PSTAT 126 Homework 3

Shaiyon Hariri

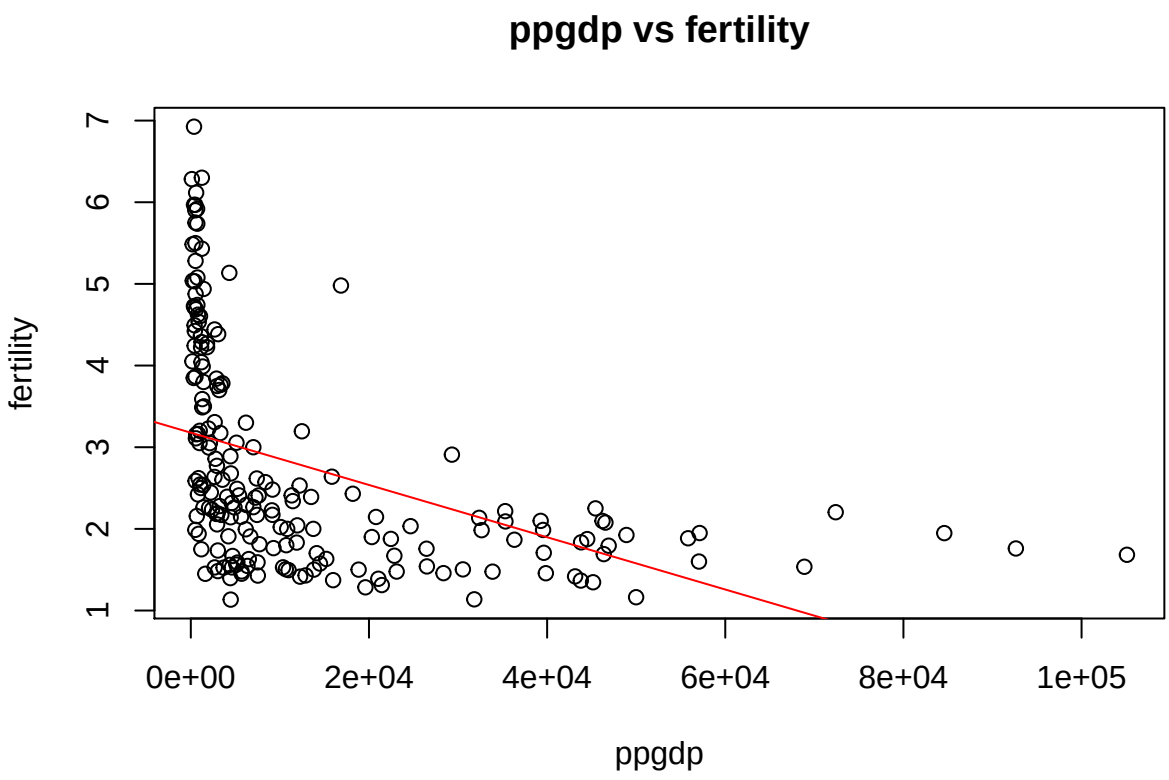
4/29/2020

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

```
library(alr4)
```

(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.

```
plot(UN11$ppgdp, UN11$fertility, main="ppgdp vs fertility", xlab="ppgdp", ylab="fertility")

model <- lm(UN11$fertility ~ UN11$ppgdp)
abline(model, col="red")
```

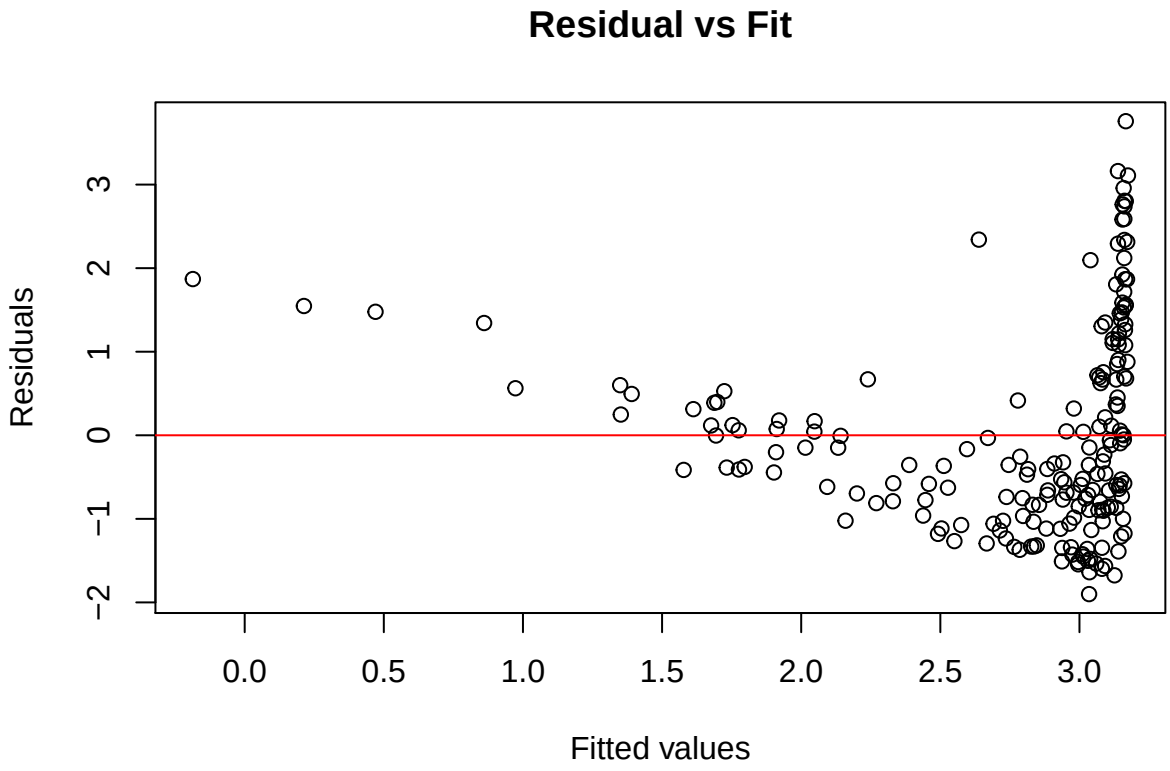


This model might capture somewhat that as ppgdb rises, fertility rates fall, but the data fails to be normally distributed around the line, leading to poor predictive power. Additionally, the model fails to capture how

extreme high fertility is correlated with low ppgdp.

(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.

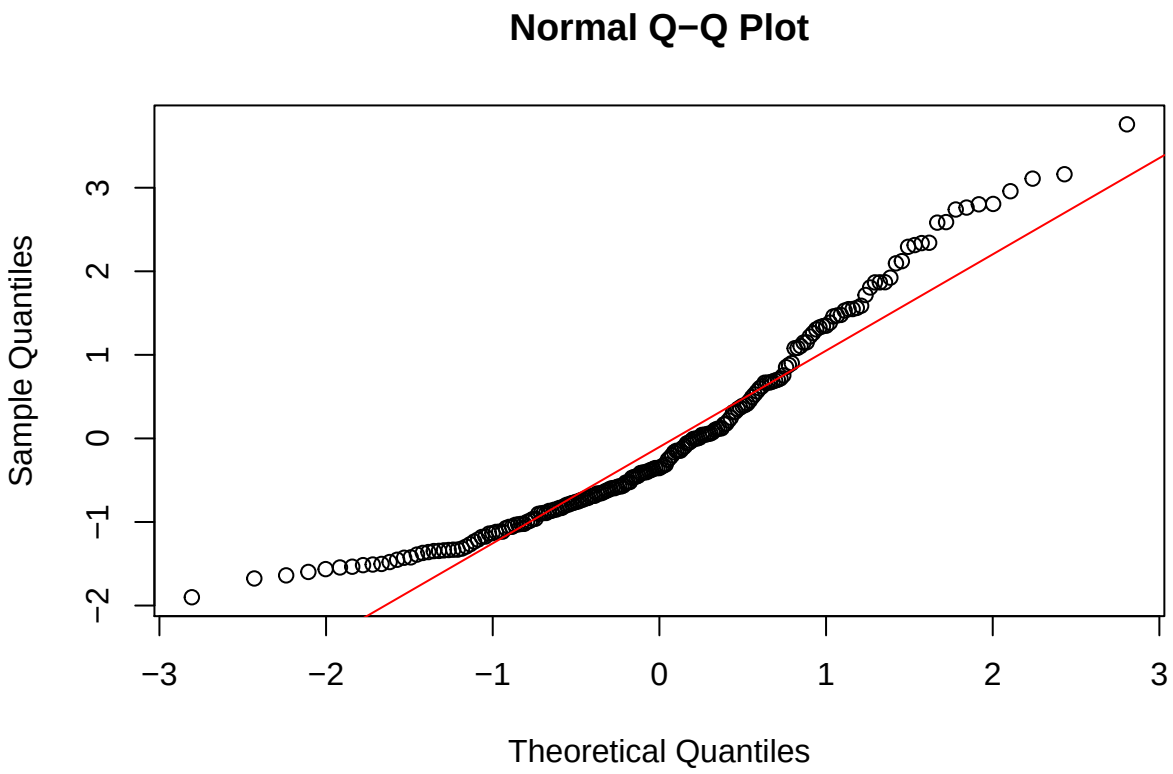
```
plot(model$fitted.values, model$residuals, main="Residual vs Fit", ylab="Residuals",
     xlab="Fitted values")
abline(0,0, col="red")
```



The main problem is that the data is simply not distributed normally around the regression line, so the residuals are all over the place. The residuals are large positive values for small fitted values, seem relatively normally distributed for medium sized values, and have extremely high variance for large fitted values. The variance in residuals does not stay constant, suggesting extreme non-linearity.

(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.

```
qqnorm(model$residuals)
qqline(model$residuals, col="red")
```



It's easy to see in the Q-Q plot that the points do not fall along the diagonal, and have a clear curved pattern, with the tails fanning out considerably away from the line. Thus, the normality assumption 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. State the p-value of this test and your conclusion given alpha = 0.05. Does the result support your conclusion in part (c)?

```
p <- shapiro.test(model$residuals)$p.value

cat("p-value: ", format(p, scientific = FALSE))
```

```
## p-value:  0.00000002708383
```

The p-value from the shapiro test is very very small considering a significance level of 0.05, implying that the distribution of the data is not normal. This supports the observation made in part (c) of this question.

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.

```
library(faraway)

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

(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.

```
preds <- data.frame(sex=0, status=mean(teengamb$status), income=mean(teengamb$income),
                    verbal=mean(teengamb$verbal))
predict(model, preds, interval="predict", level=0.95)
```

```
##        fit       lwr      upr
## 1 28.24252 -18.51536 75.00039
```

(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?

```
preds <- data.frame(sex=0, status=max(teengamb$status), income=max(teengamb$income),
                    verbal=max(teengamb$verbal))
predict(model, preds, interval="predict", level=0.95)
```

```
##        fit      lwr    upr
## 1 71.30794 17.06588 125.55
```

The confidence interval for men with maximal values is slightly wider. It makes sense that teenage males with higher socioeconomic status and income would gamble more than those without the same means, so the fit and confidence interval containing larger values is to be expected.

(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.

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

preds <- data.frame(sex=0, status=mean(teengamb$status), income=mean(teengamb$income),
                    verbal=mean(teengamb$verbal))
predict(model, preds, interval="predict", level=0.95)^2
```

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

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 Bexpend = Btakers = 0. Do any of the two predictors have an effect on the response?

Ho : Bexpend = Btakers = 0

Ha : Bexpend != 0 and/or Bexpend != 0

```
model <- lm(total ~ expend + takers, data=sat)

summary(model)
```

```
##
```

```
## 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
```

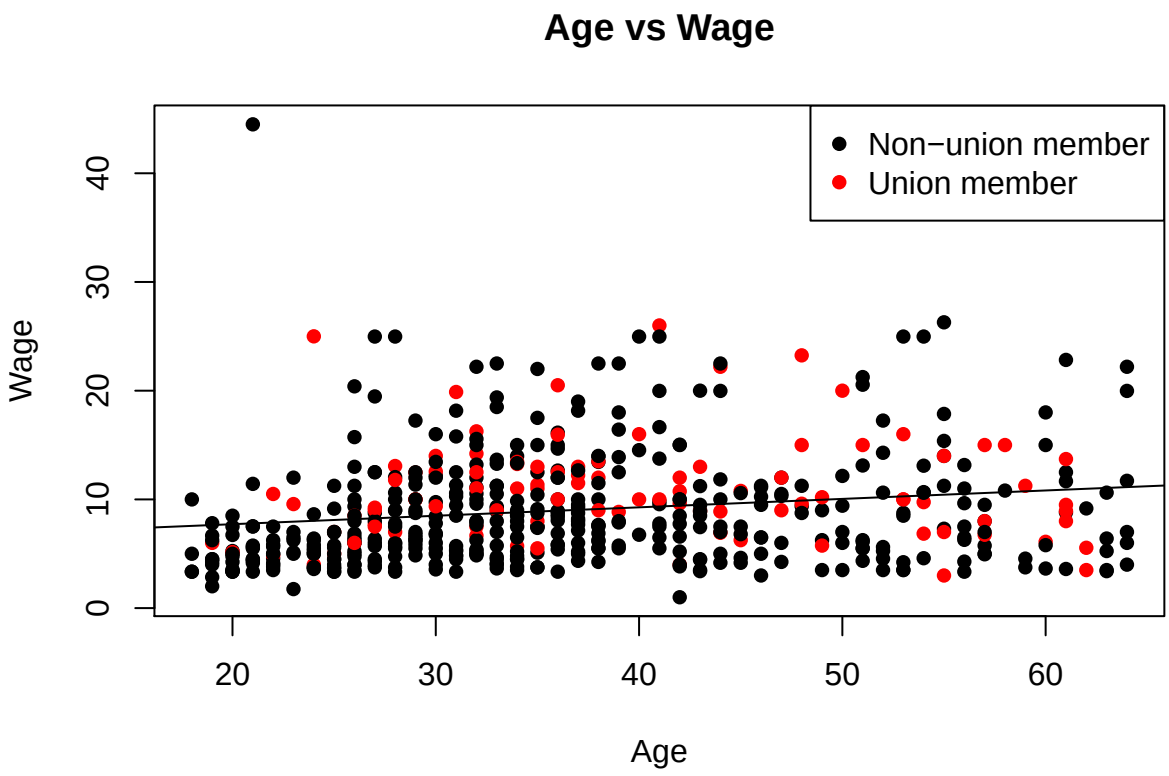
The p values for the coefficients are small, meaning it's safe to say that we can reject the null hypothesis that Bexpend = Btakers = 0, and that the two predictors have an effect on the response.

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

```
library("SemiPar")
```

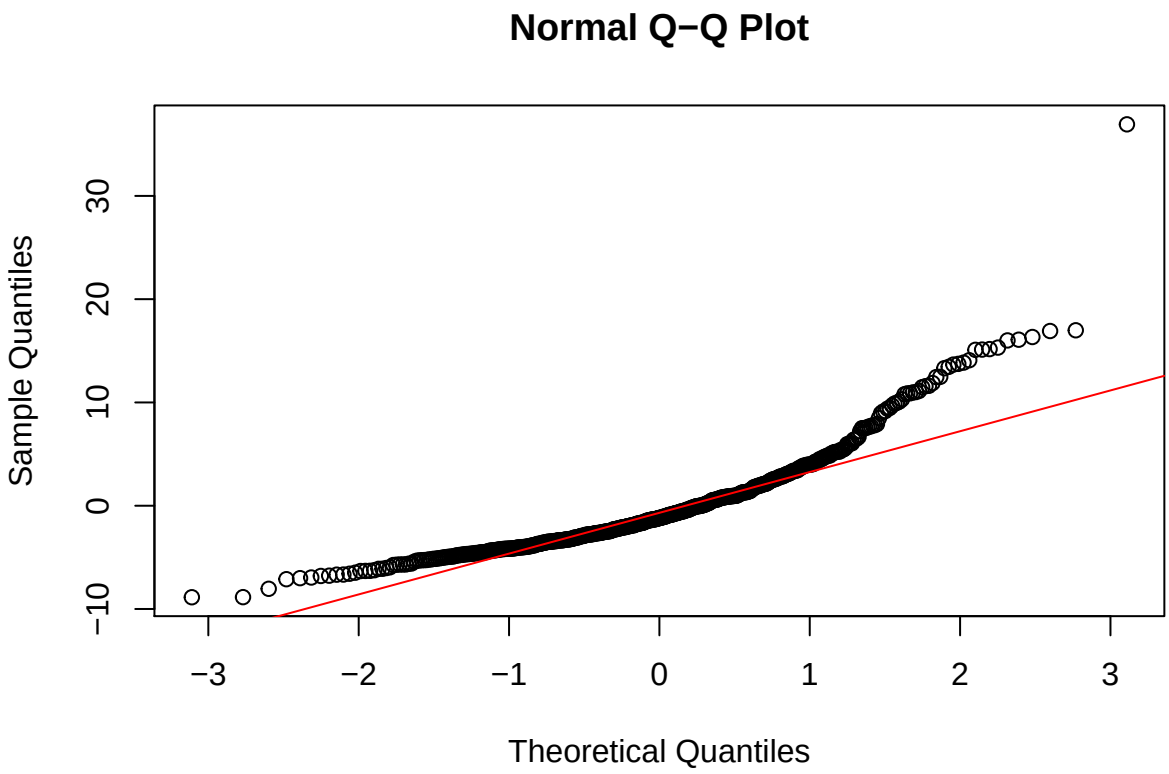
(a) Plot the wage as a function of age using a different plotting symbol for the different union membership of the world.

```
plot(trade.union$age, trade.union$wage,
     col=trade.union$union.member+1, pch=16,
     main="Age vs Wage", xlab="Age", ylab="Wage")
legend('topright', legend=c("Non-union member", "Union member"), col=1:2, pch=16)
abline(lm(trade.union$wage~trade.union$age))
```



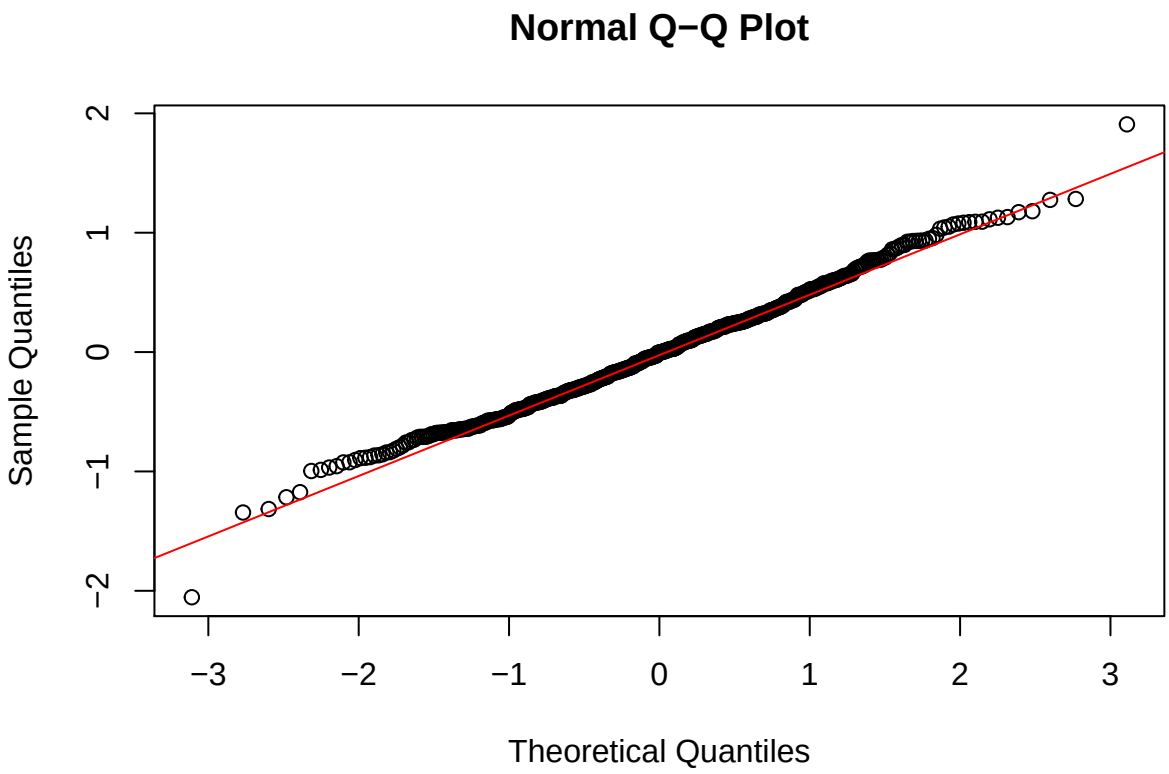
(b) Determine a transformation on the response wage to facilitate linear modeling with age and union membership as predictors.

```
untransformed <- lm(trade.union$wage ~ trade.union$age + trade.union$union.member)

qqnorm(untransformed$residuals)
qqline(untransformed$residuals, col="red")
```



The points on the Q-Q plot deviate far from the reference line, therefore the normality assumption doesn't hold and a linear regression would not give interpretable results. A transformation on our response variable here could greatly increase a linear model's effectiveness.

```
transformed <- lm(log(trade.union$wage) ~ trade.union$age + trade.union$union.member)

qqnorm(transformed$residuals)
qqline(transformed$residuals, col="red")
```

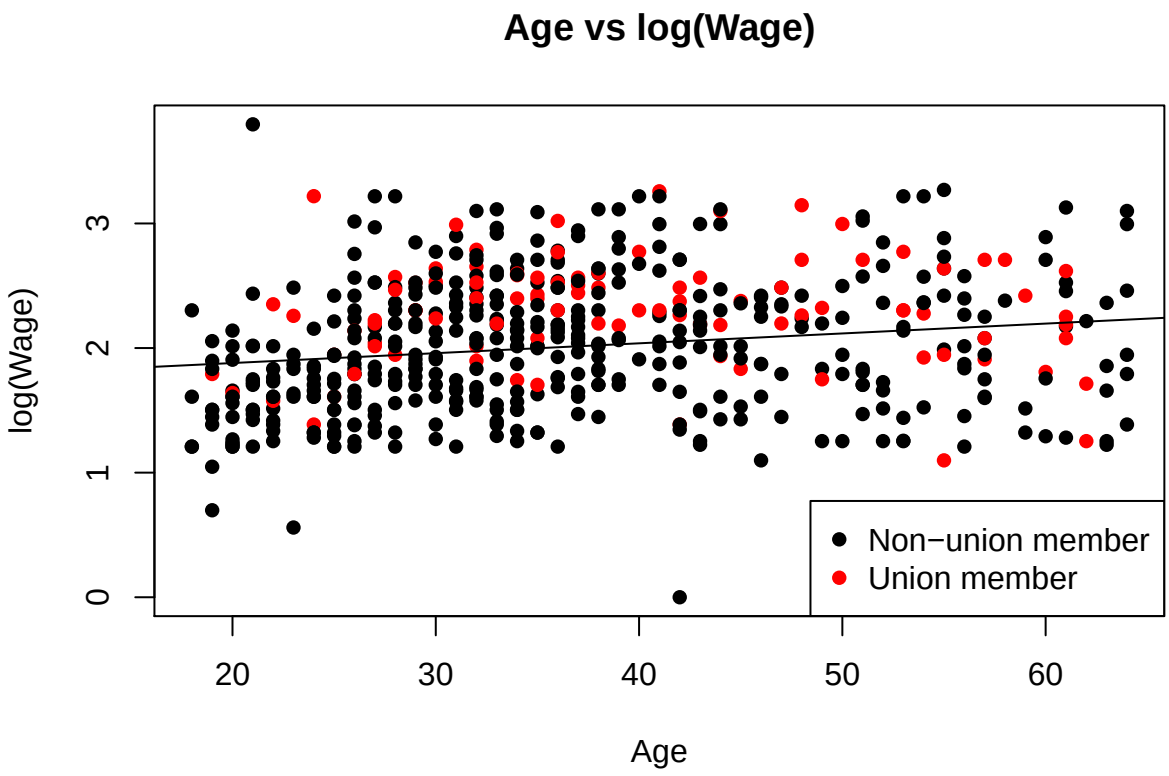


The plot when wage is transformed with a natural logarithm shows a massive improvement from the previous one. The points hug the reference line, letting us assume normality. Thus, this transformation will improve our results when fitting a regression line.

(c) Fit a linear model regressing transformed wage on age and union membership. What is the relationship of age and union membership to wage?

```
model <- transformed

plot(trade.union$age, log(trade.union$wage),
     col=trade.union$union.member+1, pch=16,
     main="Age vs log(Wage)", xlab="Age", ylab="log(Wage)")
legend('bottomright', legend=c("Non-union member", "Union member"), col=1:2, pch=16)
abline(model)
```

```
## Warning in abline(model): only using the first two of 3 regression coefficients
```




```
summary(model)
```

```
##
## Call:
## lm(formula = log(trade.union$wage) ~ trade.union$age + trade.union$union.member)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05393 -0.36727 -0.00407  0.31559  1.90779
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.721477   0.072678  23.686  < 2e-16 ***
## trade.union$age           0.007915   0.001893   4.181 3.39e-05 ***
## trade.union$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
```

According to the data, both age and being a union member are correlated with a higher wage, as the coefficients are positive. The p-values are small enough that it's safe to assume that the predictors have an effect on the response.

(d) State the null and alternative hypotheses for the overall F-test for the model in (c). Perform the test and summarize results.

Ho : Bage = Bunion.member = 0

Ha : Bage != 0 and/or Bunion.member != 0

```
p <- format(anova(model)$Pr[1:2], scientific=FALSE)

cat("Bage p value:", p[1] , "\nBunion.member p value:", p[2])
```

```
## Bage p value: 0.00000266934
## Bunion.member p value: 0.00001068375
```

The p values for the coefficients are extremely small, meaning it's safe to say that we can reject the null hypothesis that Bage = Bunion.member = 0, and that the two predictors 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):

```
n <- 6
Xi <- c(4, 1, 2, 3, 3, 4)
Yi <- c(16, 5, 10, 15, 13, 22)
```

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

a) Y'Y

```
Y <- matrix(Yi, ncol=1)
tYY <- t(Y) %*% Y
tYY
```

```
##      [,1]
## [1,] 1259
```

b) X'X

```
tXX <- matrix(c(n, sum(Xi), sum(Xi), sum(Xi^2)), nrow=2, ncol=2)
tXX
```

```
##      [,1] [,2]
## [1,]    6   17
## [2,]   17   55
```

c) X'Y

```
tXY <- matrix(c(sum(Yi), sum(Xi*Yi)), ncol=1)
tXY
```

```
##      [,1]
## [1,]   81
```

```
## [2,]  261
```

d) b0 and b1

```
b <- solve(tXX) %*% tXY
cat("b0:", b[1], "  b1:", b[2])
```

```
## b0: 0.4390244   b1: 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.

We are using the QSAR fish toxicity dataset, which contains 7 variables. There are 6 molecular descriptors (CIC0, SM1_Dz(Z), GATS1i, NdsCH, NdssC, MLOGP) and the response is LC50. All of the data is numeric. I think all or most of the variables will end up being important enough to include in the analysis, but only statistical testing can tell us exactly which ones.

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