HW3 Exercise 3

Wen Fan(wf2255), Banruo Xie(bx2168), Hanjun Li(hl3339)

4/13/2020

```r
library(readr)
library(MASS)
CpsWages <- read_table2("~/Documents/Columbia/STAT W5703/HW/HW3/CpsWages.txt")
```

```
## Parsed with column specification:
## cols(
##   education = col_double(),
##   south = col_double(),
##   sex = col_double(),
##   experience = col_double(),
##   union = col_double(),
##   wage = col_double(),
##   age = col_double(),
##   race = col_double(),
##   occupation = col_double(),
##   sector = col_double(),
##   marr = col_double()
## )
```

```r
names <- c('sex', 'race', 'marr', 'occupation', 'sector', 'south', 'union')
CpsWages[, names] <- as.data.frame(sapply(CpsWages[, names], as.factor)) # make some of the variables a
```

Problem 1

We could use a multiple linear regression to examine this dataset. That is, use wage as our target variables and all else to be the predictors. It is not a good idea to include age, education, and experience at the same time because those variables are highly correlated. i.e. as a person ages, one tend to have higher education and more experience; thus, it might lead to issues of collinearity.

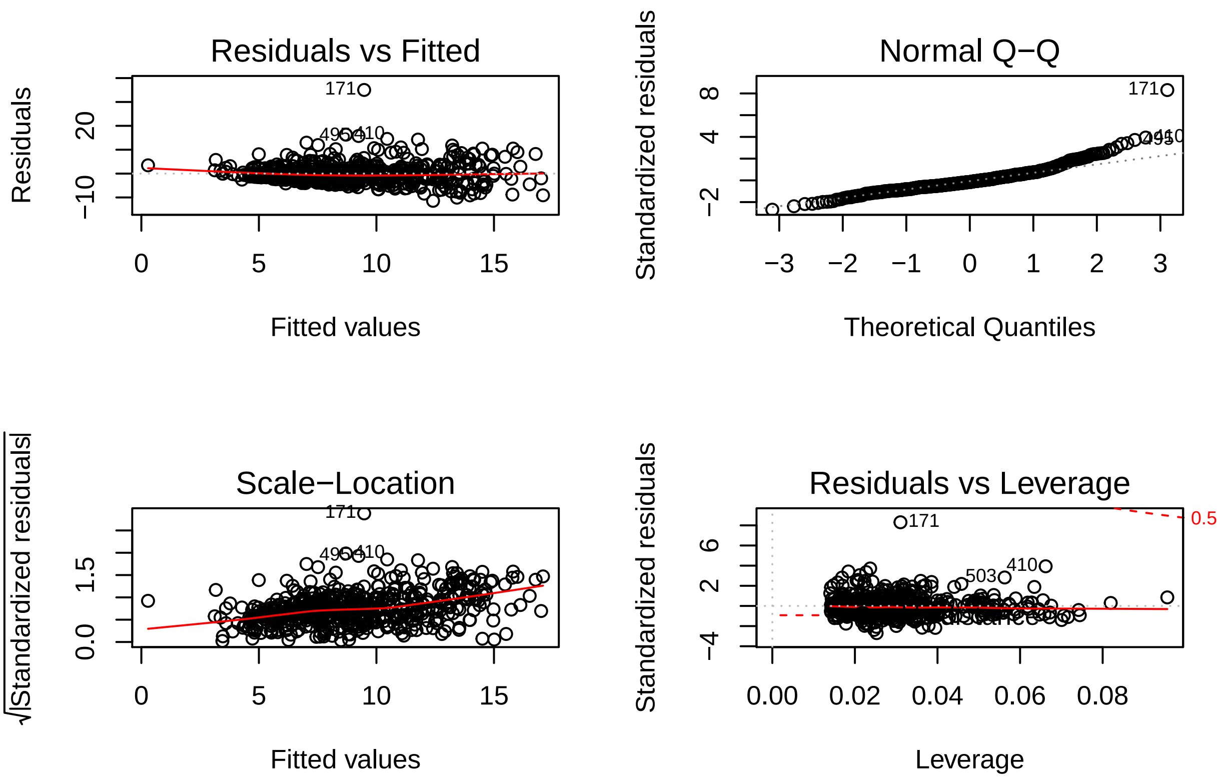
Problem 2

```r
m1 <- lm(wage ~ ., data = CpsWages)
summary(m1)
```

```
##
## Call:
## lm(formula = wage ~ ., data = CpsWages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.409  -2.486  -0.631   1.872  35.021
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2781     6.6976   0.340  0.73390
## education     0.8128     1.0869   0.748  0.45491
```

```
## south1        -0.5627     0.4198  -1.340  0.18070
## sex1          -1.9425     0.4194  -4.631 4.60e-06 ***
## experience     0.2448     1.0818   0.226  0.82103
## union1         1.6017     0.5127   3.124  0.00188 **
## age           -0.1580     1.0809  -0.146  0.88382
## race2          0.2314     0.9915   0.233  0.81559
## race3          0.8379     0.5745   1.458  0.14532
## occupation2  -4.0638     0.9159  -4.437 1.12e-05 ***
## occupation3  -3.2682     0.7626  -4.286 2.17e-05 ***
## occupation4  -3.9754     0.8108  -4.903 1.26e-06 ***
## occupation5  -1.3336     0.7289  -1.829  0.06791 .
## occupation6  -3.2905     0.8005  -4.111 4.59e-05 ***
## sector1        1.0409     0.5492   1.895  0.05863 .
## sector2        0.4774     0.9661   0.494  0.62141
## marr1          0.3005     0.4112   0.731  0.46523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.282 on 517 degrees of freedom
## Multiple R-squared:  0.3265, Adjusted R-squared:  0.3056
## F-statistic: 15.66 on 16 and 517 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(m1)
```



From the plot of residual v.s. fitted value, it seems that there's a cone-shaped pattern as the fitted values get larger, so it doesn't agree with the hypothesis of homoscedasticity. From the normal QQ-plot, although there are some points depart from the QQ-line, we see that most of the points are aligned with the normal QQ-line, so we can say that the hypothesis of normality is generally met.

Problem 3

Using $\alpha = 0.05$, it appears that only sex, union, and occupation are statistically significant according to the associated p-values from the model. To test if sector is significant, we can again use the associated p-value calculated by t-test from the model. Using $\alpha = 0.05$, we fail to reject the null hypothesis as the p-value for sector level 1 and 2 are both greater than 0.05.

Problem 4

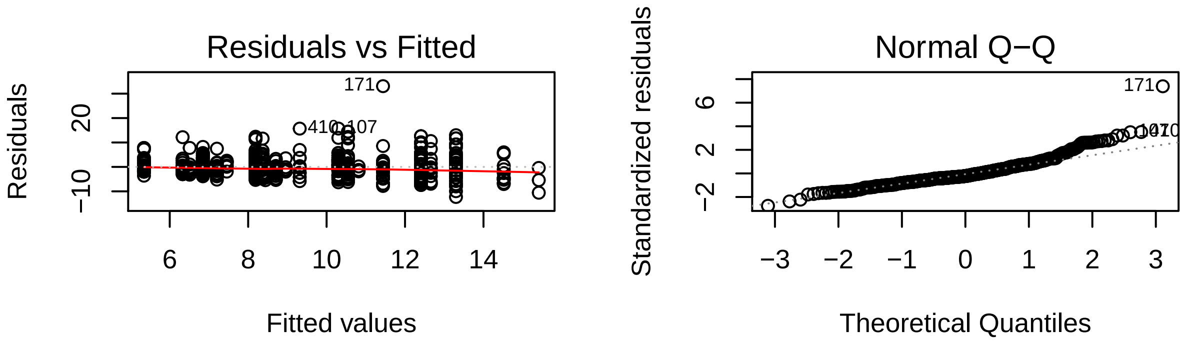
Since we see that only sex, union, and occupation are statistically significant from the model, we can fit a simpler model using only those three variables as predictors. We can also implement methods like AIC or BIC to reduce our model as some varaibles that do not appear to be significant might be significant in the reduced model.

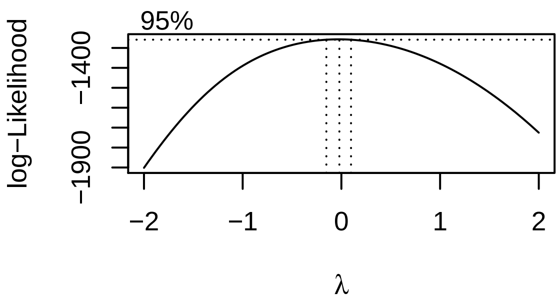
Problem 5

```r
m2 <- lm(wage ~ sex+union+occupation, data = CpsWages)
summary(m2)
```

```
##
## Call:
## lm(formula = wage ~ sex + union + occupation, data = CpsWages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.299  -2.700  -0.993   2.156  33.061
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.2991     0.6333  20.998  < 2e-16 ***
## sex1         -1.8602     0.4349  -4.278 2.25e-05 ***
## union1        2.1114     0.5279   3.999 7.26e-05 ***
## occupation2  -4.9298     0.9543  -5.166 3.40e-07 ***
## occupation3  -4.5932     0.7834  -5.863 8.03e-09 ***
## occupation4  -6.0959     0.7963  -7.655 9.29e-14 ***
## occupation5  -0.8929     0.7598  -1.175     0.24
## occupation6  -5.1104     0.7224  -7.075 4.81e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.521 on 526 degrees of freedom
## Multiple R-squared:  0.2361, Adjusted R-squared:  0.2259
## F-statistic: 23.23 on 7 and 526 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(m2, which=c(1,2))
boxcox(m2)
```



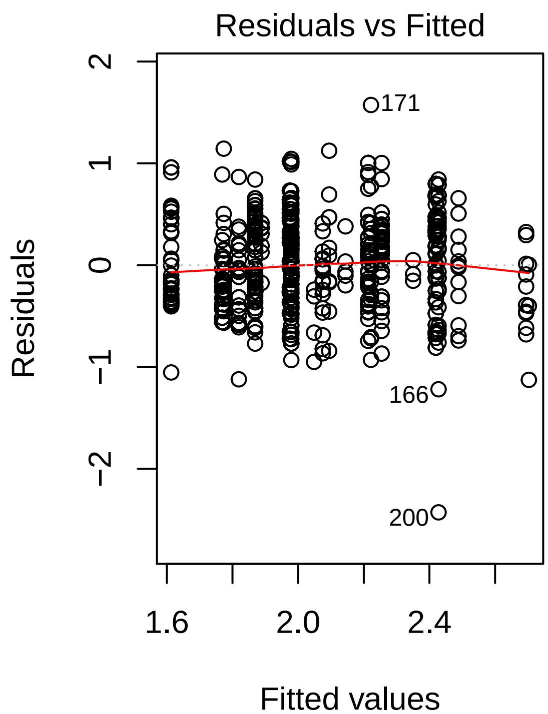


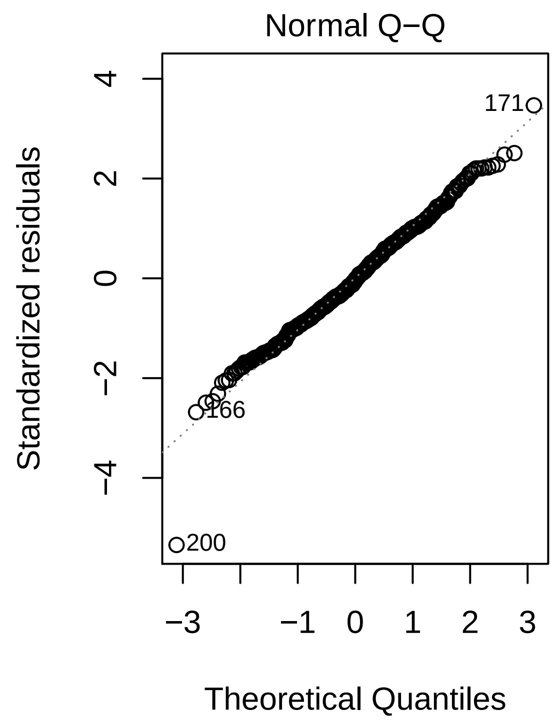
From the box-cox plot, we can see that $\lambda = 0$ lies inside the confidence interval, so we could apply log-transformation to transform the target variable wage into log(wage), and the resulting model is

```r
m3 <- lm(log(wage) ~ sex+union+occupation, data = CpsWages)
summary(m3)
```

```
##
## Call:
## lm(formula = log(wage) ~ sex + union + occupation, data = CpsWages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.42777 -0.31519 -0.01712  0.32494  1.57364
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.427768   0.064278  37.770  < 2e-16 ***
## sex1        -0.205919   0.044135  -4.666 3.91e-06 ***
## union1       0.275205   0.053578   5.136 3.95e-07 ***
## occupation2 -0.453088   0.096847  -4.678 3.68e-06 ***
## occupation3 -0.352769   0.079508  -4.437 1.11e-05 ***
## occupation4 -0.608505   0.080814  -7.530 2.23e-13 ***
## occupation5 -0.008335   0.077111  -0.108    0.914
## occupation6 -0.448551   0.073312  -6.118 1.85e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4589 on 526 degrees of freedom
## Multiple R-squared:  0.2539, Adjusted R-squared:  0.244
## F-statistic: 25.57 on 7 and 526 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(1,2))
plot(m3, which=c(1,2))
```





We can see that the cone-shaped pattern does not appear in the residual vs fitted plot, meaning homoscedasticity is met; also, there are only 2 points not aligned with the QQ-line, meaning the normality condition is met. In addition, all variables are statistically significant from the model summary; hence, this simplified model is appropriate.

Problem 6

As we can see from the QQ-plot in problem 5, point 171 and 200 appear to be outliers. Removing those two points could somehow improve our model, but it would not alter our conclusion. As shown below, sex, union, and occupation are still significant.

```r
summary(lm(log(wage) ~ sex+union+occupation, data = CpsWages[-c(171,200),]))
```

```
##
## Call:
## lm(formula = log(wage) ~ sex + union + occupation, data = CpsWages[-c(171,
##     200), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24176 -0.31202 -0.01984  0.31941  1.15792
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.45072    0.06295  38.934  < 2e-16 ***
## sex1        -0.22360    0.04261  -5.248 2.24e-07 ***
## union1       0.27223    0.05160   5.276 1.94e-07 ***
## occupation2 -0.46806    0.09399  -4.980 8.66e-07 ***
```

```
## occupation3 -0.36163    0.07749  -4.667 3.89e-06 ***
## occupation4 -0.62041    0.07870  -7.883 1.86e-14 ***
## occupation5 -0.02188    0.07516  -0.291    0.771
## occupation6 -0.46727    0.07152  -6.534 1.52e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4419 on 524 degrees of freedom
## Multiple R-squared:  0.2752, Adjusted R-squared:  0.2655
## F-statistic: 28.42 on 7 and 524 DF,  p-value: < 2.2e-16
```