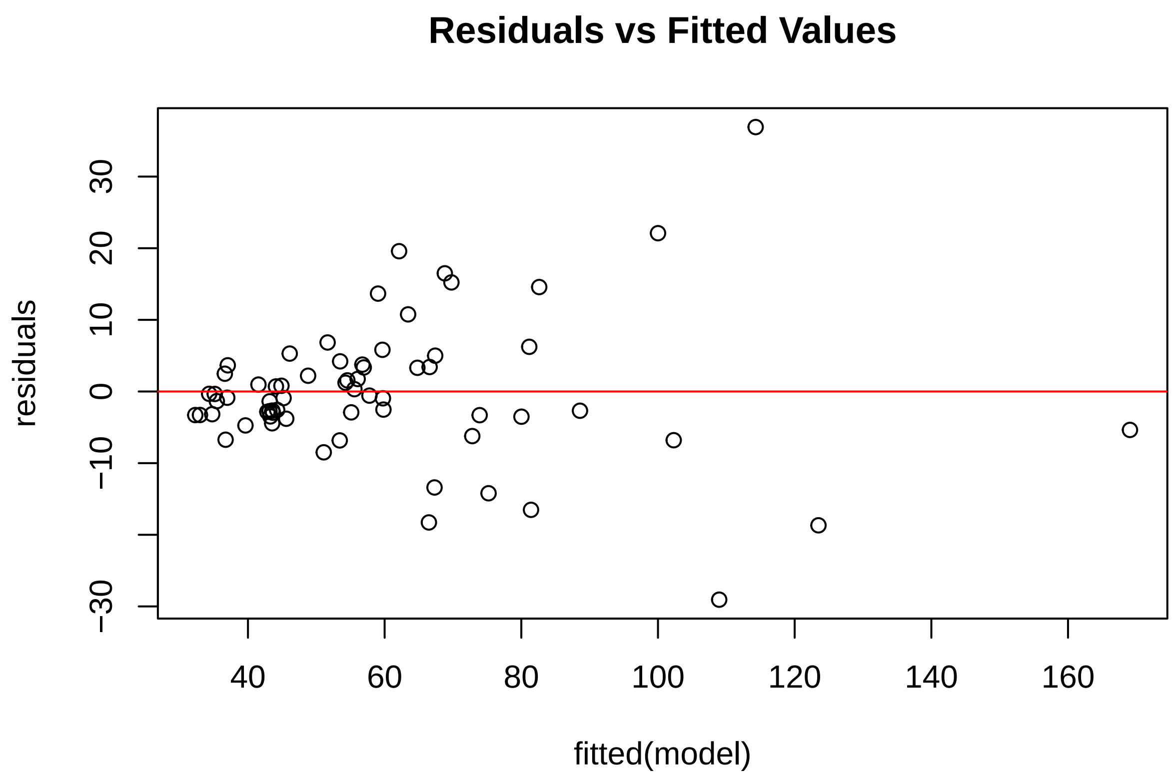
q5

Lisong He

2024-04-10

a)

```r
data <- read.csv('employee-salaries.csv')
model <- lm(Y ~ factor(data$Degree) + X3 + X4, data=data)
residuals <- resid(model)
plot(fitted(model), residuals, main="Residuals vs Fitted Values")
abline(h = 0, col = "red")
```



The plot shows unconstant variance in the models residuals increase as Y_hat increases. b)

```r
ordered_data <- data[order(fitted(model)), ]
library(ALSM)
```

```
## Loading required package: leaps
```

```
## Loading required package: SuppDists

## Loading required package: car

## Loading required package: carData
```

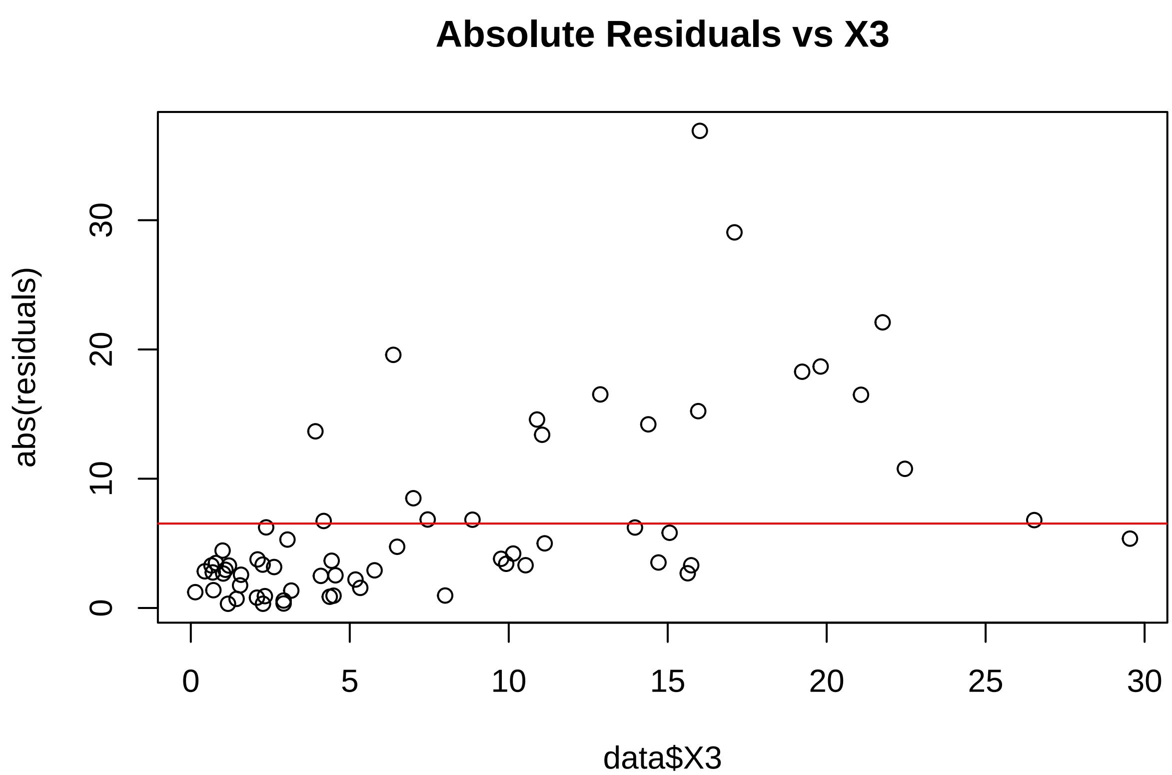
```
threshold <- data[order(fitted(model)), ][33,1]
group <- rep(1, length(data$Y))
group[fitted(model) <= threshold] <- 0
bftest(model, group, alpha=0.01)
```

```
##        t.value     P.Value alpha df
## [1,] 3.930298 0.000213524  0.01 63
```

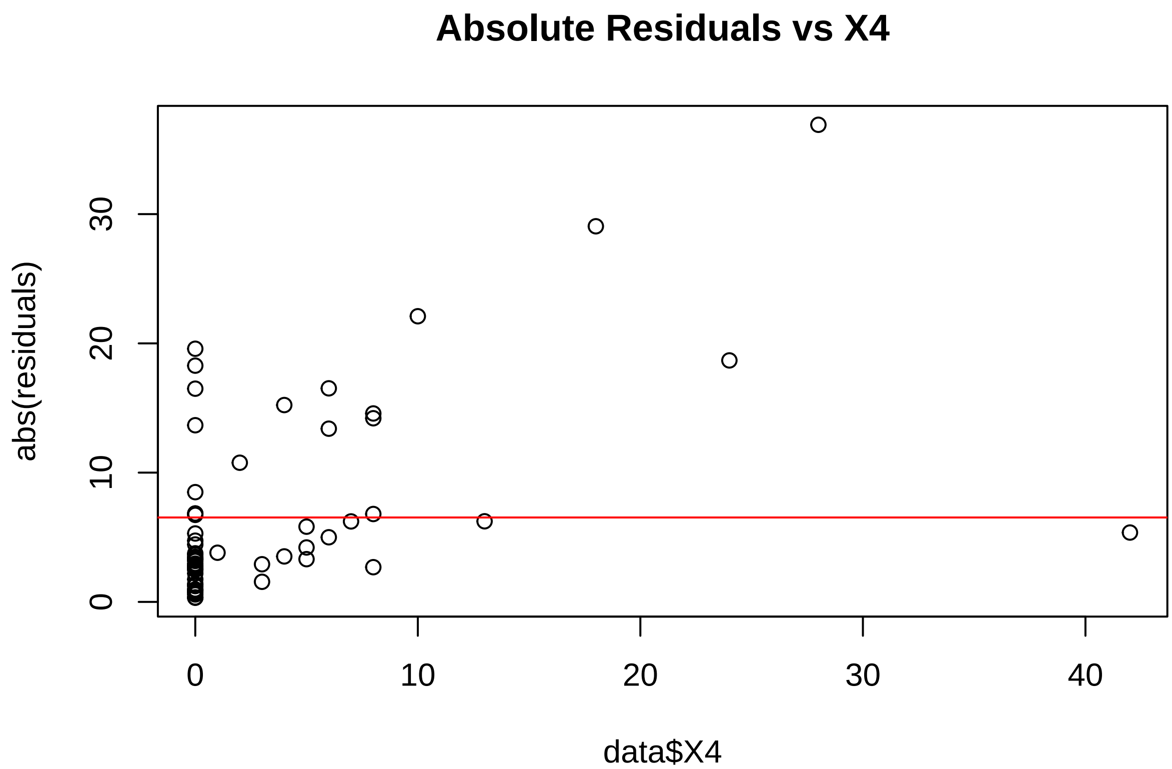
The p-value is way smaller than the significance interval alpha so we reject the null hypothesis of constant variances in favor of the alternative hypothesis of hetereoscedascitity.

c)

```
plot(data$X3, abs(residuals), main="Absolute Residuals vs X3")
abline(h = mean(abs(residuals)), col = "red")
```



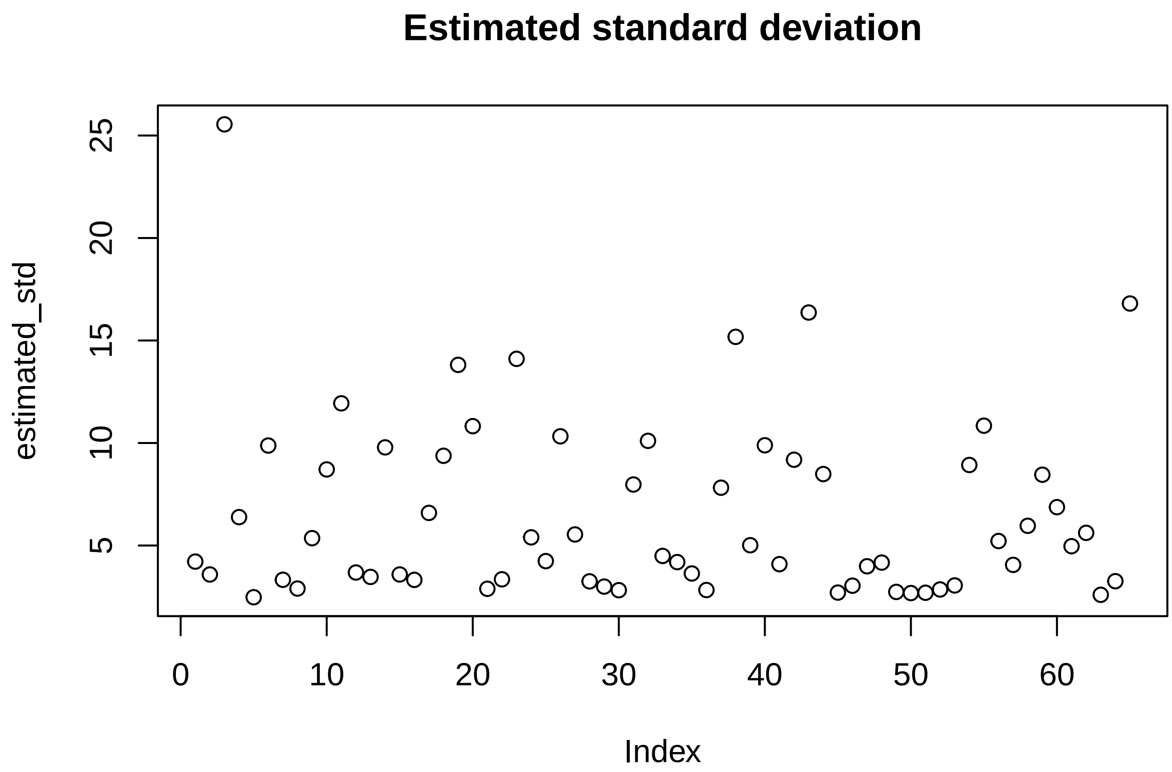
```r
plot(data$X4, abs(residuals), main="Absolute Residuals vs X4")
abline(h = mean(abs(residuals)), col = "red")
```



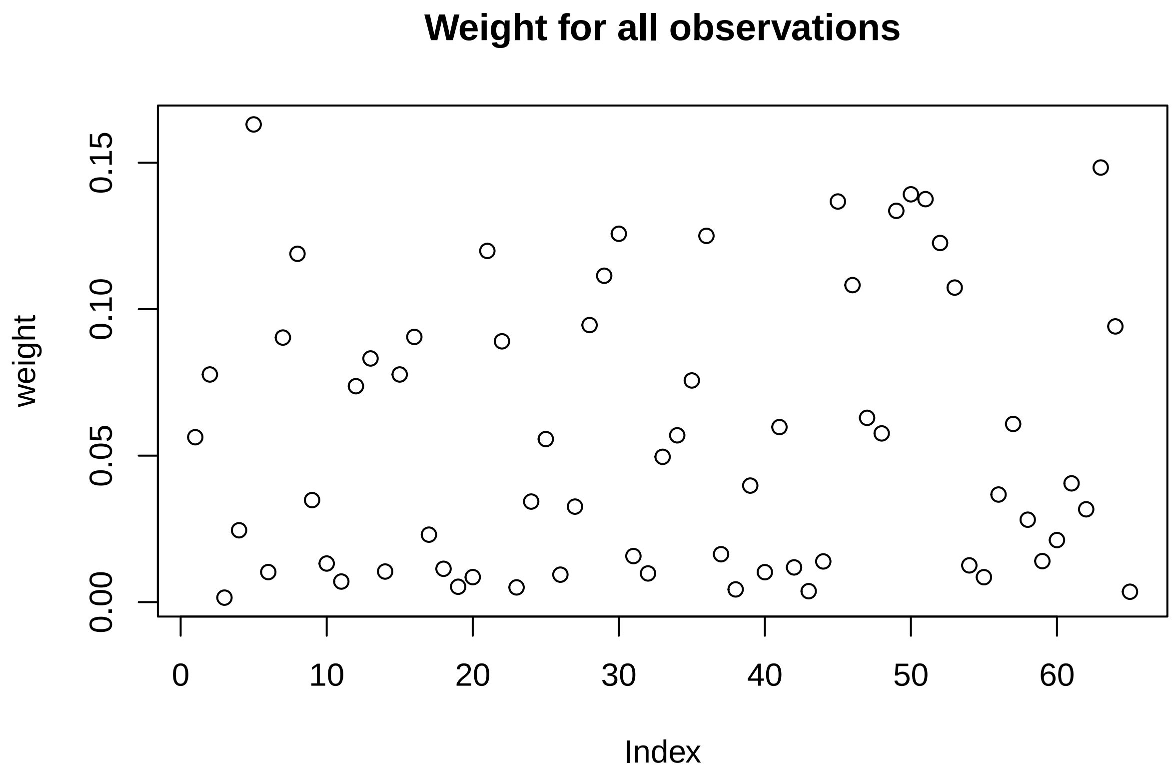
Both plots show increasing variance, ie standard deviation of error terms, as predictors increase since dots deviates from the mean of absolute residuals.

d)

```r
sd_model <- lm(abs(residuals) ~ data$X3 + data$X4)
estimated_std <- sd_model$fitted.values
weight <- 1 / (estimated_std)^2
plot(estimated_std, main = 'Estimated standard deviation')
```




```r
plot(weight, main = 'Weight for all observations')
```



e)

```
wls_model <- lm(Y ~ factor(data$Degree) + X3 + X4, data=data, weights=weight)
wls_model$coefficients
```

```
##         (Intercept) factor(data$Degree)2 factor(data$Degree)3
##           29.425494            10.899639            26.684879
##                  X3                   X4
##            1.425330             1.723949
```

```
model$coefficients
```

```
##         (Intercept) factor(data$Degree)2 factor(data$Degree)3
##           31.471430            10.811952            22.630727
##                  X3                   X4
##            1.258124             1.852313
```

All coefficients are discrepant from the ols model except that X1 and intercept don't have a big difference.

f)

```
standard_error_ols <- summary(model)$coefficients[, "Std. Error"]
standard_error_wls <- summary(wls_model)$coefficients[, "Std. Error"]
data.frame(standard_error_ols, standard_error_wls)
```

```
##                        standard_error_ols standard_error_wls
## (Intercept)                     2.8690780          1.3616587
## factor(data$Degree)2            3.2183072          1.4917607
## factor(data$Degree)3            3.4846124          1.6686145
## X3                              0.2272961          0.2002390
## X4                              0.2276418          0.3206416
```

Standard errors of ols is larger than that of wls except in X4, indicating that wls has more accruate models than ols.

g)

```r
sd_model_re <- lm(abs(residuals(wls_model)) ~ data$X3 + data$X4)
estimated_std_re <- sd_model_re$fitted.values
weight_re <- 1 / (estimated_std_re)^2
wls_model_re <- lm(Y ~ factor(data$Degree) + X3 + X4, data=data, weights=weight_re)
wls_model_re$coefficients
```

```
##          (Intercept) factor(data$Degree)2 factor(data$Degree)3
##            29.083233            11.007530            26.814222
##                   X3                   X4
##             1.490446             1.692183
```

In this iterated process, intercept and X2 show relatively large changes; all the rest are roughly the same, indicating that the coefficients are on the way to convergence. In this case, we should perform another iteration to converge even closer to the finest result.