Jankun_Dong_HW4_CS555

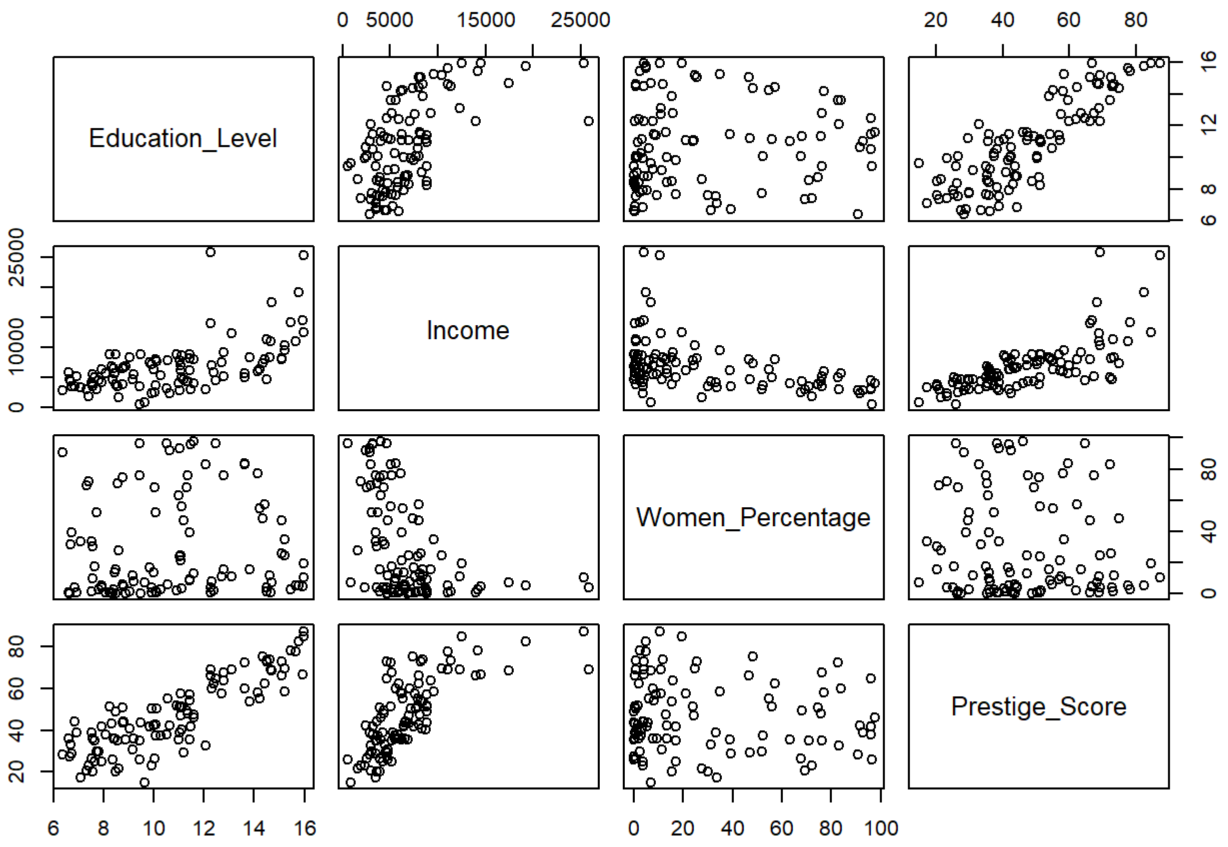
Jiankun (Bob) Dong CM3226

2023-11-05

```
hw4_data <- read.csv("./A04.csv")
attach(hw4_data)
#summary(hw4_data)
```

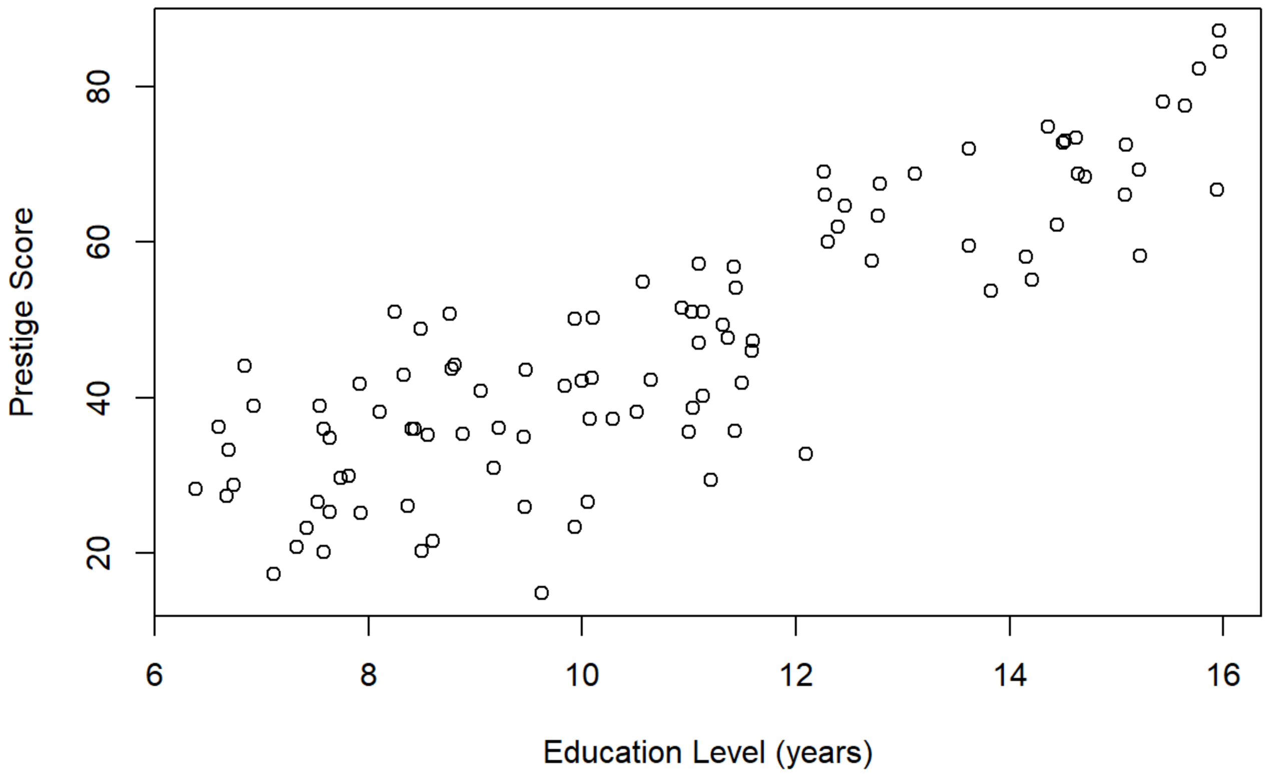
Overall summary of the data:

```
hw4_df <- data.frame(Education_Level = hw4_data$EL,
             Income = hw4_data$Inc,
             Women_Percentage = hw4_data$Perc,
             Prestige_Score = hw4_data$Score)
pairs(hw4_df)
```



Problem 1:

```
plot(EL ,Score, xlab = "Education Level (years)", ylab = "Prestige Score")
```



```
EL_Score_cor <- cor(EL,Score)
```

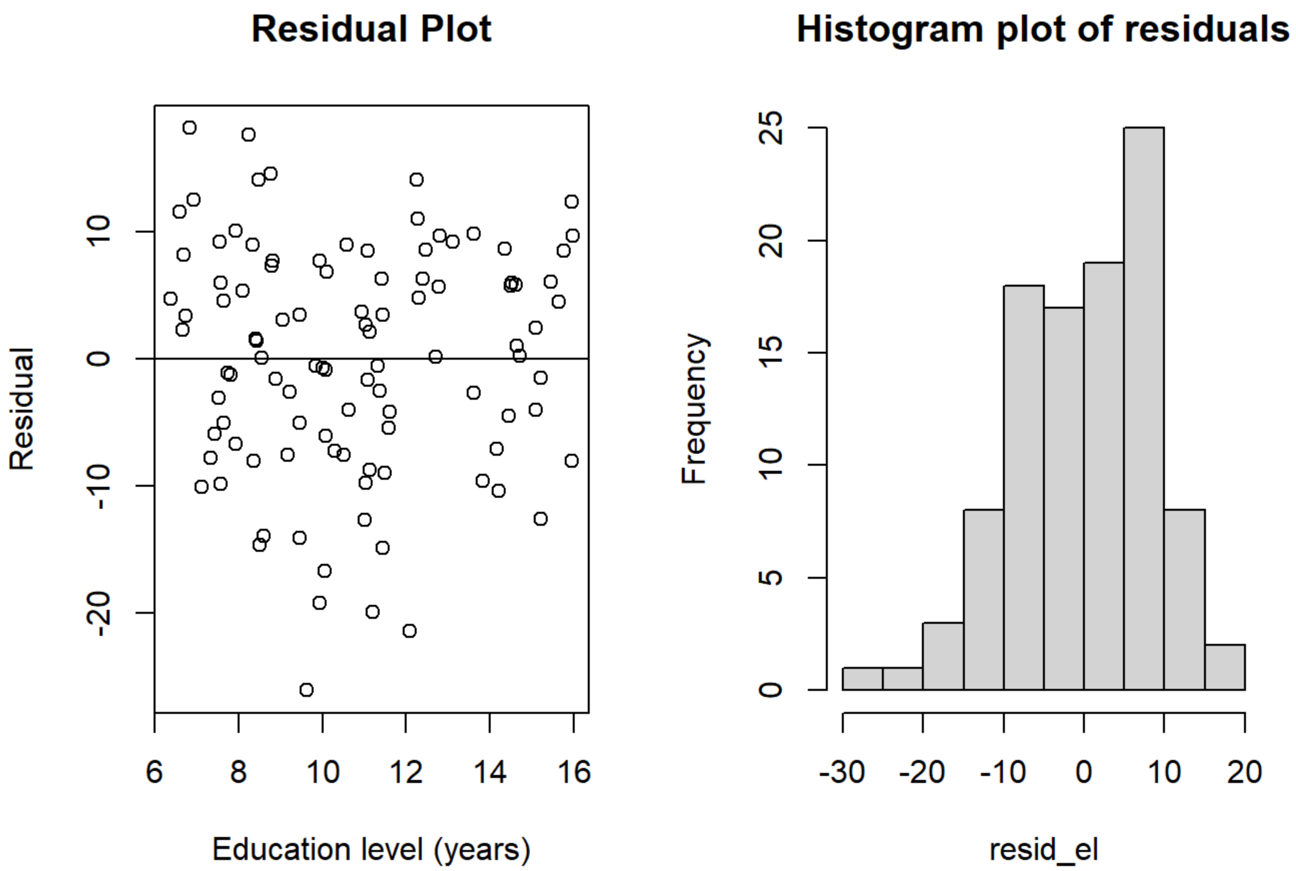
As the correlation coefficient between education level(years) and prestige score is 0.8501769, there’s a strong positive linear relationship between education level and prestige score.
Problem 2:

```
hw4_m_el <- lm(Score ~ EL,data = hw4_df)
resid_el <-resid(hw4_m_el)
summary(hw4_m_el)
```

```
## 
## Call:
## lm(formula = Score ~ EL, data = hw4_df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -26.0397  -6.5228   0.6611   6.7430  18.1636 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -10.732      3.677  -2.919  0.00434 ** 
## EL             5.361      0.332  16.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 9.103 on 100 degrees of freedom
## Multiple R-squared:  0.7228, Adjusted R-squared:   0.72 
## F-statistic: 260.8 on 1 and 100 DF,  p-value: < 2.2e-16
```

There’s a strong positive linear relationship between education level and prestige score.
Here’s the residual plot against education level:

```
par(mfrow = c(1,2))
plot(EL ,resid_el ,xlab = "Education level (years)",ylab = "Residual", main = "Residual Plot")
abline(0, 0)
hist(resid_el,main = "Histogram plot of residuals")
```



From the plot above, we can see that the residuals are roughly equal variance and linearly distributed. However, from the histogram we can see that the residual is slightly left screwed, due to a few points with low residual values. Those points will show up on the following tests for influence points and outliers. If we ignore those points, we can say that the residual also follows the normality assumption. Now for outliers and influence points:

```
outlierTest(hw4_m_el)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 53 -2.98896          0.0035306      0.36012
```

No outliers.

```
b <- influence.measures(hw4_m_el)
c <- which(apply(b$is.inf, 1, any))
hw4_data[rownames(hw4_data) %in% c, ]
```

```
##          Title    EL  Inc  Perc Score
## 41 FILE_CLERKS 12.09 3016 83.19  32.7
## 46  COLLECTORS 11.20 4741 47.06  29.4
## 53    NEWSBOYS  9.62  918  7.00  14.8
```

There are 3 influence points as show above (ID; 41, 46 and 53). From the previous scatter plot we can see that all three influence points have similar effect on the slop: They lower the value of the slop.

Problem 3:

```
hw4_lm<-lm(Score ~ EL + Inc + Perc,data = hw4_data)
residual_hw4 <- resid(hw4_lm)
summary(hw4_lm)
```

```
##
## Call:
## lm(formula = Score ~ EL + Inc + Perc, data = hw4_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8246  -5.3332  -0.1364   5.1587  17.5045
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.7943342  3.2390886  -2.098   0.0385 *
## EL           4.1866373  0.3887013  10.771  < 2e-16 ***
## Inc          0.0013136  0.0002778   4.729 7.58e-06 ***
## Perc        -0.0089052  0.0304071  -0.293   0.7702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.846 on 98 degrees of freedom
## Multiple R-squared:  0.7982, Adjusted R-squared:  0.792
## F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

Step 1:

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

$H_1 : \beta_i \neq 0$ for at least one i

$\alpha = 0.05$

$k = 3, n = 102$

Step 2:

We use $F = \frac{RegMS}{ResMS}$ Step 3:

The decision rule is reject H_0 if $F \geq F(3, 98, 0.05)$, which is F = 0.1168981 Step 4:

```
summary(hw4_lm)
```

```
##
## Call:
## lm(formula = Score ~ EL + Inc + Perc, data = hw4_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8246  -5.3332  -0.1364   5.1587  17.5045
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.7943342  3.2390886  -2.098   0.0385 *
## EL           4.1866373  0.3887013  10.771  < 2e-16 ***
## Inc          0.0013136  0.0002778   4.729 7.58e-06 ***
## Perc        -0.0089052  0.0304071  -0.293   0.7702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.846 on 98 degrees of freedom
## Multiple R-squared:  0.7982, Adjusted R-squared:  0.792
## F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

Step 5:
F value (129.2) is way bigger than 0.1168981, reject H_0.
There is evidence for a linear association between the prestige score and (education level, income and percent of women) at $\alpha = 0.05$ level. (The overall model is significant)

Problem 4:

```
tcrit <- qt(1-0.05/2,102-1-3)
summary(hw4_lm)
```

```
## 
## Call:
## lm(formula = Score ~ EL + Inc + Perc, data = hw4_data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -19.8246  -5.3332  -0.1364   5.1587  17.5045 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -6.7943342  3.2390886  -2.098   0.0385 *  
## EL           4.1866373  0.3887013  10.771  < 2e-16 ***
## Inc          0.0013136  0.0002778   4.729 7.58e-06 ***
## Perc        -0.0089052  0.0304071  -0.293   0.7702    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.846 on 98 degrees of freedom
## Multiple R-squared:  0.7982, Adjusted R-squared:  0.792 
## F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

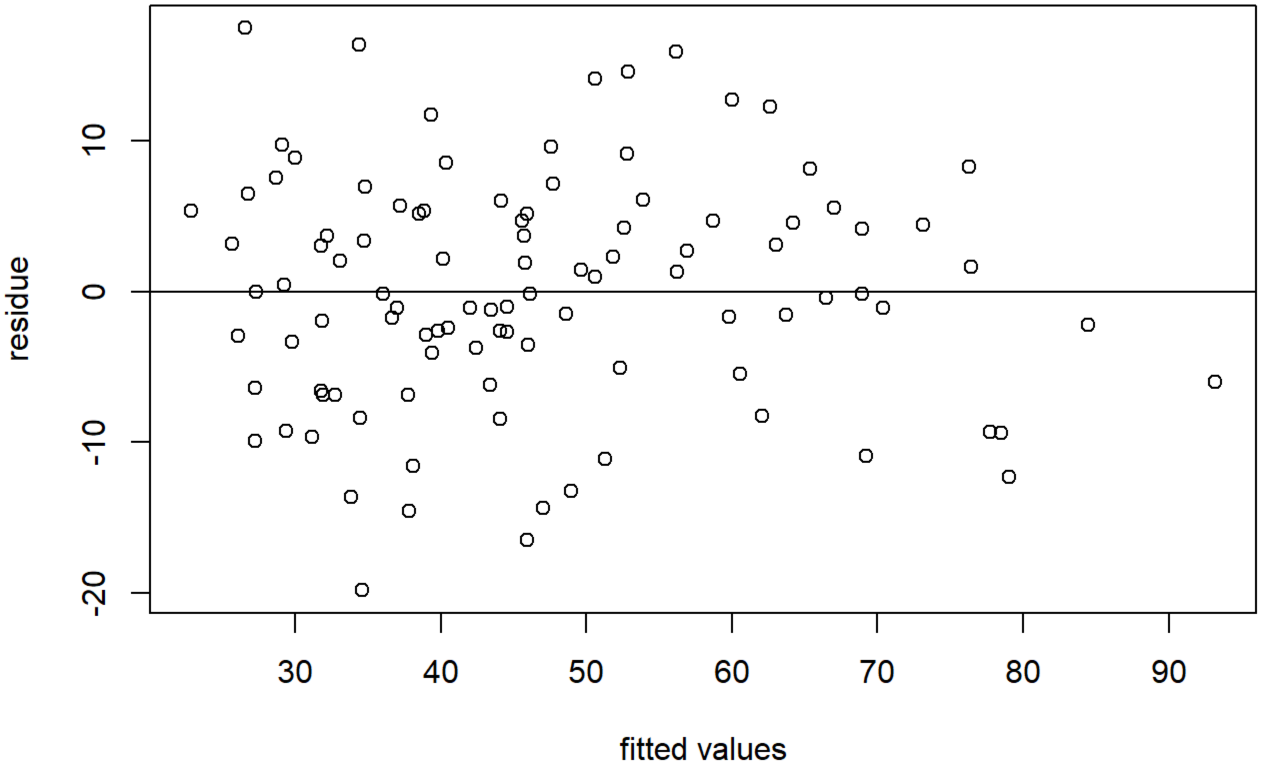
```
confhw <- confint(hw4_lm,level =0.95)
```

The critical value of t for $\alpha = 0.05$ is 1.9844675. From the summary we can see that:
Education level is a significant predictor of prestige level after adjusting for income and percent of women.
After controlling income and percent of women, for 1 year increase in the education level, the prestige level increase by 4.1866373. The 95% confidence interval is (3.4152723,4.9580023).

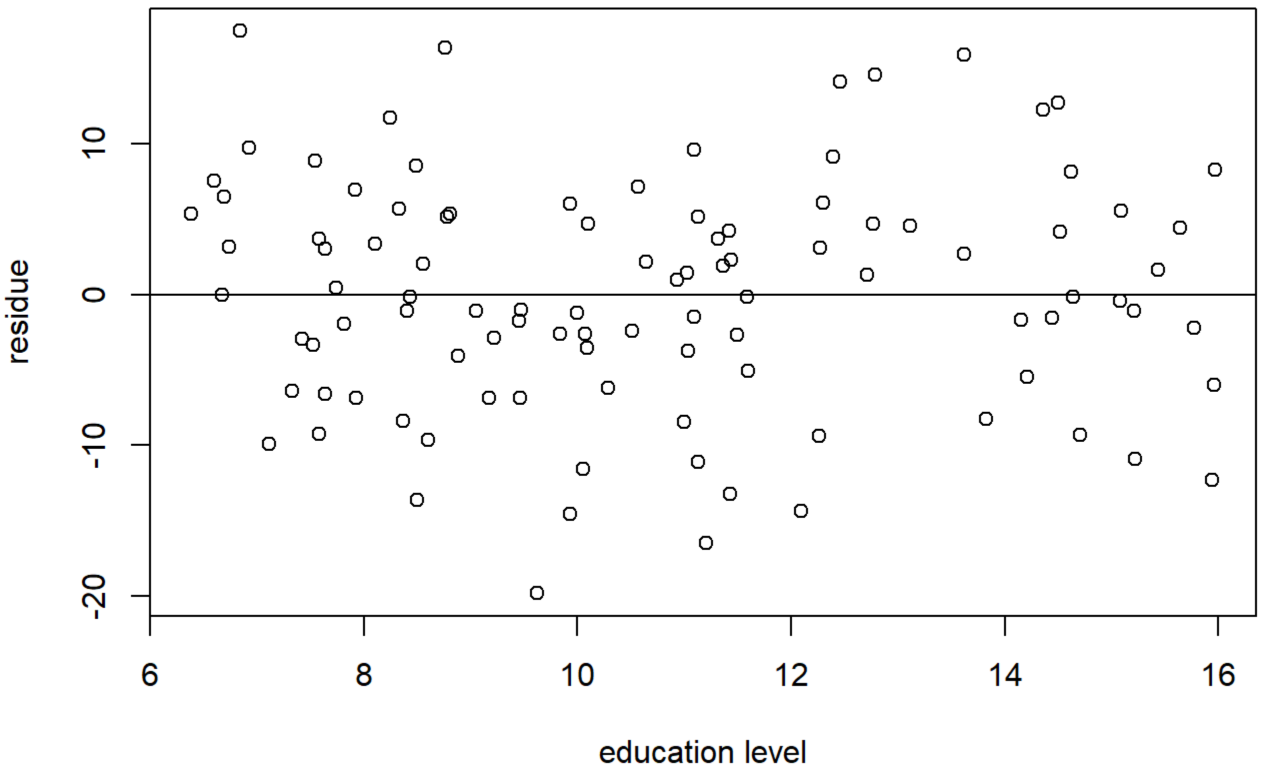
Income is a significant predictor of prestige level after adjusting for education level and percent of women.
After controlling education level and percent of women, for 1 dollar increase in income, the prestige level increase by 0.0013136. The 95% confidence interval is (7.6231268^{-4},0.0018648).

At $\alpha = 0.05$, percentage of women is not a significant predictor for prestige level, after adjusting for education level and income.
Problem 5:

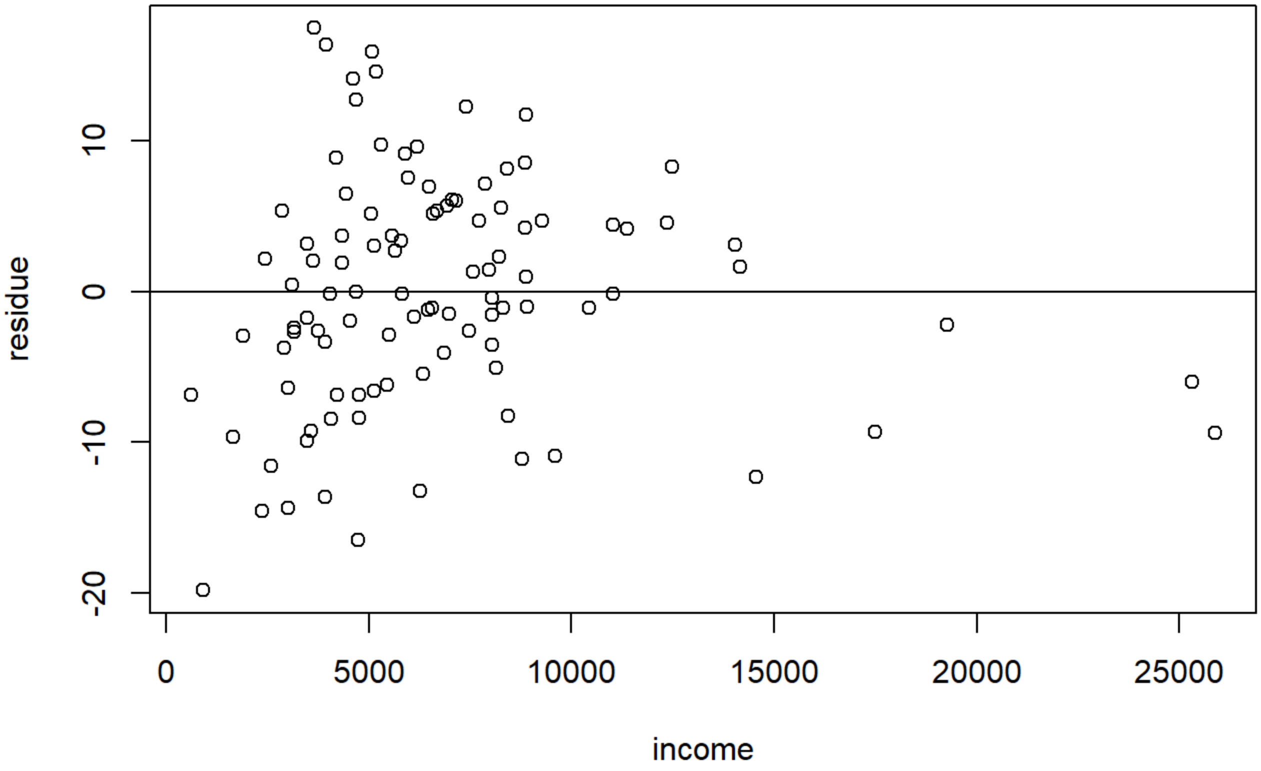
```
plot(fitted(hw4_lm),residual_hw4,axes=TRUE, frame.plot=TRUE, xlab='fitted values', ylab='residue')
abline(0,0)
```



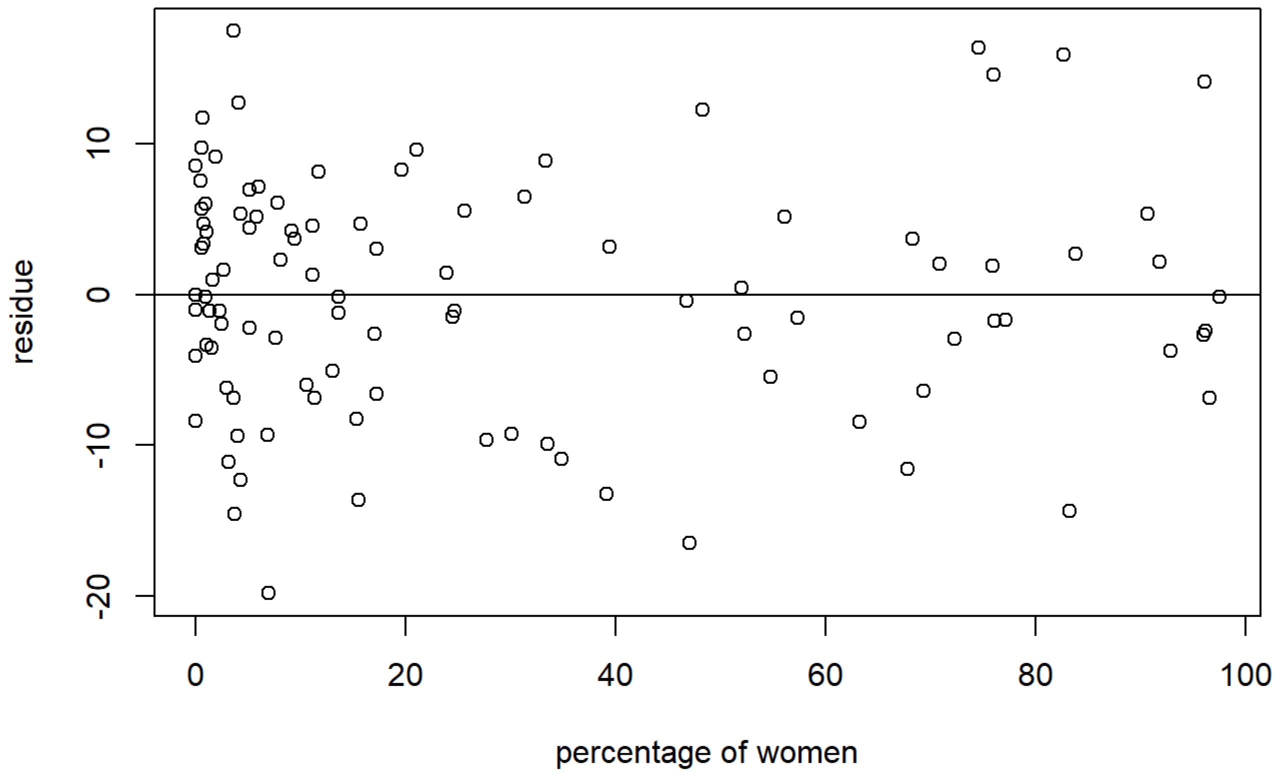
```
plot(EL, residual_hw4, axes=TRUE, frame.plot=TRUE, xlab='education level', ylab='residue')
abline(0,0)
```



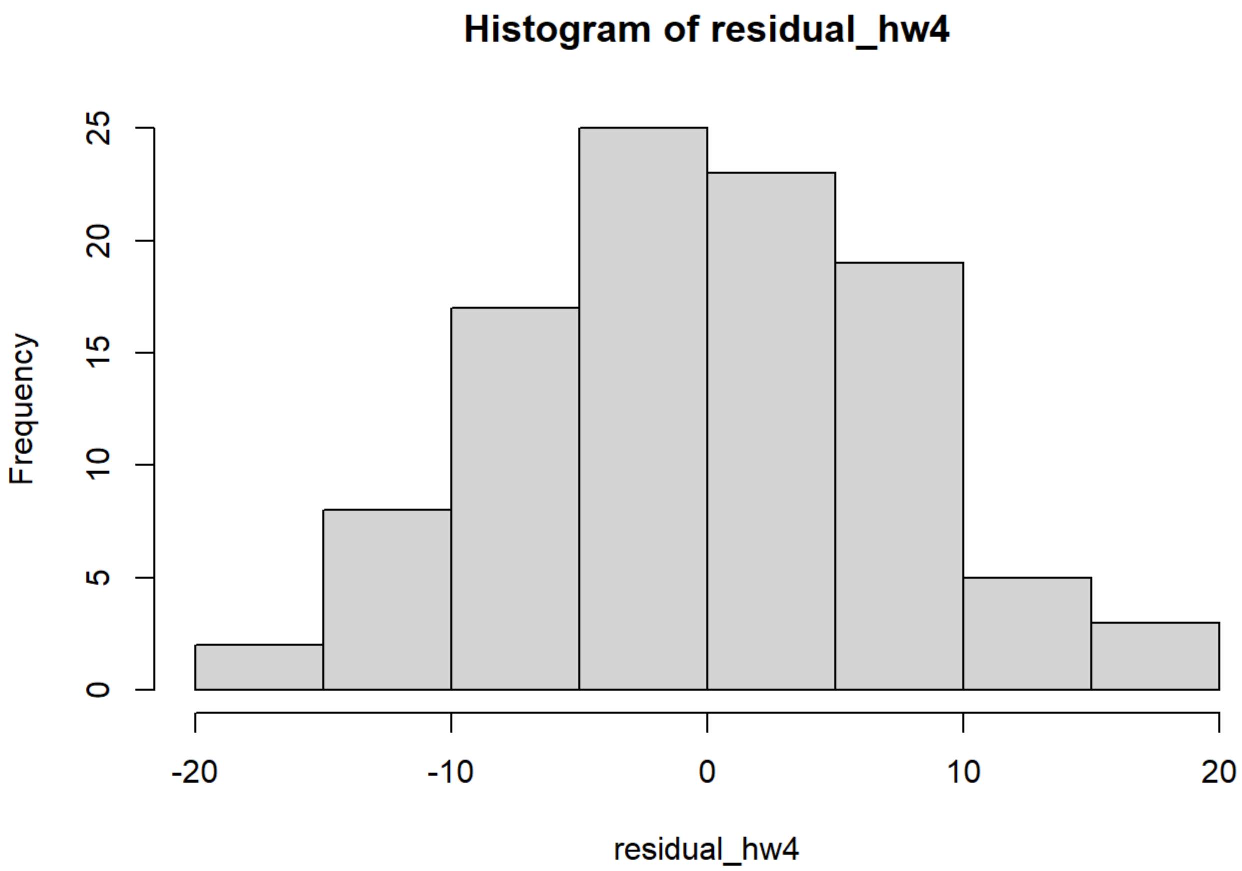
```
plot(Inc, residual_hw4, axes=TRUE, frame.plot=TRUE, xlab='income', ylab='residue')
abline(0,0)
```



```
plot(Perc, residual_hw4, axes=TRUE, frame.plot=TRUE, xlab='percentage of women', ylab='residue')
abline(0,0)
```



```
hist(residual_hw4)
```



Based on the graph, I think the model is reasonable as for each variable, the residual follows: linearity and equal variance. And the residual is roughly normal distributed. However, we might want to gather more data on people with higher values of incomes, as there’s currently very few data points for us to say for certain that it follows equal variance.

```
outlierTest(hw4_lm)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 53 -2.694442          0.0083101      0.84763
```

No outliers.

```
b <- influence.measures(hw4_lm)
c <- which(apply(b$is.inf, 1, any))
hw4_data[rownames(hw4_data) %in% c, ]
```

```
##                    Title    EL   Inc  Perc Score
## 2       GENERAL_MANAGERS 12.26 25879  4.02  69.1
## 17               LAWYERS 15.77 19263  5.13  82.3
## 24            PHYSICIANS 15.96 25308 10.56  87.2
## 46            COLLECTORS 11.20  4741 47.06  29.4
## 53              NEWSBOYS  9.62   918  7.00  14.8
## 67               FARMERS  6.84  3643  3.60  44.1
## 84 SEWING_MACH_OPERATORS  6.38  2847 90.67  28.2
```

There are 7 influence points as shown.