STAT-242 Final Project

Hanfei Su

2022/4/16

Introduction

I chose the data set "Wife Working Hours" from the Github link(containing data sets) provided. Working wife has been such a popular topic for a long time; it is closely related to woman's independence. This data set includes comprehensive information that may affect a woman's decision to work. I think it would be an appropriate data set for me to analyze.

This data set has 3382 cases and 12 variables. The variables are wife working hours per year, household income in hundreds of dollars, age of the wife, education years of the wife, number of children for ages 0 to 5, number of children for ages 6 to 13, number of children for ages 14 to 17, non-white, whether the home owned by the household or not, whether the house on a mortgage or not, occupation of the husband, and local unemployment rate.

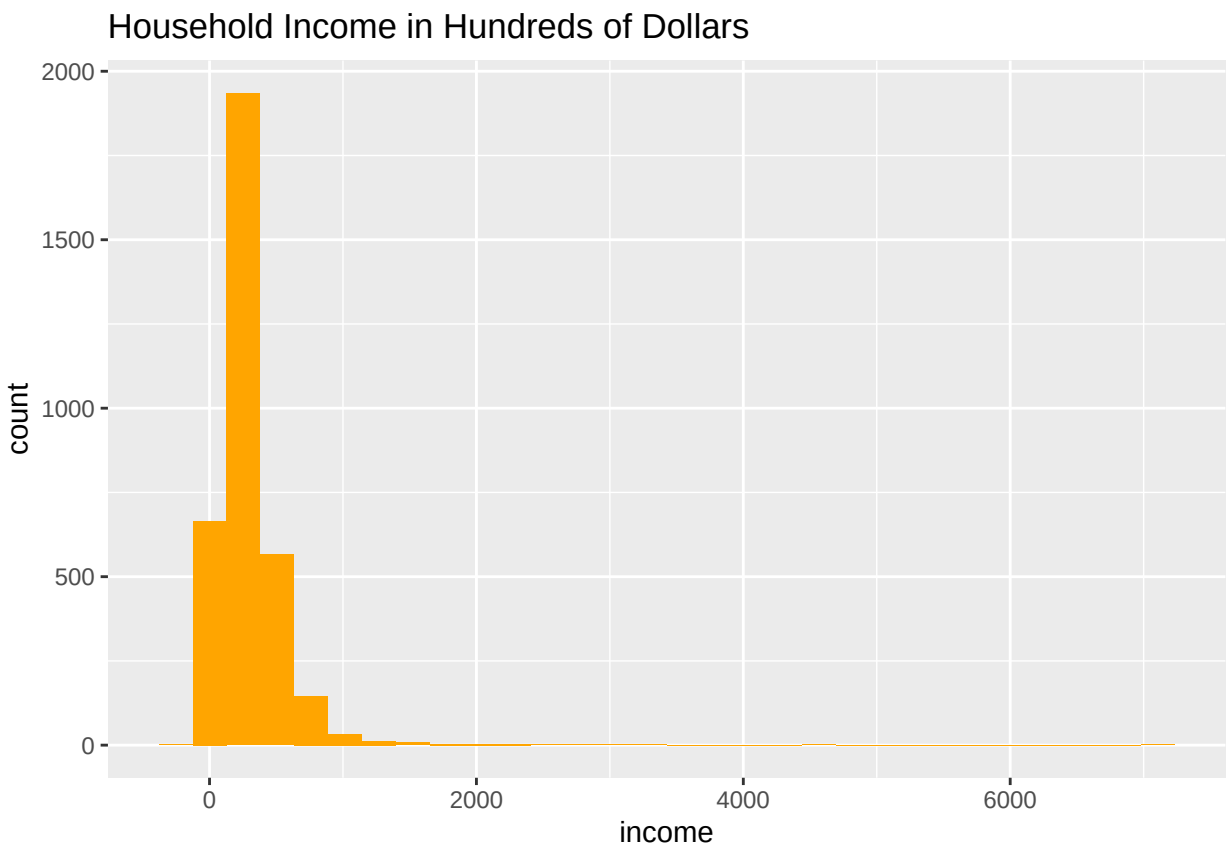
Among these 12 variables, all of them except the husband's occupation are represented as numerical variables, while the husband's occupation is represented as a categorical variable. But I would consider nonwhite, owned, and mortgage as categorical variables, because they are represented by 0(no) and 1(yes).

```
wifeWorkingHours <-
  read.csv("https://vincentarelbundock.github.io/Rdatasets/csv/Ecdat/Workinghours.csv",
           stringsAsFactors = TRUE)
glimpse(wifeWorkingHours)
```

```
## Rows: 3,382
## Columns: 13
## $ rownames   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
## $ hours      <int> 2000, 390, 1900, 0, 3177, 0, 0, 1040, 2040, 0, 1432, 1544, ~
## $ income     <int> 350, 241, 160, 80, 456, 390, 181, 726, -5, 78, 195, 95, 351~
## $ age        <int> 26, 29, 33, 20, 33, 22, 41, 31, 33, 30, 41, 23, 32, 56, 35,~
## $ education  <int> 12, 8, 10, 9, 12, 12, 9, 16, 12, 11, 12, 14, 16, 12, 12, 14~
## $ child5     <int> 0, 0, 0, 2, 0, 2, 0, 2, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0,~
## $ child13    <int> 1, 1, 2, 0, 2, 0, 0, 1, 3, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0,~
## $ child17    <int> 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ nonwhite   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0,~
## $ owned      <int> 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0,~
## $ mortgage   <int> 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0,~
## $ occupation <fct> swcc, other, swcc, other, swcc, other, swcc, mp, fr, other,~
## $ unemp      <int> 7, 4, 7, 7, 7, 7, 7, 3, 4, 5, 7, 3, 7, 12, 12, 7, 7, 6, 7, ~
```

Data Descriptives

```
ggplot(data = wifeWorkingHours, aes(x = income)) +
  geom_histogram(fill = "orange") +
  ggtitle("Household Income in Hundreds of Dollars")
```




```
summary(wifeWorkingHours$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -139.0   146.0   247.0   296.9   368.8  7220.0
```

The center of the data is around 247.0(median). The data ranges from -139.0 to 7220, this is a wide range and I am surprised the min is negative. The distribution is unimodal and right-skewed. There is a unusual value in the data, which is the maximum value 7220, although we almost cannot see it in this histogram.

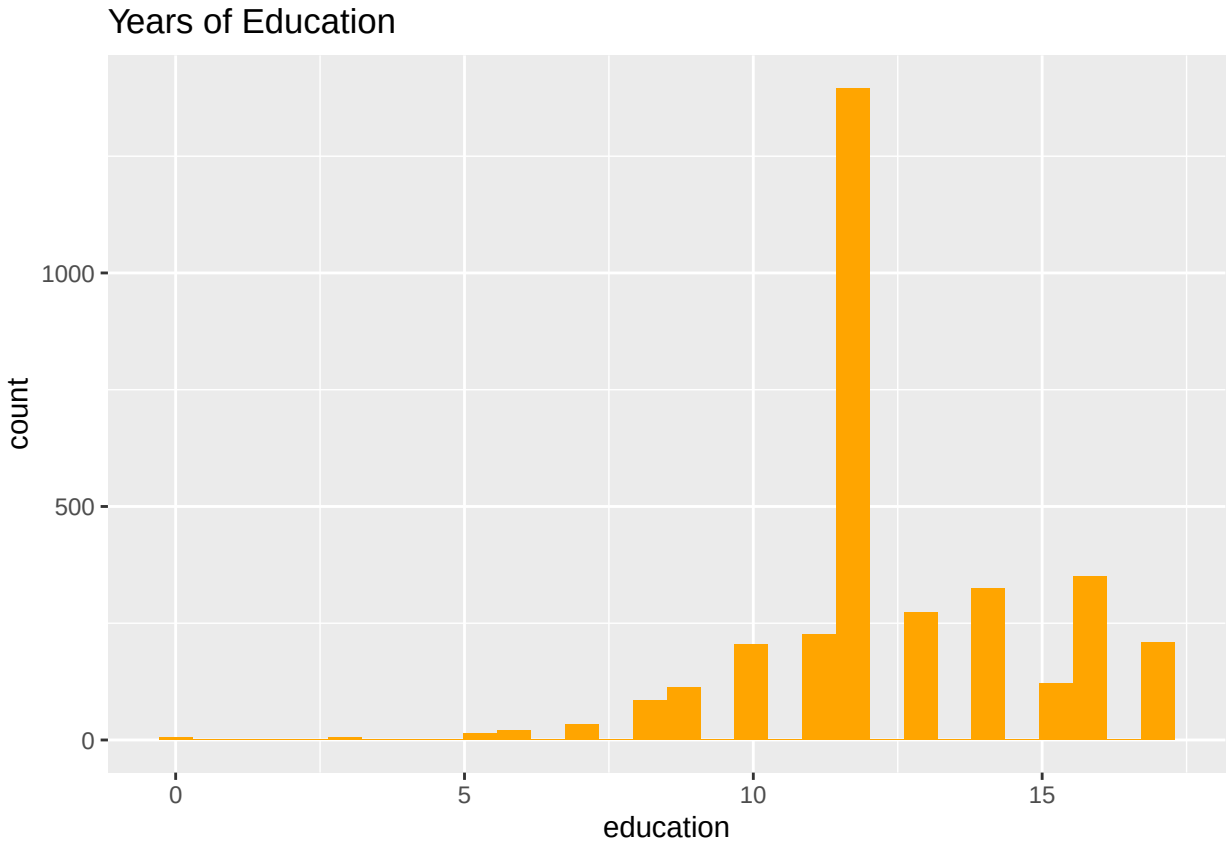
```
ggplot(data = wifeWorkingHours, aes(x = hours)) +
  geom_histogram(fill = "orange") +
  ggtitle("Hours of Wife Working per Year")
```


```
summary(wifeWorkingHours$hours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0    1304    1135    1944    5840
```

The center of the data is around 1304(median). The data ranges from 0 to 5840. If a wife's hour of working is 0, this means that she is full-time housewife. There are usually 52 weeks of a year, which means the wife who works 5840 hours per year has to work 112.3 hours per week and 16 hours per day(no resting weekend)! The distribution is bimodal and left-skewed, but there are nearly 1000 full-time housewives. There is a unusual value in the data, which is the maximum value 5840.

```
ggplot(data = wifeWorkingHours, aes(x = education)) +
  geom_histogram(fill = "orange") +
  ggtitle("Years of Education")
```

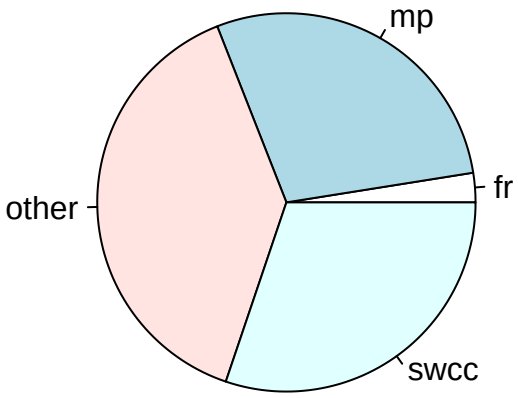



```
summary(wifeWorkingHours$education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.00   12.00   12.55   14.00   17.00
```

The center of the data is around 12(median), I think this makes sense because this is the sum of years of primary school(6) and years of high school(6), and we know the wives who got 12 years of education didn't go to the colleges. The data ranges from 0 to 17. The distribution is unimodal. The values under 2.5 can be considered as unusual values in the data.

```
pie(table(wifeWorkingHours$occupation))
```

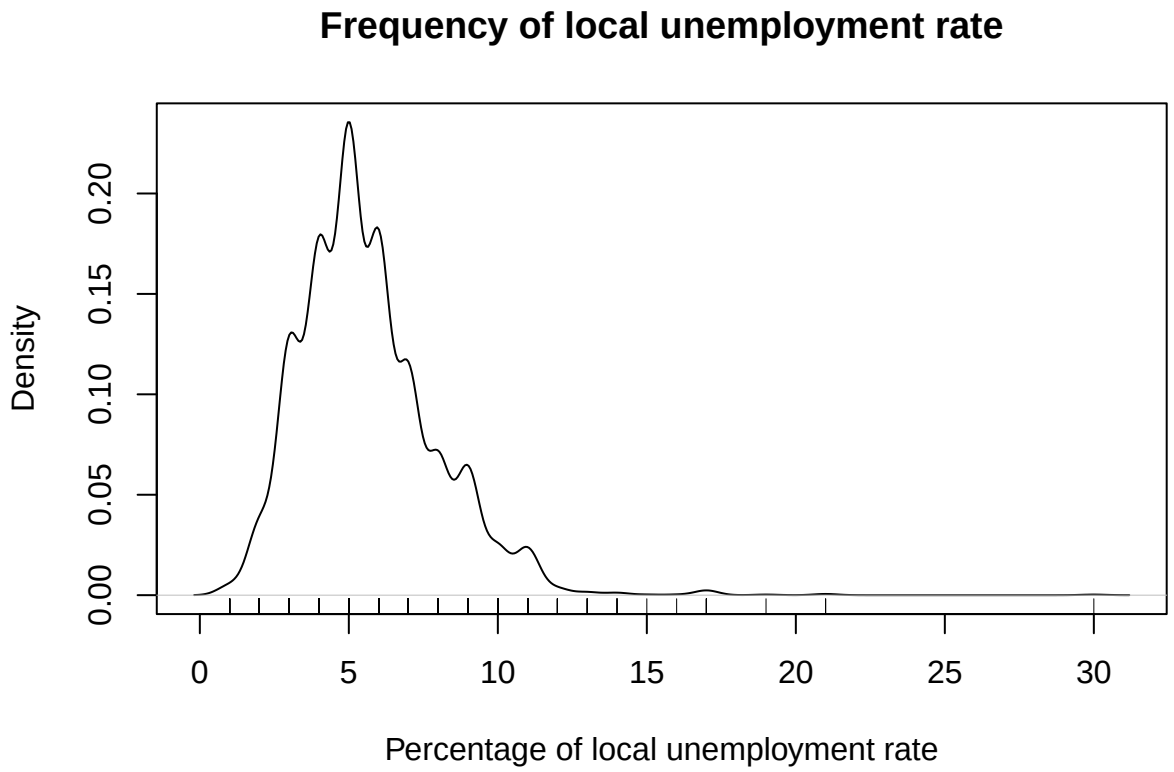



```
summary(wifeWorkingHours$occupation)
```

```
##    fr    mp other  swcc
##    85   962  1314  1021
```

In the sample, 962 husbands are manager or professional; 1021 husbands are sales or workers or clericals or craftmen; 85 husbands are farm workers; and rest of them are in other occupation.

```
plot(density(wifeWorkingHours$unemp),
     main="Frequency of local unemployment rate",
     xlab="Percentage of local unemployment rate")
rug(wifeWorkingHours$unemp)
```




```
summary(wifeWorkingHours$unemp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   5.000   5.641   7.000  30.000
```

The most frequent percentage of local unemployment rate is between 3% to 8%. The highest percentage of local unemployment rate is 30%.

Simple Linear Regression

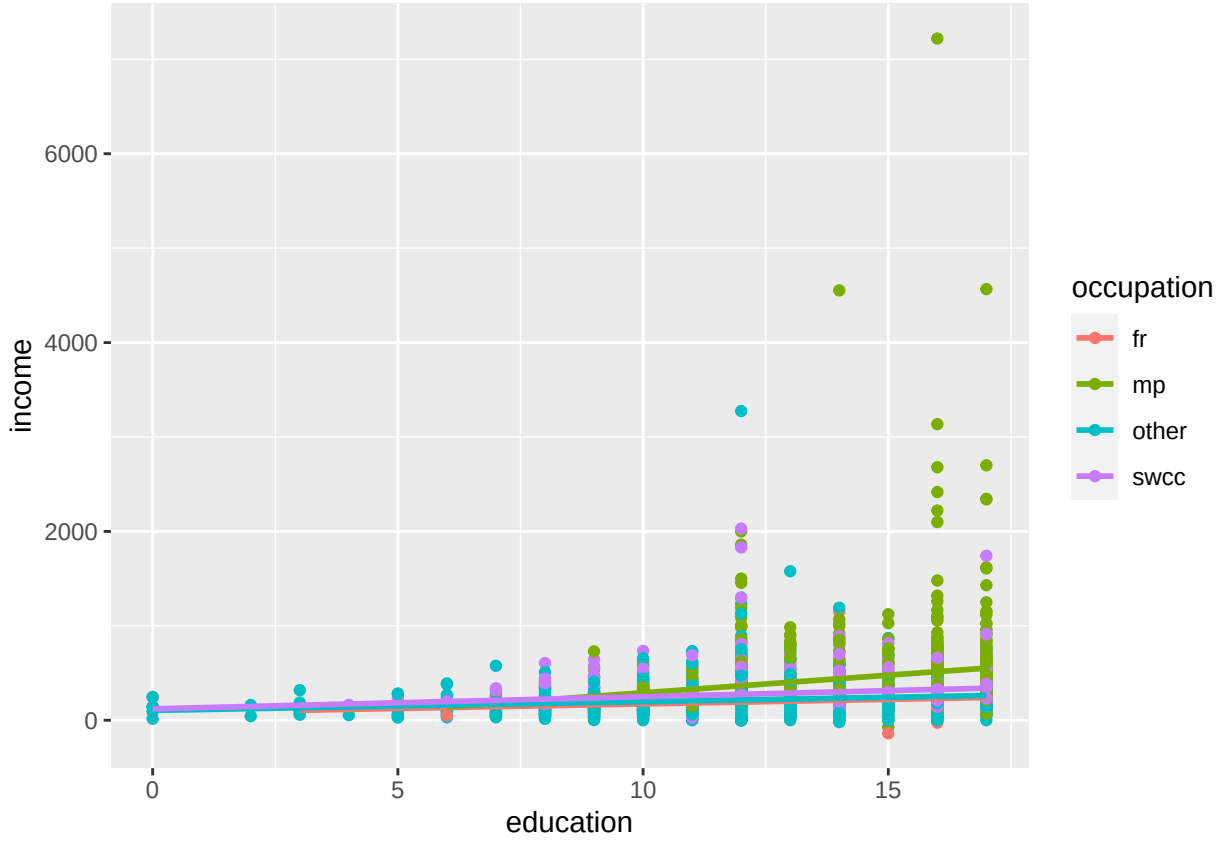
```
lm1 = lm(hours ~ age, data = wifeWorkingHours)
summary(lm1)
```

```
##
## Call:
## lm(formula = hours ~ age, data = wifeWorkingHours)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1354.3  -923.9   173.3   770.4  4683.4
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1563.580     51.522  30.348   <2e-16 ***
## age          -11.629      1.337  -8.695   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 882.9 on 3380 degrees of freedom
## Multiple R-squared:  0.02188,    Adjusted R-squared:  0.02159
## F-statistic:  75.6 on 1 and 3380 DF,  p-value: < 2.2e-16
```

The model above shows a negative relationship between age and working hours. Since the p value for t test is almost zero, we are confident there is relationship between age and working hours under 1% significant level.

```
wifeWorkingHours %>%
  group_by(occupation) %>%
  ggplot(aes(x = education, y = income,
             color = occupation)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



Statistical Inference

```
wifeWorkingHours_lm = lm(income ~ age + hours + education, data = wifeWorkingHours)
```

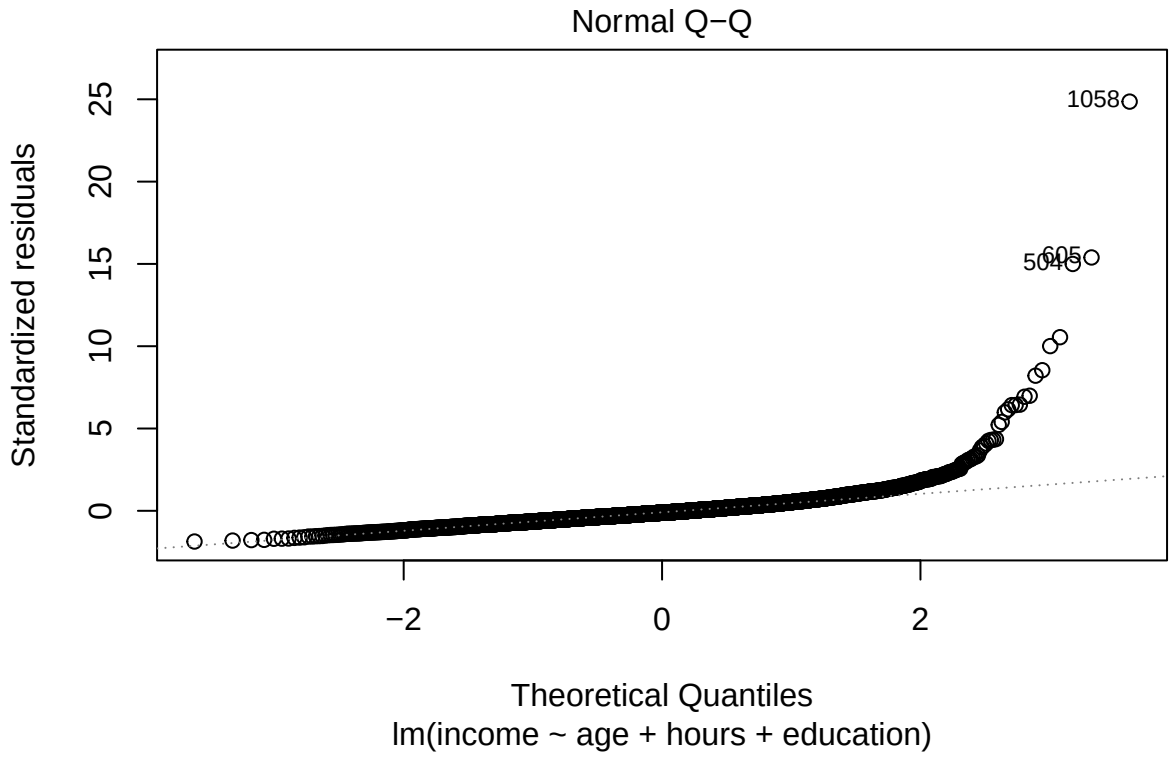
Hypotheses

H_0: Nothing is useful in this model.

H_A: At least one predictor is useful in this model.

Functions and Transformations

```
wifeWorkingHours_lm %>% plot(which = 2)
```



Normality: We can use the Normal QQ plot of the residuals to check normality. The points are more or less along the line, but when theoretical quantile is larger than 2, the points show an increasing trend.

Since the normality wasn't satisfied, I need to apply transformation for income.

```
trans_income <- function(x) {
  result_income = x + 139.000001 # since the smallest value for income is -139
  return(result_income)
}
# transform the income
translated_income <- c()
for(i in 1:3382){
  translated_income[i] = trans_income(wifeWorkingHours$income[i])
}
# add updated variable into the dataset
wifeWorkingHours <- wifeWorkingHours %>%
    mutate(translated_income = translated_income)
glimpse(wifeWorkingHours)
```

```
## Rows: 3,382
## Columns: 14
## $ rownames          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1~
## $ hours             <int> 2000, 390, 1900, 0, 3177, 0, 0, 1040, 2040, 0, 1432,~
## $ income            <int> 350, 241, 160, 80, 456, 390, 181, 726, -5, 78, 195, ~
## $ age               <int> 26, 29, 33, 20, 33, 22, 41, 31, 33, 30, 41, 23, 32, ~
```

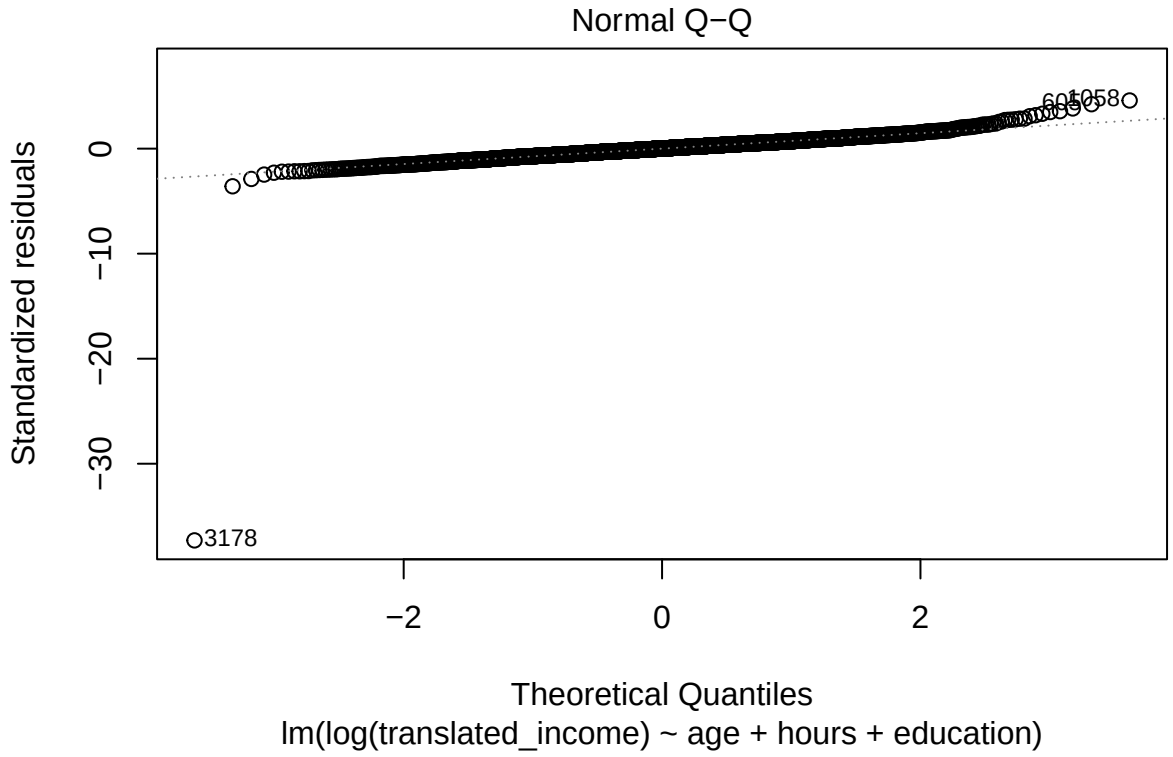
```
## $ education         <int> 12, 8, 10, 9, 12, 12, 9, 16, 12, 11, 12, 14, 16, 12,~
## $ child5            <int> 0, 0, 0, 2, 0, 2, 0, 2, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1~
## $ child13           <int> 1, 1, 2, 0, 2, 0, 0, 1, 3, 1, 1, 0, 0, 0, 1, 0, 0, 1~
## $ child17           <int> 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ nonwhite          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0~
## $ owned             <int> 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0~
## $ mortgage          <int> 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0~
## $ occupation        <fct> swcc, other, swcc, other, swcc, other, swcc, mp, fr,~
## $ unemp             <int> 7, 4, 7, 7, 7, 7, 7, 3, 4, 5, 7, 3, 7, 12, 12, 7, 7,~
## $ translated_income <dbl> 489, 380, 299, 219, 595, 529, 320, 865, 134, 217, 33~
```

After that, we can rebuild the linear model using the new variable.

```
wifeWorkingHours_lm_updated = lm(log(translated_income) ~ age + hours + education,
                                 data = wifeWorkingHours)
```

Analysis

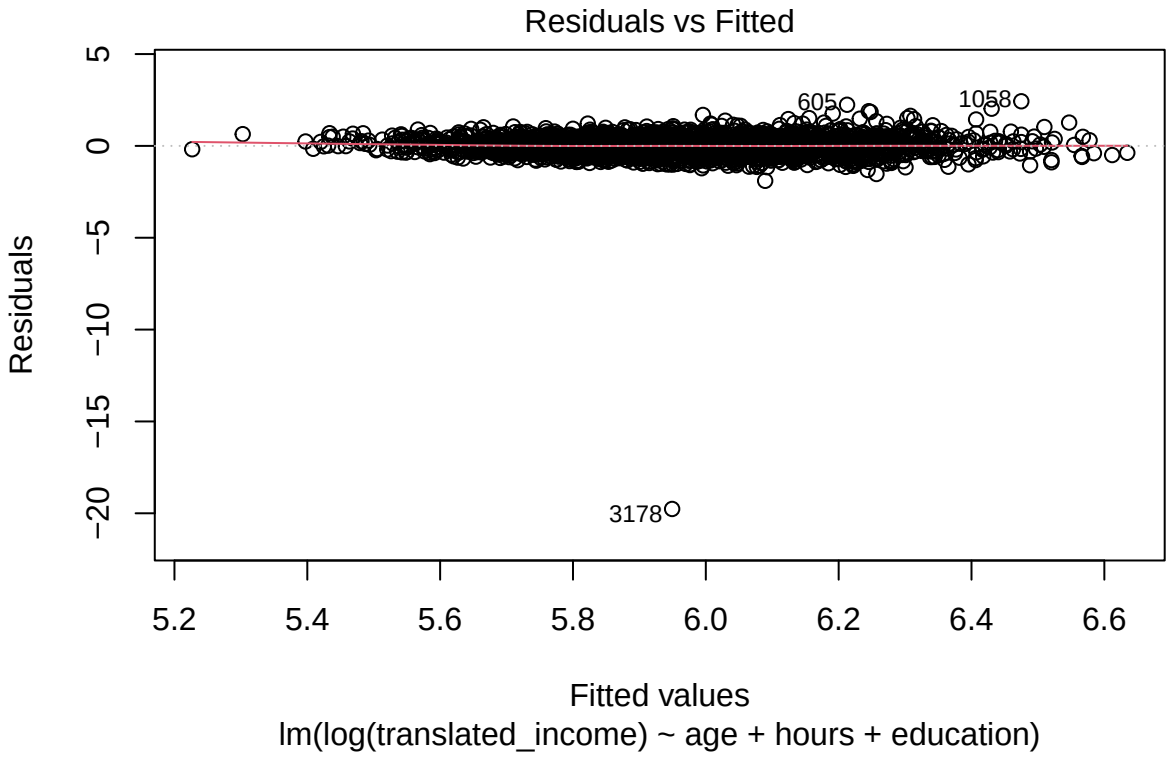
```
wifeWorkingHours_lm_updated %>% plot(which = 2)
```



Normality: This looks much better!

Linearity: We can look at the plot of the residuals vs. the fitted values and check for any bends there.

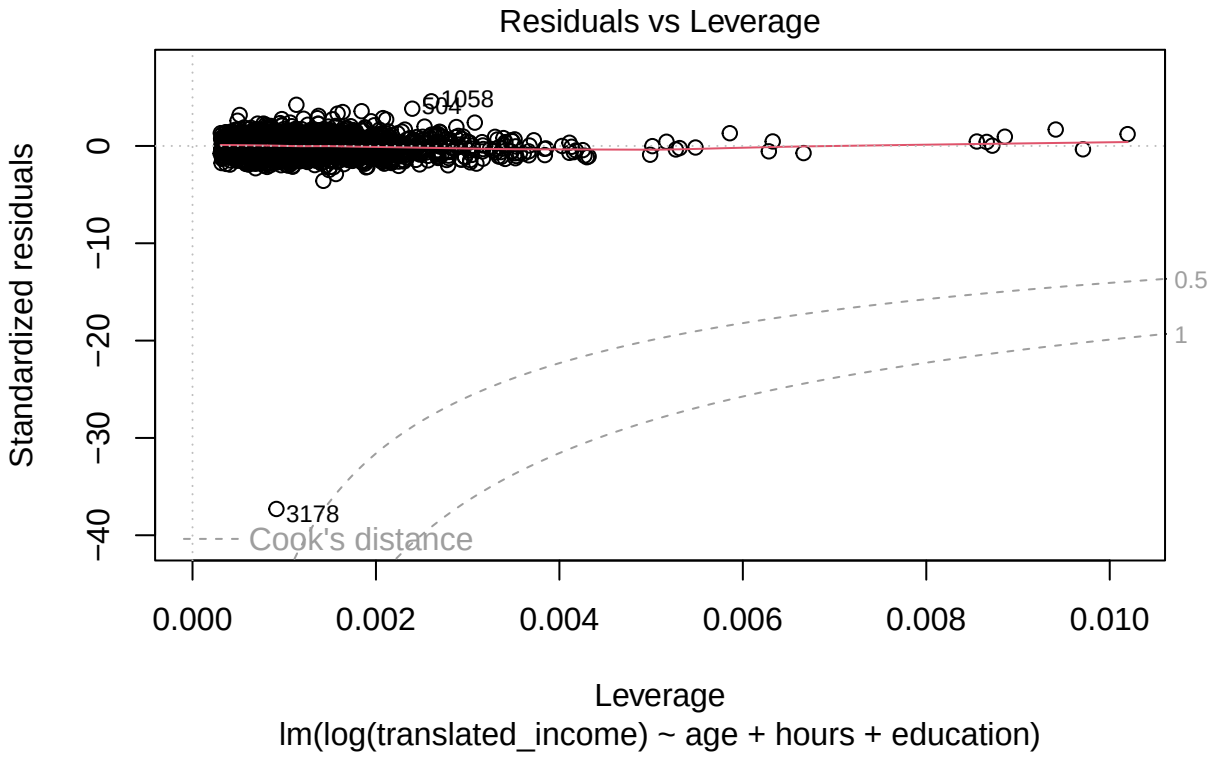
```
wifeWorkingHours_lm_updated %>% plot(which = 1)
```



This looks fine. So it meets linearity condition.

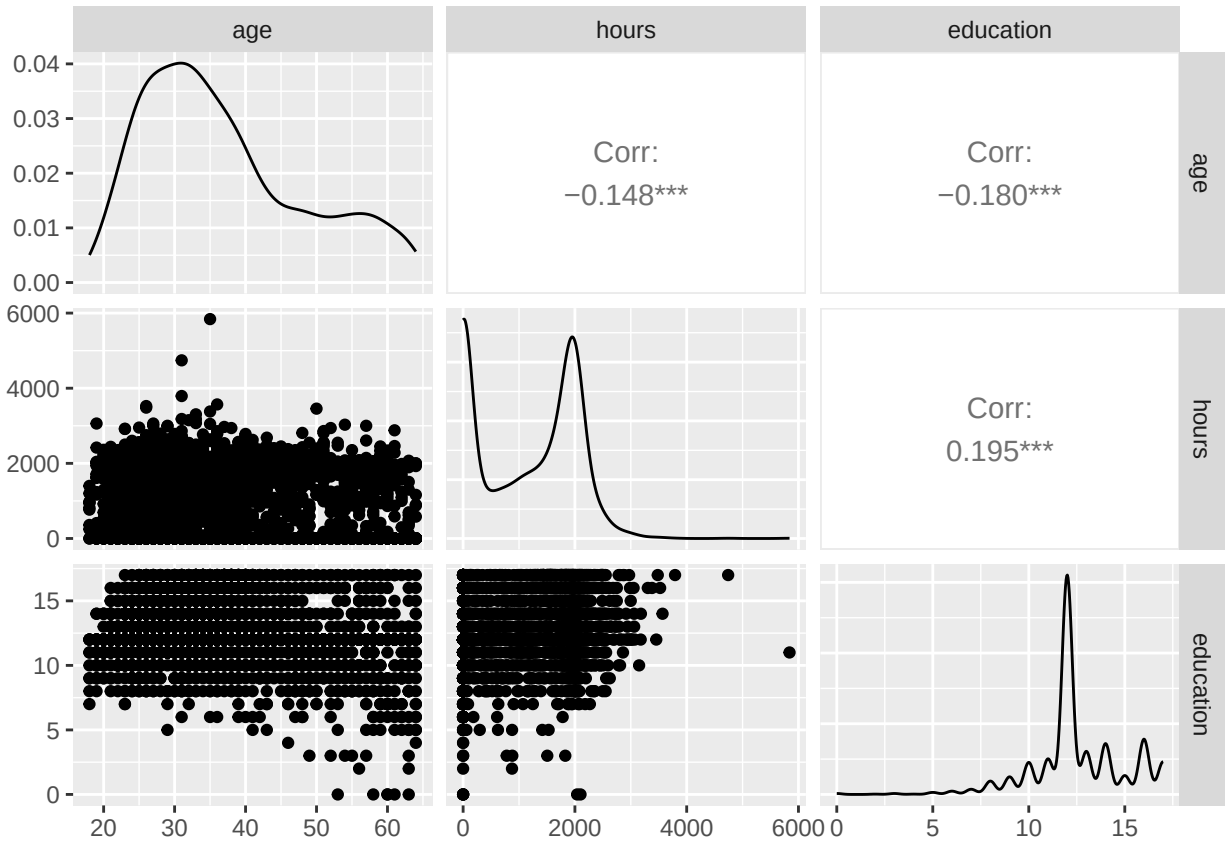
Equal Variance: We can check equal variance by looking at the plot of the residuals vs. the fitted values. The vertical spread of the residuals seems about the same all the way across, except the residual of 1058. This condition is met if we ignore this residual.

```
wifeWorkingHours_lm_updated %>% plot(which = 5)
```



Outliers: We can see that point 3178 does have a very high residual, but it doesn't have much leverage, so it probably isn't that influential on the slope.

```
wifeWorkingHours %>%
  dplyr::select(age, hours, education) %>%
  ggpairs()
```



Multicollinearity: Looks like these two predictors are not linearly related, or indeed associated in any way.

Independence: There's no obvious multicollinearity within this model, so we are able to assume that it is independent.

Test Statistics

```
summary(wifeWorkingHours_lm_updated)
```

```
##
## Call:
## lm(formula = log(translated_income) ~ age + hours + education,
##     data = wifeWorkingHours)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.7648  -0.2617   0.0046   0.2625   2.4287
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.739e+00  6.227e-02  76.114  < 2e-16 ***
## age          1.135e-02  8.221e-04  13.802  < 2e-16 ***
## hours       -5.587e-05  1.049e-05  -5.328 1.06e-07 ***
## education    6.877e-02  3.896e-03  17.653  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5301 on 3378 degrees of freedom
```

```
## Multiple R-squared:  0.1182, Adjusted R-squared:  0.1174
## F-statistic: 150.9 on 3 and 3378 DF,  p-value: < 2.2e-16
```

Conclusion

Assuming we're using a typical α of 0.01, this p-value is less than α. We reject the null hypothesis that nothing in the model is useful. Our model is better than just guessing y^- for every prediction we make.