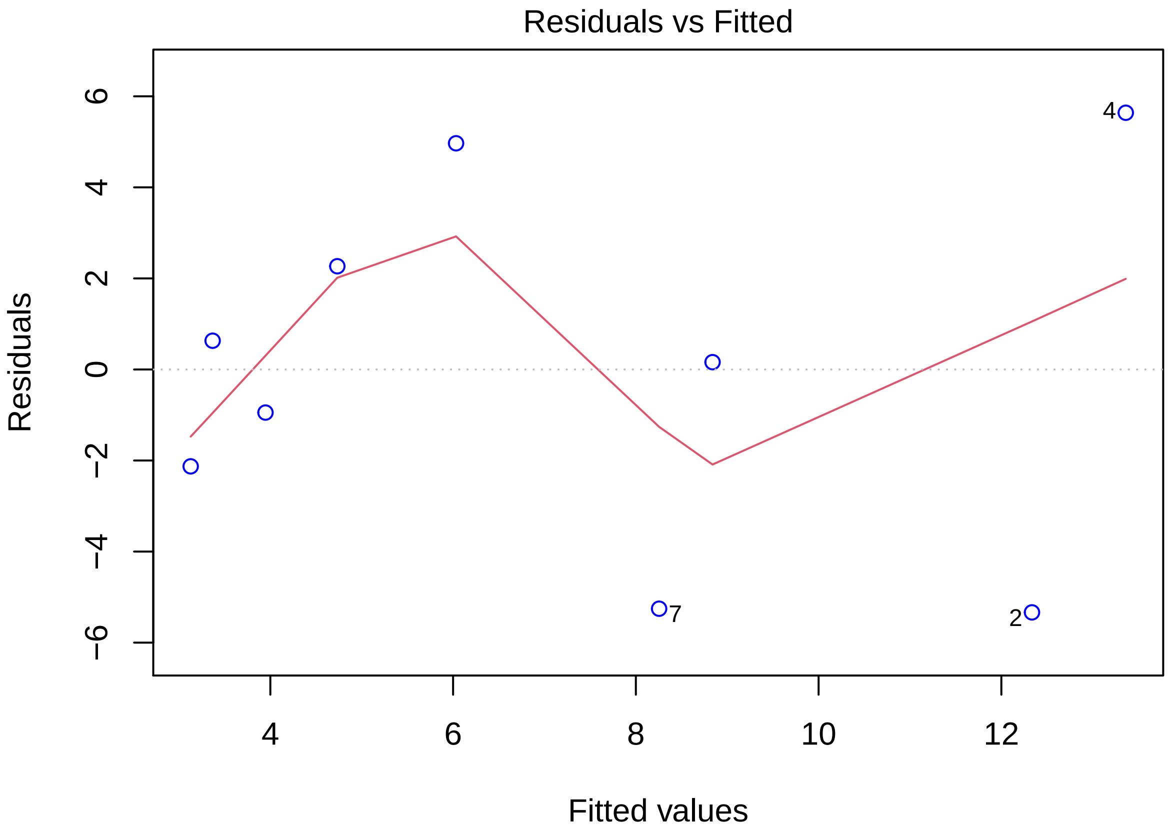
Part 1

```
ModelYNFit <- lm(Y ~ N, data = Table6.6)
ModelYN <- summary(ModelYNFit)
FitsYN <- ModelYNFit$fitted.values
```

```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(x = Table6.6$N, y = Table6.6$Y, col = "blue",
     main = "Fitted Line", xlab = "N", ylab = "Y", cex.main = 1)
lines(x = Table6.6$N, y = FitsYN, col = "red", lwd = 1)
text(0.125, 17.25, col = "purple",
     paste("Y =", round(coefficients(ModelYNFit)[[1]],3), "+", round(coefficients(ModelYNFit)[[2]],3),
```



```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(ModelYNFit, which = 1, col = c("blue"))
```



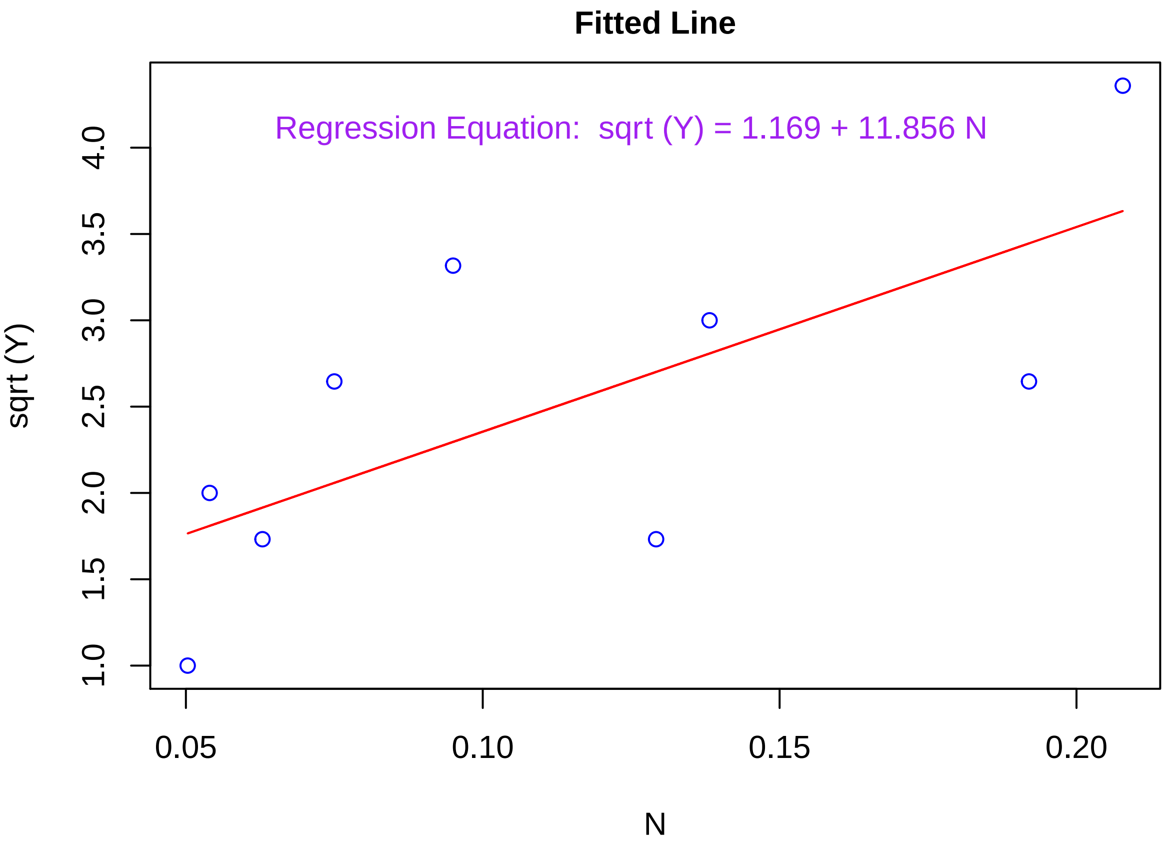
Conclusion: The error is heteroscedastic, violating the "E" assumption. There is a pattern of the error increasing in magnitude from left to right (megaphone shape). This suggests that there is no equal variance. Additionaly, the residual line has a pattern of going up-down-up; if the "E" assumption was not violated, we would expect random residual errors with no patterns.

Part 2

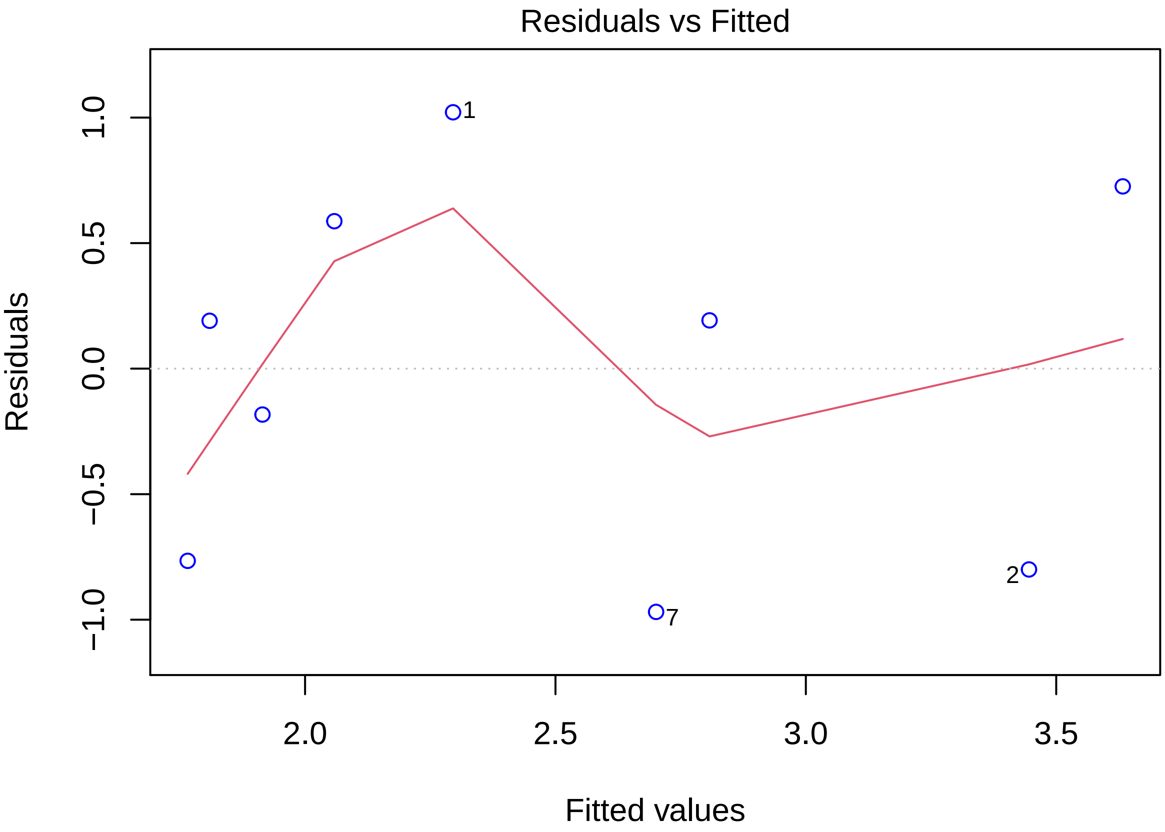
```
XN <- Table6.6$N
YY <- sqrt(Table6.6$Y)
```

```
SqrtModelFitYN <- lm(YY ~ XN)
SqrtModelYN <- summary(SqrtModelFitYN)
SqrtFitsYN <- SqrtModelFitYN$fitted.values
```

```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(XN, YY, col = "blue",
     main = "Fitted Line", xlab = "N", ylab = "sqrt (Y)", cex.main = 1)
lines(XN, SqrtFitsYN, col = "red", lwd = 1)
text(0.125, 4.1, col = "purple",
     paste("Regression Equation:  sqrt (Y) =",
           round(coefficients(SqrtModelFitYN)[[1]],3), "+", round(coefficients(SqrtModelFitYN)[[2]],3),
```



```
par(mar = c(4.1, 4.1, 2.1, 2.1))
plot(SqrtModelFitYN, which = 1, col = "blue")
```



Conclusion: The residuals no longer have a megaphone shape as variance neither decreases or incrases from left to right. Hence, there is not clear evidence of heteroscedasticity. At any rate, N still does not seem to be a strong linear predictor of Y as the fitted values are not consistently close to the regression line.

Problem 4

Instructions Given the data in Table 6.9, the response variable Y is the number of supervisors, and the predictor variable X is the number of supervised workers. Based on empirical observations, it is hypothesized that the standard deviation of the error term ϵ_i is proportional to x_i:

$$\sigma_i^2 = k^2 x_i^2 \ , \ k > 0$$

Part 1 Use the weighted least squared (WLS) method to fit the model. Provide the regression equation.

Part 1 Use data transformation method to transform Y to $Y' = Y/X$, and transform X to $X' = 1/X$ (see equations 6.11 and 6.12), and then use the OLS method to regress Y' on X'. Provide the regression equation.

Part 3 Compare the results from the above two methods and conclude if the two methods are equivalent. You can compare the residual vs. fitted value plot side by side and conclude if they have the same effect in terms of removing heteroscedasticity.

```
Table6.9 <- read_tsv("Table6.9.txt")
```

Part 1

Given the hypothesized relationship. $w_i = 1/x_i^2$.

Thus, the parameters are determined by minimizing:

$$\sum_{i=1}^{n} \frac{1}{x_i^2}(y_i - \beta_0 - \beta_1 x_i)^2$$

This methods yields the following coefficients:

```
w <- 1 / (Table6.9$X)^2
FitWLS <- lm(Y ~ X, data = Table6.9, weights = 1 / X^2)
FitWLS_Model <- summary(FitWLS)
WLS_Fits <- FitWLS$fitted.values
FitWLS_Model$coefficients
```

```
             Estimate  Std. Error    t value     Pr(>|t|)
(Intercept) 3.8032958 4.569745381  0.8322774 4.131324e-01
X           0.1209903 0.008998637 13.4454039 6.043603e-13
```

The Corresponding WLS Regression Equation: $Y = 3.803 + 0.121X$

Part 2

```
Ynew = Table6.9$Y / Table6.9$X
Xnew = 1 / Table6.9$X
```

The transformation methods yields the following regression coefficients:

```
FitOLS <- lm(Ynew ~ Xnew)
FitOLS_Model <- summary(FitOLS)
OLS_Fits <- FitOLS$fitted.values
FitOLS_Model$coefficients
```

```
             Estimate  Std. Error    t value     Pr(>|t|)
(Intercept) 0.1209903 0.008998637 13.4454039 6.043603e-13
Xnew        3.8032958 4.569745381  0.8322774 4.131324e-01
```

The Corresponding OLS Regression Equation: $Y' = 0.121 + 3.803X'$

Part 3

$\sigma_i^2 = k^2 x_i^2$, $k > 0 \Rightarrow w_i = 1/x_i^2 \Rightarrow Var(\varepsilon) = W^{-1}\sigma^2 \Rightarrow \varepsilon' = W^{1/2}\varepsilon$

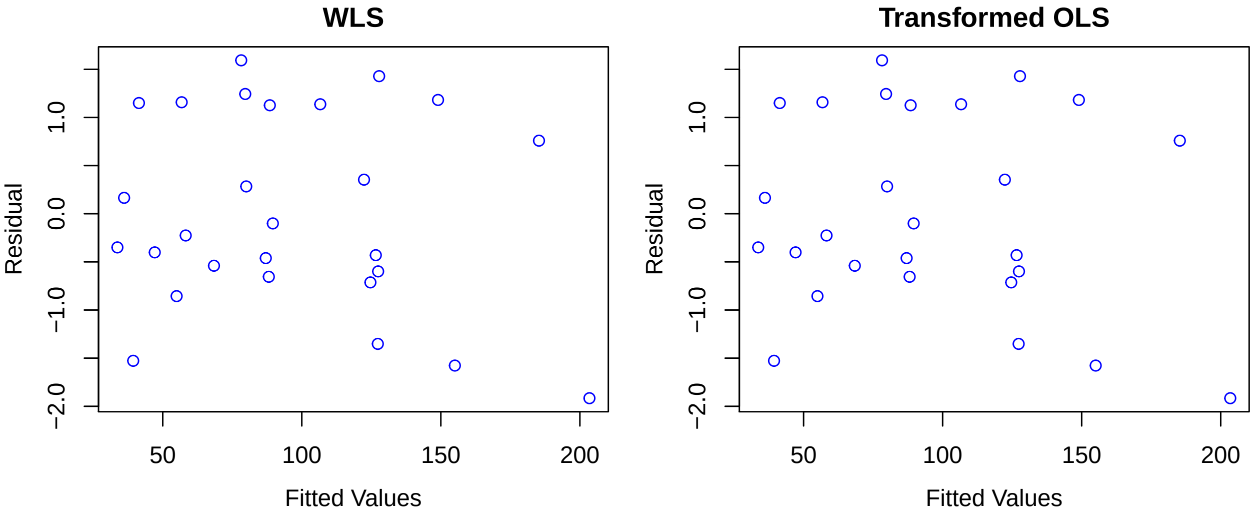
Therefore, I used the standardized residuals for the residual plots.

In terms of the corresponding regression equations, if we reverse the transformations in the transformed OLS method, the resulting equation is the same as the WLS regression equation. Given $Y' = Y/X$, I multiplied the fitted values by X in the residual plot for the OLS method.

```r
par(mfrow = c(1, 2), mar = c(5.1, 4.1, 2.1, 0.5))

plot(FitWLS$fitted.values, rstandard(FitWLS), col = "blue", #xaxt = "none",
     xlab = "", ylab = "", main = "WLS")
# axis(1, seq(250, 1650, 250))
mtext(side=1, line = 2.5, "Fitted Values")
mtext(side=2, line = 2.5, "Residual")

plot(FitOLS$fitted.values * Table6.9$X, rstandard(FitOLS), col = "blue",
     xlab = "", ylab = "", main = "Transformed OLS")
mtext(side=1, line = 2.5, "Fitted Values")
mtext(side=2, line = 2.5, "Residual")
```



Conclusion: The resulting residual vs fitted value plots have the same effect in terms of removing heteroscedasticity. The residual plots are equivalent.

Problem 5

Instructions For the data in Problem 4, use OLS without data transformation to fit the model, i.e., directly regress Y on X, and compare the variances of the coefficients $Var(\hat{\beta}_0)$ and $Var(\hat{\beta}_1)$ with their counterparts obtained using WLS, conclude which method yields smaller variances.

OLS Model without Data Transformations

```
Fit5OLS <- lm(Y ~ X, data = Table6.9)
Fit5OLS_Model <- summary(Fit5OLS)
OLS5_Fits <- Fit5OLS$fitted.values
Fit5OLS_Model
```

```

Call:
lm(formula = Y ~ X, data = Table6.9)

Residuals:
    Min      1Q  Median      3Q     Max 
-53.294  -9.298  -5.579  14.394  39.119 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 14.44806    9.56201   1.511    0.143    
X            0.10536    0.01133   9.303 1.35e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.73 on 25 degrees of freedom
Multiple R-squared:  0.7759,    Adjusted R-squared:  0.7669 
F-statistic: 86.54 on 1 and 25 DF,  p-value: 1.35e-09
```

```
meanX <- mean(Table6.9$X)
meanY <- mean(Table6.9$Y)
```

```
B1 <- sum((Table6.9$X - meanX) * (Table6.9$Y - meanY)) / sum((Table6.9$X - meanX)^2)
noquote(paste("B1_ols =", round(B1,4)))
```

```
[1] B1_ols = 0.1054
```

```
B0 <- meanY - B1 * meanX
noquote(paste("B0_ols =", round(B0,4)))
```

```
[1] B0_ols = 14.4481
```

```
sigmaOLS <- Fit5OLS_Model$sigma

VarB1 <- sigmaOLS^2 / sum((Table6.9$X - meanX)^2)
noquote(paste("Var(B1_ols) =", round(VarB1,6)))
```

```
[1] Var(B1_ols) = 0.000128
```

```
VarB0 <- sigmaOLS^2 * ( (1 / length(Table6.9$X) ) + (meanX^2 / sum((Table6.9$X - meanX)^2)) )
noquote(paste("Var(B0_ols) =", round(VarB0,3)))
```

```
[1] Var(B0_ols) = 91.432
```

```
tcrit <- 1.70814

lwB1 <- B1 - (tcrit * sqrt(VarB1))
upB1 <- B1 + (tcrit * sqrt(VarB1))
noquote(paste("95% C.I. B1_ols: (", round(lwB1,3), ", ", round(upB1,3), ")"))
```

```
[1] 95% C.I. B1_ols: ( 0.086 ,  0.125 )
```

```
lwB0 <- B0 - (tcrit * sqrt(VarB0))
upB0 <- B0 + (tcrit * sqrt(VarB0))
noquote(paste("95% C.I. B0_ols: (", round(lwB0,3), ", ", round(upB0,3), ")"))
```

```
[1] 95% C.I. B0_ols: ( -1.885 ,  30.781 )
```

WLS Model

```
FitWLS_Model
```

```
Call:
lm(formula = Y ~ X, data = Table6.9, weights = 1/X^2)

Weighted Residuals:
      Min        1Q    Median        3Q       Max
-0.041477 -0.013852 -0.004998  0.024671  0.035427

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.803296   4.569745   0.832    0.413
X           0.120990   0.008999  13.445 6.04e-13 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.02266 on 25 degrees of freedom
Multiple R-squared:  0.8785,    Adjusted R-squared:  0.8737
F-statistic: 180.8 on 1 and 25 DF,  p-value: 6.044e-13
```

```
w <- 1 / (Table6.9$X)^2
meanXw <- sum(w * Table6.9$X) / sum(w)
meanYw <- sum(w * Table6.9$Y) / sum(w)
```

```
B1w <- sum(w * (Table6.9$X - meanXw) * (Table6.9$Y - meanYw)) / sum(w * (Table6.9$X - meanXw)^2)
noquote(paste("B1_wls =", round(B1w, 5)))
```

```
[1] B1_wls = 0.12099
```

```
B0w <- meanYw - B1w * meanXw
noquote(paste("B0_wls =", round(B0w, 4)))
```

```
[1] B0_wls = 3.8033
```

```r
sigmaWLS <- FitWLS_Model$sigma

VarB1w <- sigmaWLS^2 / sum(w * (Table6.9$X - meanXw)^2)
noquote(paste("Var(B1_wls) =", round(VarB1w,8)))
```

```
[1] Var(B1_wls) = 8.098e-05
```

```r
VarB0w <- sigmaWLS^2 * ( (1 / sum(w)) + (meanXw^2 / sum(w * (Table6.9$X - meanXw)^2)) )
noquote(paste("Var(B0_wls) =", round(VarB0w,3)))
```

```
[1] Var(B0_wls) = 20.883
```

```r
tcrit <- 1.70814

lwB1w <- B1w - (tcrit * sqrt(VarB1w))
upB1w <- B1w + (tcrit * sqrt(VarB1w))
noquote(paste("95% C.I. B1_wls: (", round(lwB1w,3), ", ", round(upB1w,3), ")"))
```

```
[1] 95% C.I. B1_wls: ( 0.106 ,  0.136 )
```

```r
lwB0w <- B0w - (tcrit * sqrt(VarB0w))
upB0w <- B0w + (tcrit * sqrt(VarB0w))
noquote(paste("95% C.I. B0_wls: (", round(lwB0w,3), ", ", round(upB0w,3), ")"))
```

```
[1] 95% C.I. B0_wls: ( -4.002 ,  11.609 )
```

$Var(\hat{\beta}_{1_{wls}})$ was 4.7e-05 (~ 37%) less than $Var(\hat{\beta}_{1_{ols}})$, and $Var(\hat{\beta}_{0_{wls}})$ was 70.5 (~ 77%) less than $Var(\hat{\beta}_{0_{ols}})$.

As a result, the confidence interval for $\hat{\beta}_{1_{wls}}$ was approximately 17 less than the one for $\hat{\beta}_{1_{ols}}$, and $\hat{\beta}_{0_{ols}}$'s was 0.008 less than the one $\hat{\beta}_{0_{wls}}$.

Also, the overall residual standard error in WLS was 21.7 less than OLS's residual standard error.

Conclusion: WLS yielded smaller variances of the coefficients compared to the OLS method.