PSTAT 126 HW 3

Tamjid Islam

4/29/2020

1. This problem uses the UN11 data in the alr4 package.

```r
# install.packages('alr4')
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

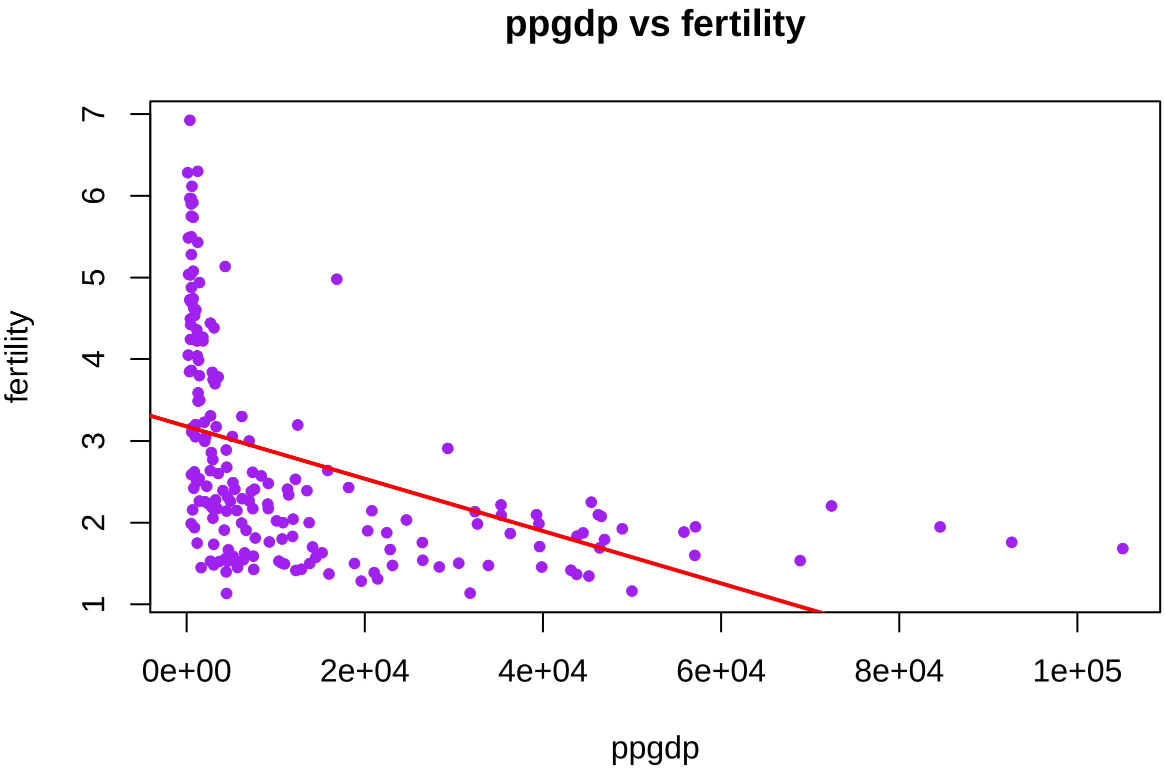
```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

(a) Plot fertility against ppgdp. Fit a linear model regressing fertility on ppgdp and add the fit on the plot. Comment on why this model is not good.

```r
library(alr4)
data(UN11)
plot(UN11$ppgdp, UN11$fertility, xlab = 'ppgdp', ylab = 'fertility', main =
       'ppgdp vs fertility', pch = 20, col = 'purple')
lm <- lm(UN11$fertility ~ UN11$ppgdp)
lm
```

```
##
## Call:
## lm(formula = UN11$fertility ~ UN11$ppgdp)
##
## Coefficients:
## (Intercept)   UN11$ppgdp
##   3.178e+00   -3.201e-05
```

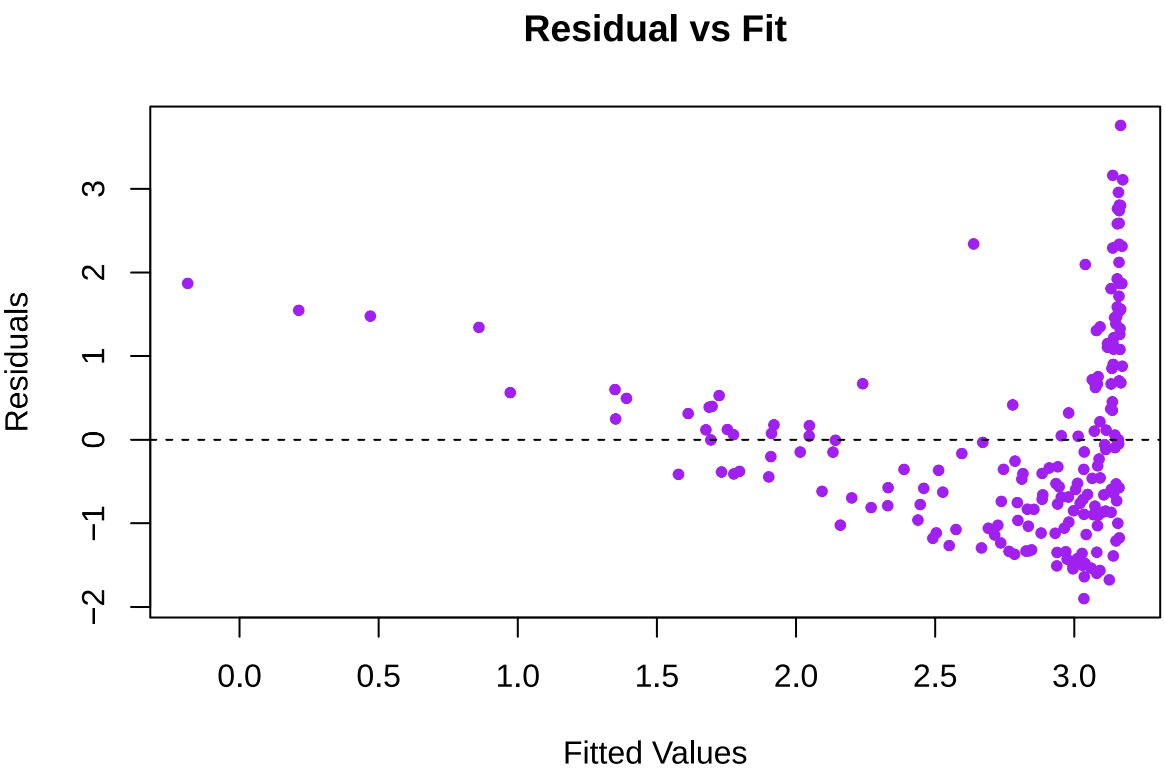
```
abline(lm, col = 'red', lwd = 2)
```



This is not a good model. It is not following the LINE properties when checking if the model is appropriate. From this gragh we can automatically fail the idea of being a good model by being able to tell that the population regression function is not linear but rather skewed right. In other words, majority of the data is clustered in one area.

(b) Use a "residuals vs fit" plot to check if there is any non-constant variance or non-linearity problem. State the main problem and explain why in one or two sentences.

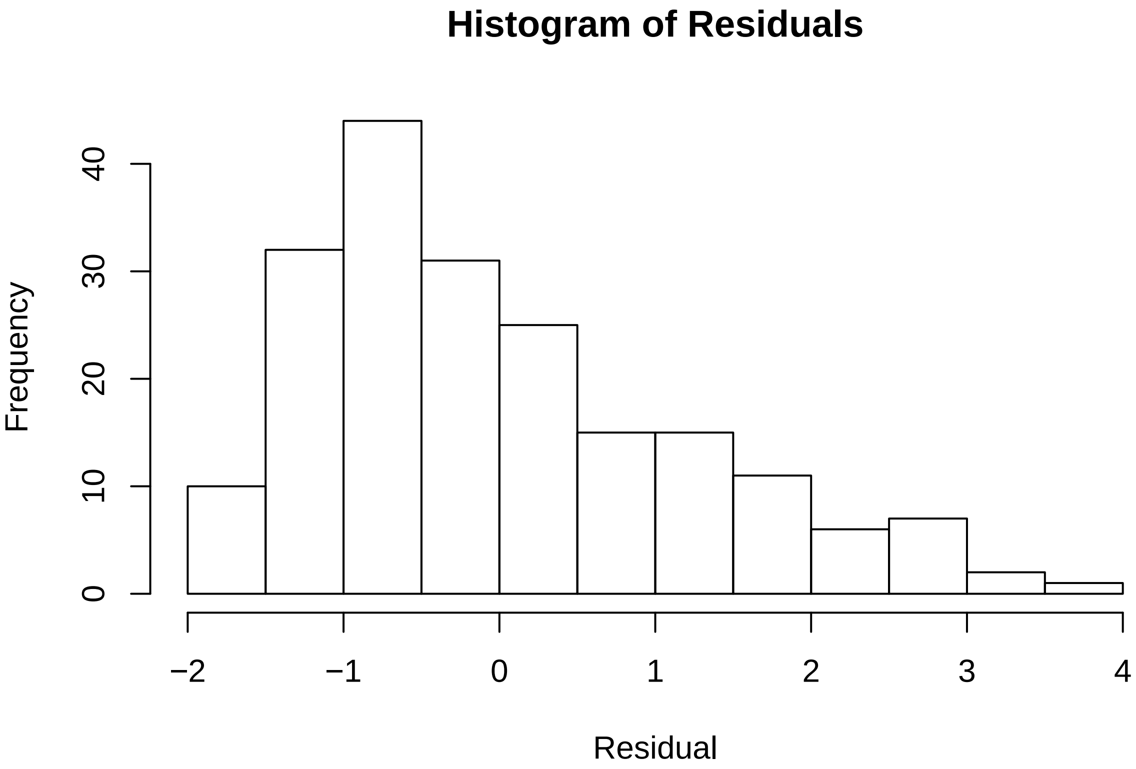
```
y_hat <- fitted(lm)
residual <- UN11$fertility - y_hat
plot(y_hat, residual, xlab = 'Fitted Values', ylab = 'Residuals', main = 'Residual vs Fit',
     pch = 20, col = 'purple')
abline(h = 0, lty = 2)
```



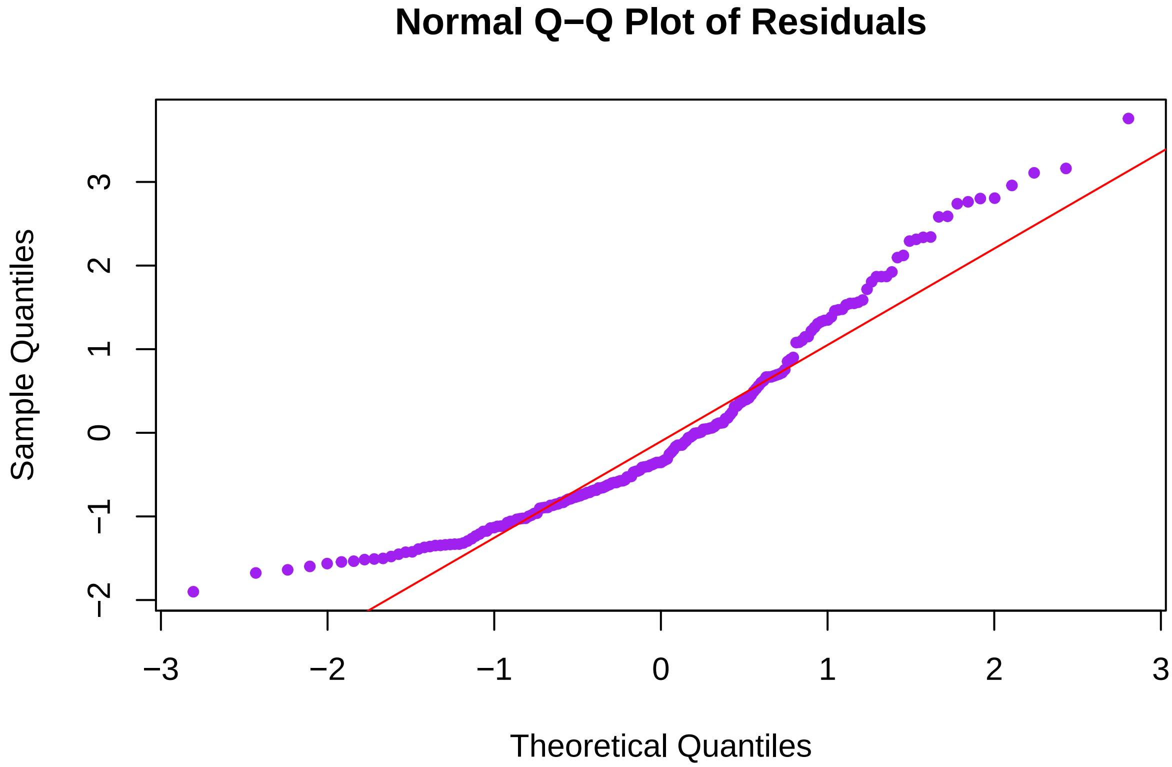
There is a non-linearity and a non- constant varience problem in this model. There are mainly several spreadout positive x values then a heavily negative cluster of x values and then a slight cluster of positive x values. Stated from before, we can tell that there is skewness issue, thus the model doesn't have normality.

(c) Use a normal Q-Q probability plot to check if the normality assumption is met. State the main problem and explain why in one or two sentences.

```
hist(residual, xlab = 'Residual', main = 'Histogram of Residuals')
```



```
qqnorm(residual, main = 'Normal Q-Q Plot of Residuals', pch = 20, col = 'purple')
qqline(residual, col = 'red')
```



The histogram or residuals shows that the residuals and error terms are skewed and thus not normally distributed. In the normal Q-Q plot, the relationship is far from being linear, which implies that the condition of the error terms are normally distributed is not met.

(d) Shapiro-Wilk test is a test of normality of a numeric variable. The null hypothesis for this test is that the variable is normally distributed. Use the R function shapiro.test() to test if the residuals of the linear fit in part (a) is normally distributed. State the p-value of this test and your conclusion given $\alpha = 0.05$. Does the result support your conclusion in part (c)? (Use the code ?shapiro.test or help(shapiro.test) to understand how to use this function.)

```
shapiro.test(residual)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residual
## W = 0.92844, p-value = 2.708e-08
```

P-value is 2.708e-08. The p-value is almost 0 which is less than ($\alpha = 0.05$.) Reject H_0. The residuals are not normally distrinuted. Thus the variable is not normally distributed which supports the conclusion in part (c).

2. This problem uses the teengamb data set in the faraway package. Fit a model with gamble as the response and the other variables as predictors.

```
# install.packages('faraway')
library(faraway)
```

```
##
## Attaching package: 'faraway'

## The following objects are masked from 'package:alr4':
##
##     cathedral, pipeline, twins

## The following objects are masked from 'package:car':
##
##     logit, vif
```

(a) Predict the amount that men with average (given the data) status, income and verbal score would gamble along with an appropriate 95% confidence interval for the mean amount.

```
library(faraway)
data(teengamb)
fit <- lm(gamble ~ sex + status + income + verbal, data = teengamb)
fit
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Coefficients:
## (Intercept)          sex       status       income       verbal
##    22.55565    -22.11833      0.05223      4.96198     -2.95949
```

```
predictors <- data.frame(sex=0, status = mean(teengamb$status), income = mean(teengamb$income),
                         verbal = mean(teengamb$verbal))
predictors
```

```
##   sex   status   income   verbal
## 1   0 45.23404 4.641915 6.659574
```

```
ci_95 <- predict(fit, predictors, se.fit = TRUE, interval = 'confidence', level = 0.95,
type = 'response')
ci_95
```

```
## $fit
##        fit      lwr      upr
## 1 28.24252 18.78277 37.70227
##
## $se.fit
## [1] 4.687496
##
## $df
## [1] 42
##
## $residual.scale
## [1] 22.69034
```

We are 95% confident that the mean predictors for the males are between 18.78277 and 37.70227 would gamble.

(b) Repeat the prediction for men with maximal values (for this data) of status, income and verbal score. Which confidence interval is wider and why is the result expected?

```
library(faraway)
data(teengamb)
fit <- lm(gamble ~ sex + status + income + verbal, data = teengamb)
fit
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Coefficients:
## (Intercept)          sex       status       income       verbal
##    22.55565    -22.11833      0.05223      4.96198     -2.95949
```

```
predictors <- data.frame(sex=0, status = max(teengamb$status), income = max(teengamb$income),
                         verbal = max(teengamb$verbal))
predictors
```

```
##   sex status income verbal
## 1   0     75     15     10
```

```
ci_95 <- predict(fit, predictors, se.fit = TRUE, interval = 'confidence', level = 0.95,
type = 'response')
ci_95
```

```
## $fit
##        fit      lwr      upr
## 1 71.30794 42.23237 100.3835
##
## $se.fit
## [1] 14.40753
##
## $df
## [1] 42
##
```

```
## $residual.scale
## [1] 22.69034
```

We are 95% confident that the max predictors for the males are between 42.23237 and 100.3835 would gamble. This confidence is wilder than the mean predictors because it makes sense that males with the best status, income and verbal would gamble more than the average. The max confidence interval takes the maximum in all predictor cases making it larger, where the mean confidence interval is much more condensed compared to this confidence interval by taking into account the average predictor values.

(c) Fit a model with sqrt(gamble) as the response but with the same predictors. Now predict the response and give a 95% prediction interval for an individual in (a). Take care to give your answer in the original units of the response.

```
library(faraway)
data(teengamb)
fit <- lm(sqrt(gamble) ~ sex + status + income + verbal, data = teengamb)
fit
```

```
##
## Call:
## lm(formula = sqrt(gamble) ~ sex + status + income + verbal, data = teengamb)
##
## Coefficients:
## (Intercept)          sex       status       income       verbal
##     2.97707     -2.04450      0.03688      0.47938     -0.42360
```

```
predictors <- data.frame(sex=0, status = mean(teengamb$status), income = mean(teengamb$income),
                         verbal = mean(teengamb$verbal))
predictors
```

```
##   sex   status   income   verbal
## 1   0 45.23404 4.641915 6.659574
```

```
ci_95 <- predict(fit, predictors, interval = 'prediction')
ci_95
```

```
##        fit       lwr      upr
## 1 4.049523 -0.245035 8.344082
```

```
ci_95^2
```

```
##        fit        lwr     upr
## 1 16.39864 0.06004216 69.6237
```

The 95% prediction interval from a is that the mean predictors for the males are between 0.06004216 and 69.6237 would gamble.

3. Using the sat data in the faraway package:

(a) Fit a model with total sat score as the response and expend and takers as predictors. Test the hypothesis that $\beta_{expend} = \beta_{takers} = 0$. Do any of the two predictors have an effect on the response?

$H_0 : \beta_{expend} = \beta_{takers} = 0$
$H_1 : \beta_{expend} \neq 0$ and/or $\beta_{takers} \neq 0$

```
library(faraway)
data(sat)
fit <- lm(total ~ expend + takers, sat)
fit
```

```
##
## Call:
## lm(formula = total ~ expend + takers, data = sat)
##
## Coefficients:
## (Intercept)       expend       takers
##     993.832       12.287       -2.851
```

```r
summary(fit)
```

```
##
## Call:
## lm(formula = total ~ expend + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.400 -22.884   1.968  19.142  68.755
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 993.8317    21.8332  45.519  < 2e-16 ***
## expend       12.2865     4.2243   2.909  0.00553 **
## takers       -2.8509     0.2151 -13.253  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.46 on 47 degrees of freedom
## Multiple R-squared:  0.8195, Adjusted R-squared:  0.8118
## F-statistic: 106.7 on 2 and 47 DF,  p-value: < 2.2e-16
```

Since the p-value for both expend and takers are less than α, we reject the H_0. Therefore the null, $\beta_{expend} = \beta_{takers} = 0$, is not true. Thus, the two predictors have an affect on the response and they are both statistically significant.

4. This problem uses the trade.union data in the SemiPar package.

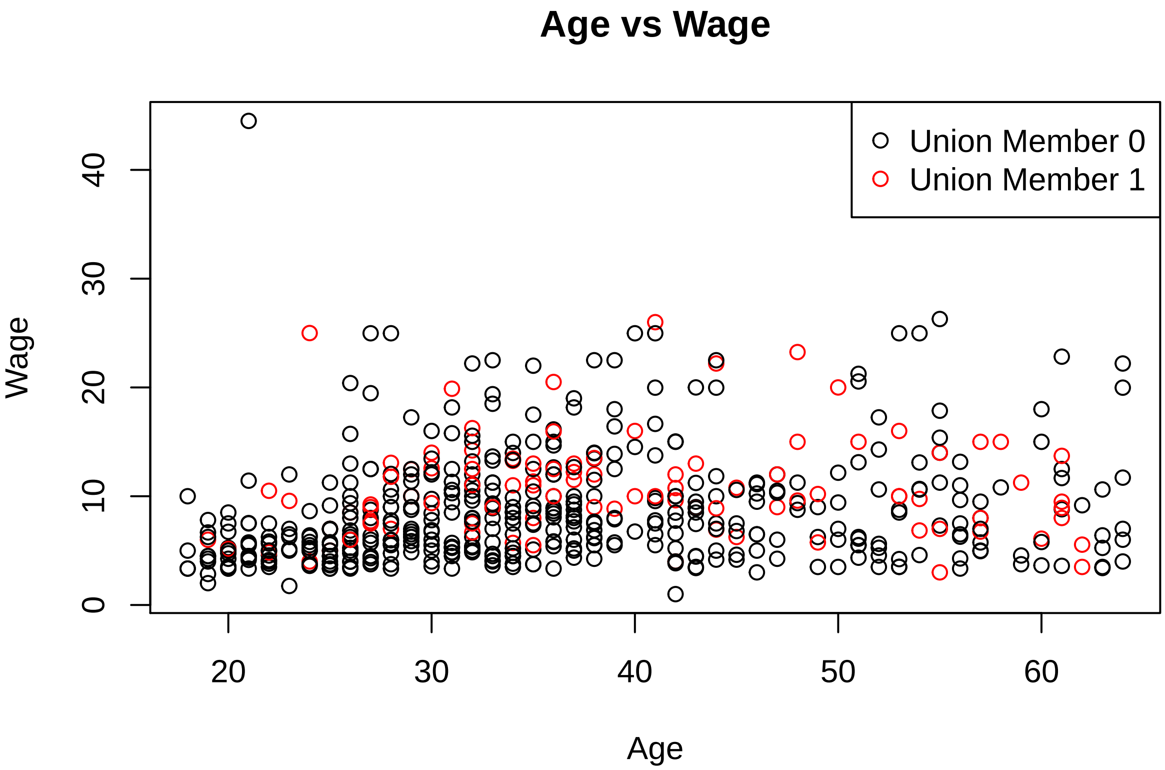
```r
# install.packages('SemiPar')
library(SemiPar)
```

```
##
## Attaching package: 'SemiPar'

## The following object is masked from 'package:car':
##
##     spm
```

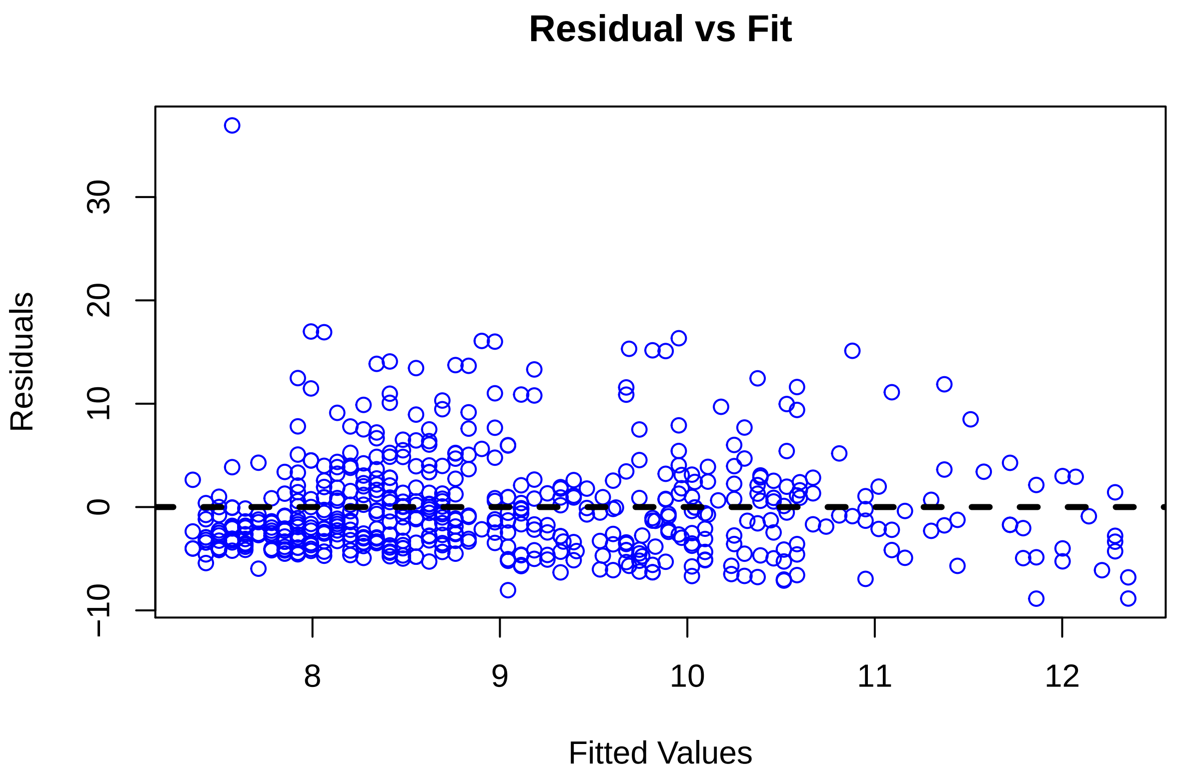
(a) Plot the wage as a function of age using a different plotting symbol for the different union membership of the world.

```r
library(SemiPar)
data(trade.union)
plot(trade.union$age,trade.union$wage, xlab = 'Age', ylab = 'Wage', main = 'Age vs Wage',
     col = trade.union$union.member+1, pch = 1)
legend('topright', legend = paste('Union Member', 0:1), col = 1:2, pch = 1, bty = 'o')
```

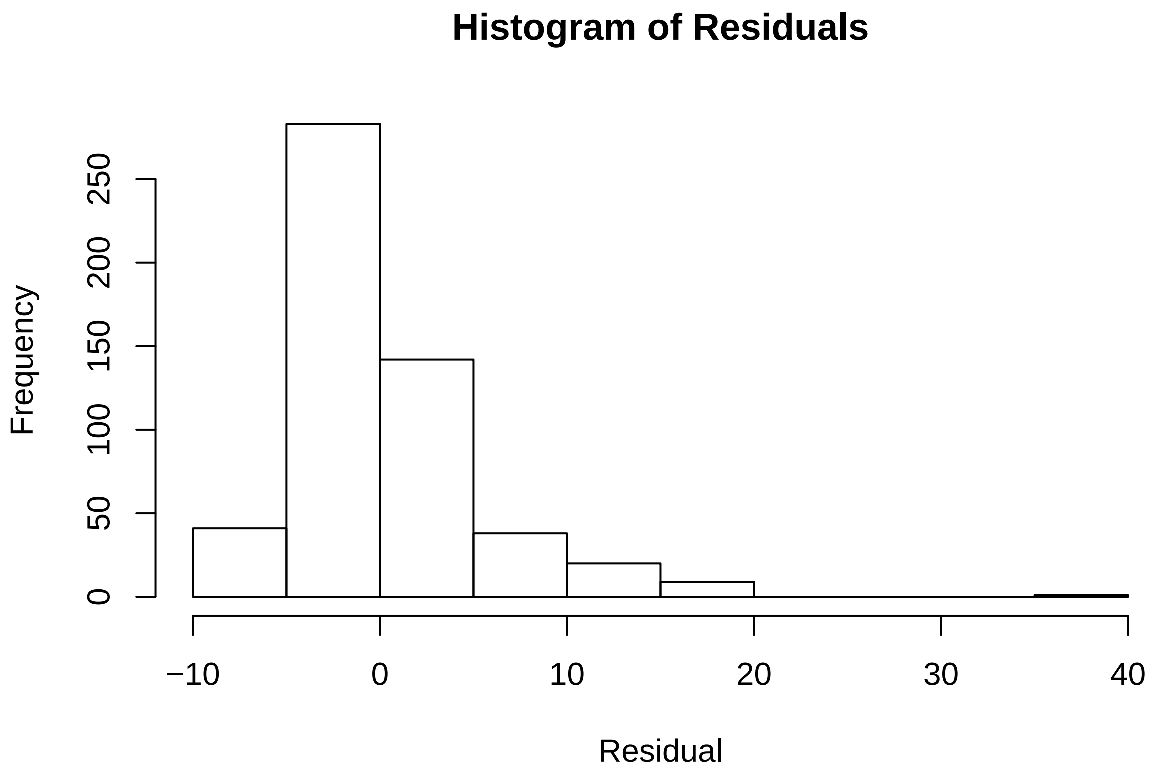


(b) Determine a transformation on the response wage to facilitate linear modeling with age and union membership as predictors.

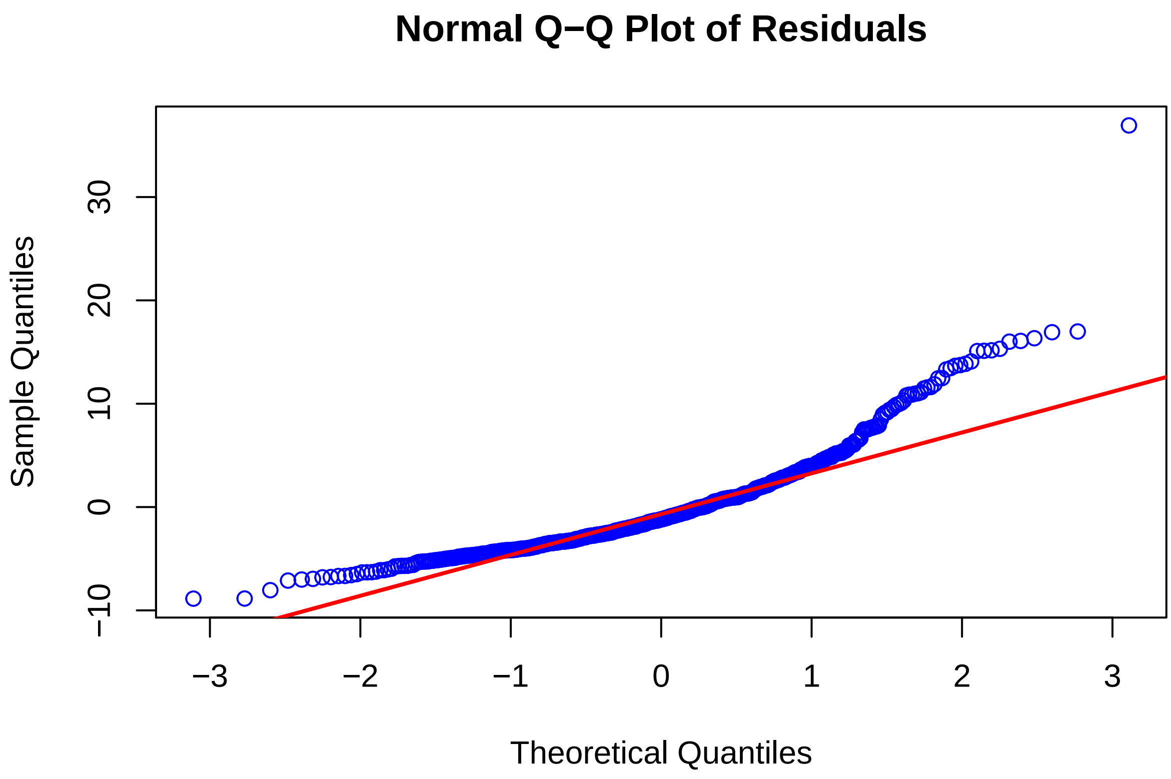
```r
fit <- lm(wage ~ age + union.member, data = trade.union)
yhat <- fitted(fit)
residual <- trade.union$wage - yhat
plot(yhat, residual, xlab = 'Fitted Values', ylab = 'Residuals', main = 'Residual vs Fit',
     pch = 1, col = 'blue')
abline(h = 0, lty = 2, lwd = 3)
```



```
hist(residual, xlab = 'Residual', main = 'Histogram of Residuals')
```



```
qqnorm(residual, main = 'Normal Q-Q Plot of Residuals', pch = 1, col = 'blue')
qqline(residual, col = 'red', lwd = 2)
```



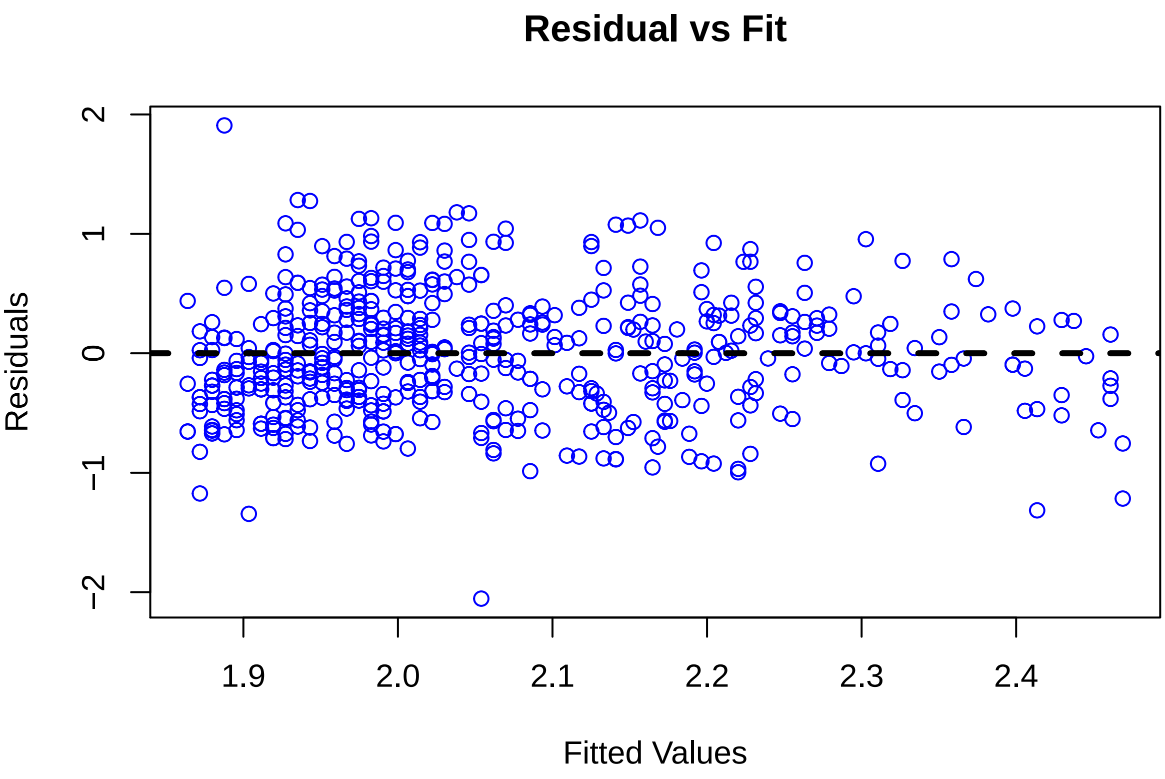
There is a non-linearity and an unequal varience issue in this problem. Therefore, the transformation needed is to take the log of the wage variable to resolve both issues.

Therefore, we make our transformation.

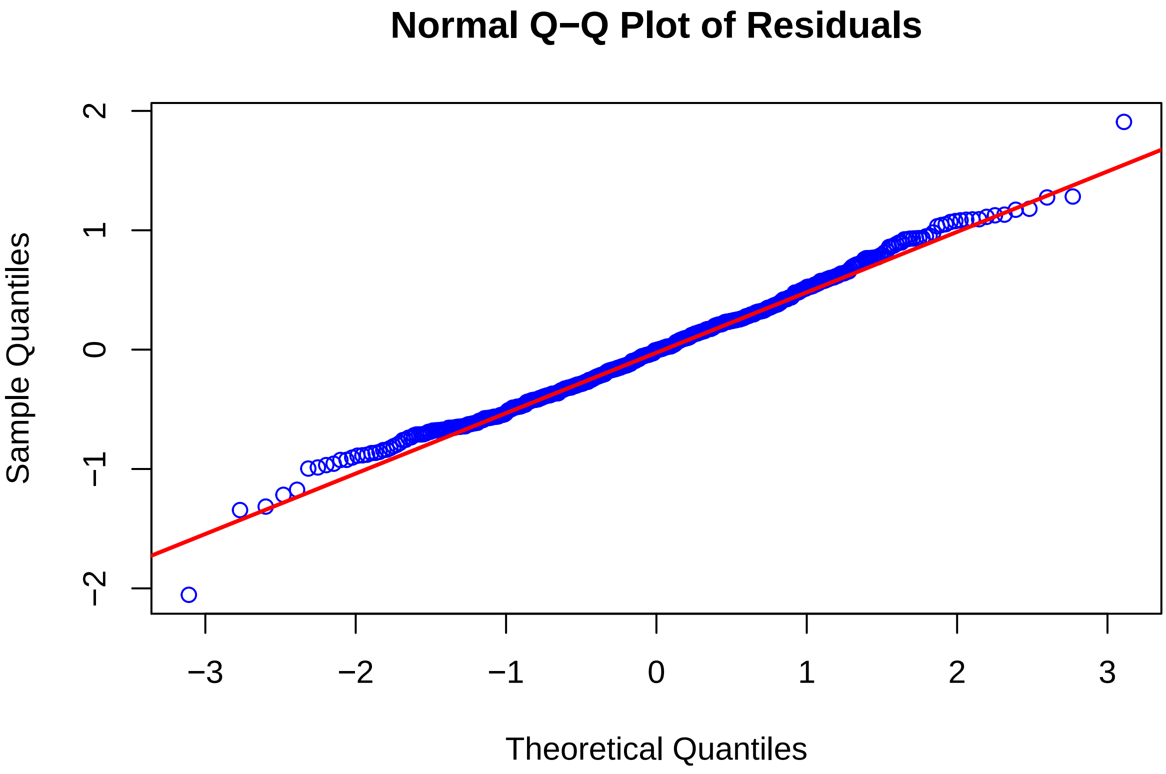
```
fit <- lm(log(wage) ~ age + union.member, data = trade.union)
fit
```

```
##
## Call:
## lm(formula = log(wage) ~ age + union.member, data = trade.union)
##
## Coefficients:
##  (Intercept)           age  union.member
##     1.721477      0.007915      0.256765
```

```
yhat <- fitted(fit)
residual <- log(trade.union$wage) - yhat
plot(yhat, residual, xlab = 'Fitted Values', ylab = 'Residuals', main = 'Residual vs Fit',
     pch = 1, col = 'blue')
abline(h = 0, lty = 2, lwd = 3)
```



```r
qqnorm(residual, main = 'Normal Q-Q Plot of Residuals', pch = 1, col = 'blue')
qqline(residual, col = 'red', lwd = 2)
```



With the log transformation on wage, our linear model is fixed.

(c) Fit a linear model regressing transformed wage on age and union membership. What is

the relationship of age and union membership to wage?

```r
fit <- lm(log(wage) ~ age + union.member, data = trade.union)
fit
```

```
##
## Call:
## lm(formula = log(wage) ~ age + union.member, data = trade.union)
##
## Coefficients:
##  (Intercept)           age  union.member
##     1.721477      0.007915      0.256765
```

```r
summary(fit)
```

```
##
## Call:
## lm(formula = log(wage) ~ age + union.member, data = trade.union)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05393 -0.36727 -0.00407  0.31559  1.90779
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.721477   0.072678  23.686  < 2e-16 ***
## age          0.007915   0.001893   4.181 3.39e-05 ***
## union.member 0.256765   0.057759   4.445 1.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5089 on 531 degrees of freedom
## Multiple R-squared:  0.07376,    Adjusted R-squared:  0.07028
## F-statistic: 21.14 on 2 and 531 DF,  p-value: 1.461e-09
```

There is a positive linear relationship between the the two predictors (age and union.member) and the response(log(wage)). The p-values are small enough for the predictors to have an effect on the response.

5.The data below shows, for a consumer finance company operating in six cities, the number of competing loan companies operating in the city (X) and the number per thousand of the company's loans made in that city that are currently delinquent (Y):

i :	1	2	3	4	5	6
X_i :	4	1	2	3	3	4
Y_i :	16	5	10	15	13	22

Assume that a simple linear regression model is applicable. Using matrix methods, find

a) Y'Y

$$Y'Y = \begin{pmatrix} Y_1 & Y_2 \cdots & Y_n \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 16 & 5 & 10 & 15 & 13 & 22 \end{pmatrix} \begin{pmatrix} 16 \\ 5 \\ 10 \\ 15 \\ 13 \\ 22 \end{pmatrix} = \begin{pmatrix} 1259 \end{pmatrix}$$

a) X'X

$$X'X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} = \begin{pmatrix} 6 & 17 \\ 17 & 55 \end{pmatrix}$$

a) X'Y

$$X'Y = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{pmatrix} = \begin{pmatrix} 81 \\ 261 \end{pmatrix}$$

a) b_0 and b_1

$$b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = (X'X)^{-1}X'Y = \begin{pmatrix} 6 & 17 \\ 17 & 55 \end{pmatrix}^{-1} \begin{pmatrix} 81 \\ 261 \end{pmatrix} = \begin{pmatrix} 55/41 & -17/41 \\ -17/41 & 6/41 \end{pmatrix} = \begin{pmatrix} 18/41 \\ 189/41 \end{pmatrix}$$

b_0 is $18/41 = 0.4390244$ and b_1 is $189/41 = 4.609756$

6.Briefly describe the dataset you would be using for your project. Give its source also. Then write down what the response is and a few important independent variables which you think should be included in the analysis.

The dataset that we will be using is QSAR Fish Toxicity from the UCI Machine Learning Respitory. The data set comes from the "qsar_fish_toxicity.csv" file. The response is LC50 [-LOG(mol/L)] which is the toxicity levels and a few independent variables that will be used in our analysis is are 6 molecular descriptors (CIC0, SM1_Dz(Z), GATS1i, NdsCH, NdssC, MLOGP). We feel like majority of the data will be significant but only when we do our tests we will know what is.

Source: https://archive.ics.uci.edu/ml/datasets/QSAR+fish+toxicity