THE DIFFERENT WEEKLY EARNINGS BETWEEN BLACK AND NOT BLACK PEOPLE

Aims

To investigate whether race influences weekly earnings

Background

There is different race in the world. This project would like to investigate whether race affects weekly earnings. The dataset ex1029 in the Sleuth3 package of R studio are used in this project for the investigation. The dataset includes the following personal information: weekly earnings in $, region (Midwest, Northeast, South, or West), Metropolitan Status (Whether a person live in a metropolitan area), year of experience, year of education, race (whether a person is Black). There are 25437 sample in the dataset, 1978 are black people and 23459 are not.

Findings

The weekly earnings of not black people are 1.264 times more than black people.

Discussion

The estimate of not black people is 2.352e-01 (Table 3), this helps the project to calculate the difference of weekly earning between not black and black people, which not black people are 1.264 times more than black people. 63 years of experience is not usual, so this outlier is removed. Log transformation can help this project. Figure 5 shows in the metropolitan area, with less than seven years of education, black and not black people have the similar weekly earning, with more than seven years of education, not black people have more weekly earning than black people. In the not metropolitan area, with the same education, not black people have more weekly earnings than not black people.

Statistical methods and results

Computational methods

- R 4.0.2 and RStudio Desktop 1.3.1073 are used to do the analysis in this project
- plot() for creating a plot for a data.
- lm() for fitting linear models.
- library() for loading a packages.
- summary() for producing a summarized result of a data
- layout() for showing data in a matrix format..
- pairs() for making a matrix of scatterplots
- boxplot() for making a box-and-whisker plot
- aes() for Aesthetic mappings
- ggplot() for initializes a ggplot
- log() for calculating a value after log
- geom_bin2d() for plane rectangle divide
- facet_grid() for produce a matrix of panels
- data.frame() for creating data frames
- lines() for adding line in a plot
- exp() for log calculation
- confint() for calculating confidence intervals
- Package ggplot2, MASS, hexbin is installed

Results

Table 1: Summary of simply linear regression on comparison of weekly earning and education and experience for the dataset ex1029

Summary of simply linear regression on comparison of weekly earning and education and experience for the dataset ex1029

```
Call:
lm(formula = WeeklyEarnings ~ Educ + Exper, data = ex1029)

Residuals:
    Min      1Q  Median      3Q     Max
-1122.0  -214.3   -51.8   146.5 18114.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -387.1435    13.6058  -28.45   <2e-16 ***
Educ          63.6179     0.8964   70.97   <2e-16 ***
Exper         10.5982     0.2112   50.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 399.2 on 25434 degrees of freedom
Multiple R-squared:  0.1916,    Adjusted R-squared:  0.1916
F-statistic:  3015 on 2 and 25434 DF,  p-value: < 2.2e-16
```

Table 1 shows that the Pr(>|t|) is <2e-16, which means the null hypothesis that the intercept term is zero can be rejected as there is sufficient evidence. The Pr(>|t|) of education and experience is <2e-16, which means the null hypothesis that the coefficient of education and experience are zero and can be rejected as there is strong evidence. Therefore, the null hypothesis that there is no relationship can be rejected. There is a low adjusted R-squared, which is 0.1916.

Figure 1: Residual plot for Regression in Table 1

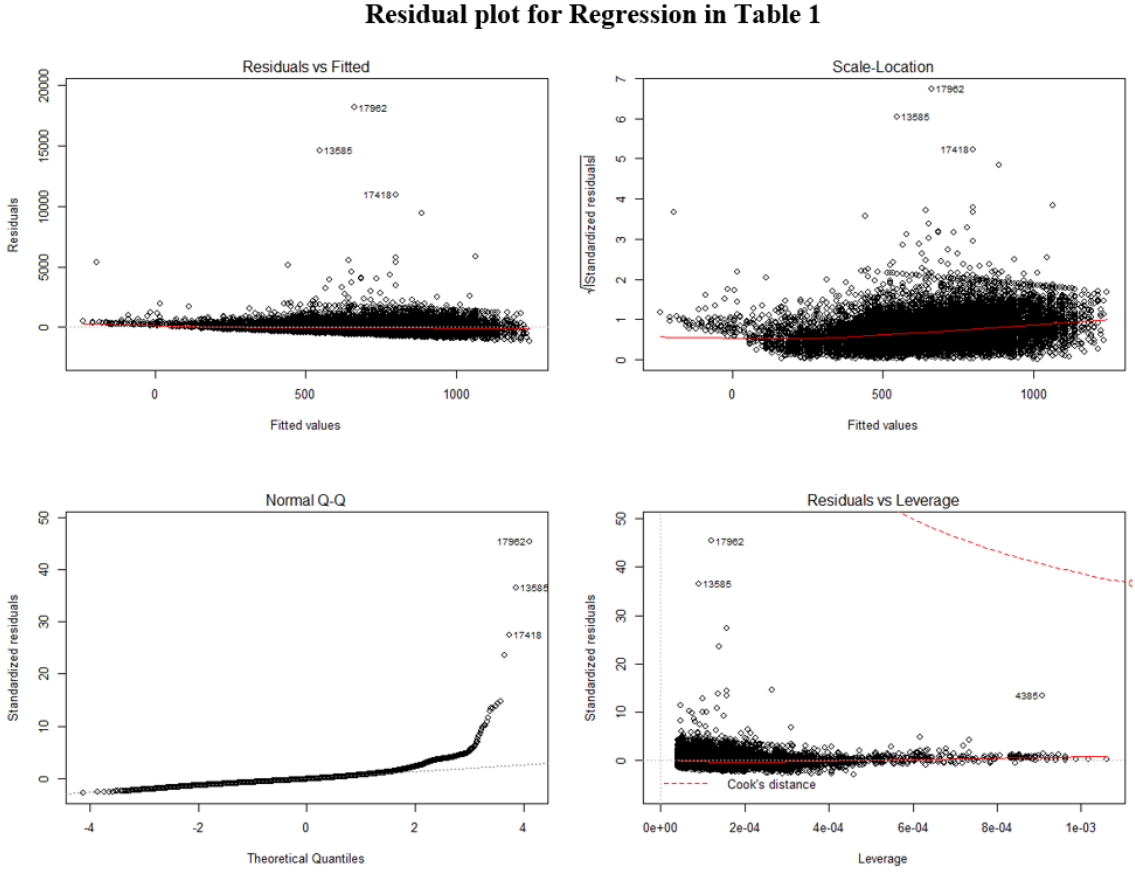


Figure 1 shows there is not enough homoscedasticity in the top-left figure (Residuals vs Fitted) and top-right figure (Scale-Location). The bottom-left figure (Normal Q-Q) shows a little curve which means the data are skewed a little bit. No influential outlier is shown in the bottom-right figure (Residuals vs Leverage).

Table 2: Summary of log regression on comparison of weekly earning and education and experience for the dataset ex1029

Summary of log regression on comparison of weekly earning and education and experience for the dataset ex1029

```
Call:
lm(formula = log(WeeklyEarnings) ~ Educ + Exper, data = ex1029)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6930 -0.3084  0.0461  0.3524  3.6055

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.6176140  0.0183499  251.64   <2e-16 ***
Educ        0.1015493  0.0012089   84.00   <2e-16 ***
Exper       0.0179128  0.0002848   62.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5385 on 25434 degrees of freedom
Multiple R-squared:  0.2556,    Adjusted R-squared:  0.2555
F-statistic:  4366 on 2 and 25434 DF,  p-value: < 2.2e-16
```

Table 2 shows that the Pr(>|t|) is <2e-16, which means the null hypothesis that the intercept term is zero can be rejected as there is sufficient evidence. The Pr(>|t|) of education and experience are <2e-16, which means the null hypothesis that the coefficient of education and experience are zero and can be rejected as there is strong evidence. Therefore, the null hypothesis that there is no relationship can be rejected. There is a low adjusted R-squared, which is 0.2555.

Figure 2: Residual plot for Regression in Table 2

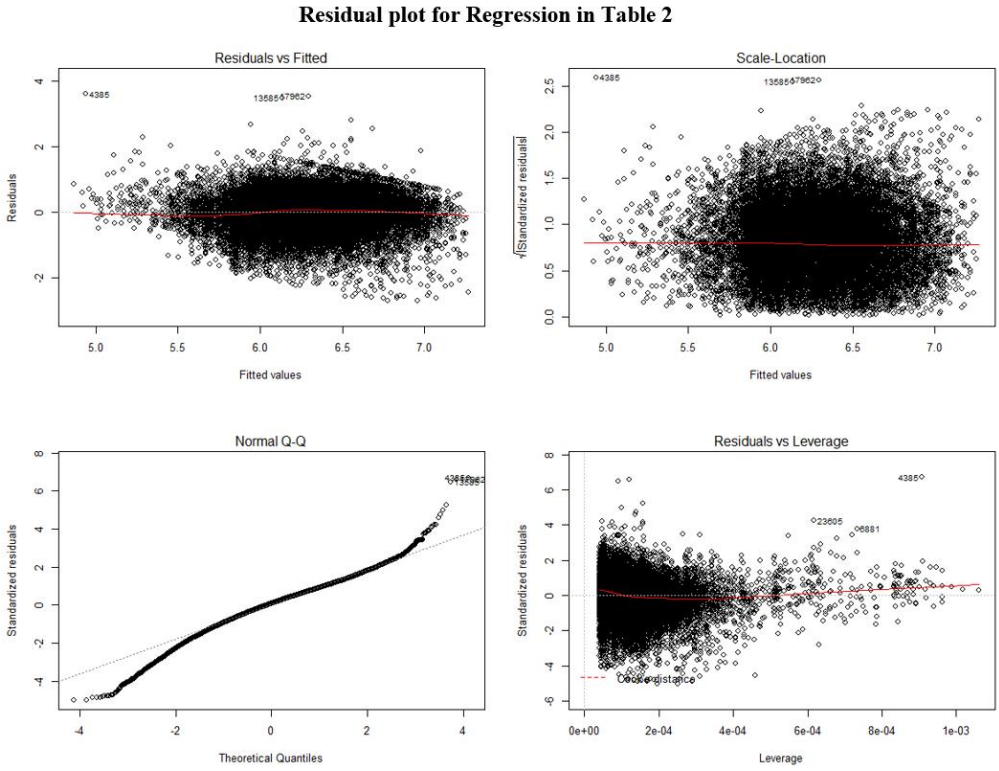
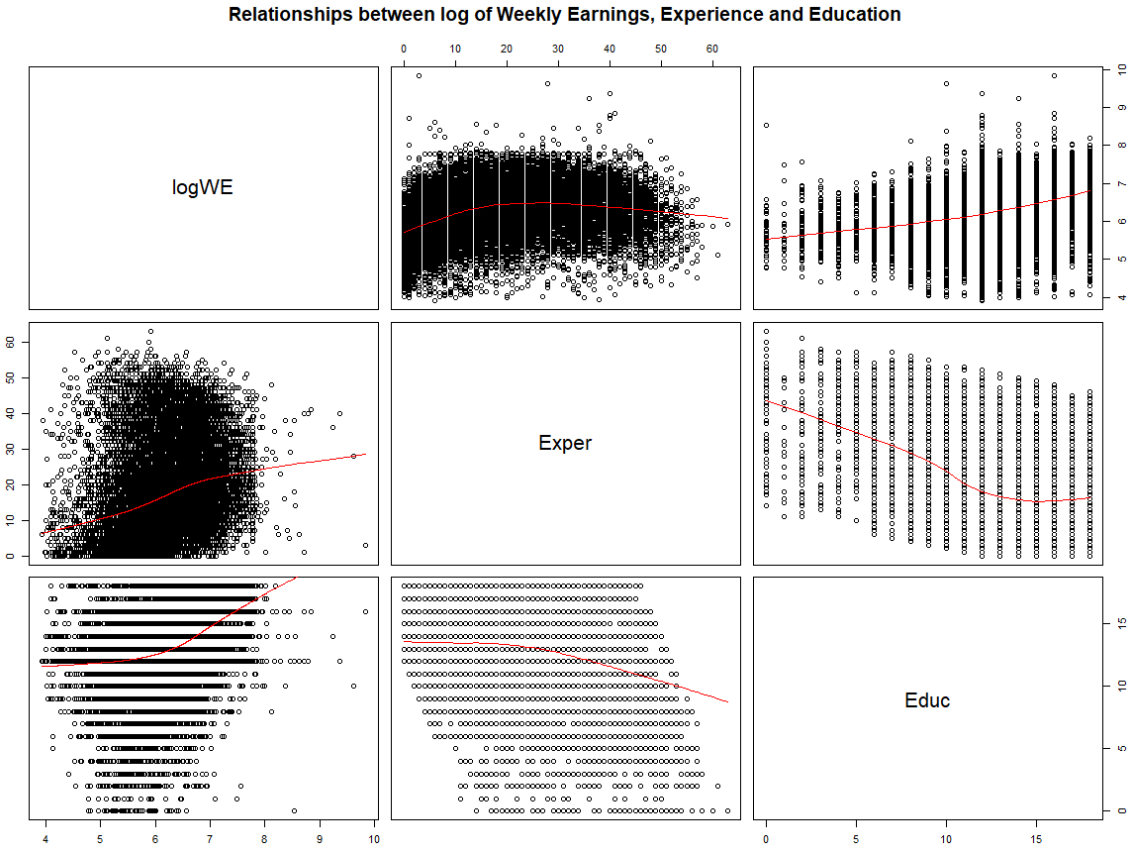


Figure 2 shows there is particular pattern (homoscedasticity) in the top-left figure (Residuals vs Fitted) and top-right figure (Scale-Location). The residuals are distributed normally as the bottom-left figure (Normal Q-Q) almost show a straight line. No influential outlier is shown in the bottom-right figure (Residuals vs Leverage).

Since the adjusted R-squared of log regression (0.2555) (Table 1) is 0.0639 better than the adjusted R-squared of simply linear regression (0.1916) (Table 2), and the log regression provides a better residual plot (Figure 2). Therefore, the log regression on comparison of weekly earning and education and experience for the dataset ex1029 will be used in this project.

Figure 3: Pairs plot for log of weekly earnings, experience and education



In Figure 3, the pair plot of log of weekly earnings and experience (the middle plot of first row) shows a cursive line. Therefore, a polynomial regression model will be used in this project.

Figure 4: Boxplot for log of weekly earnings and metropolitan status of black and not black people

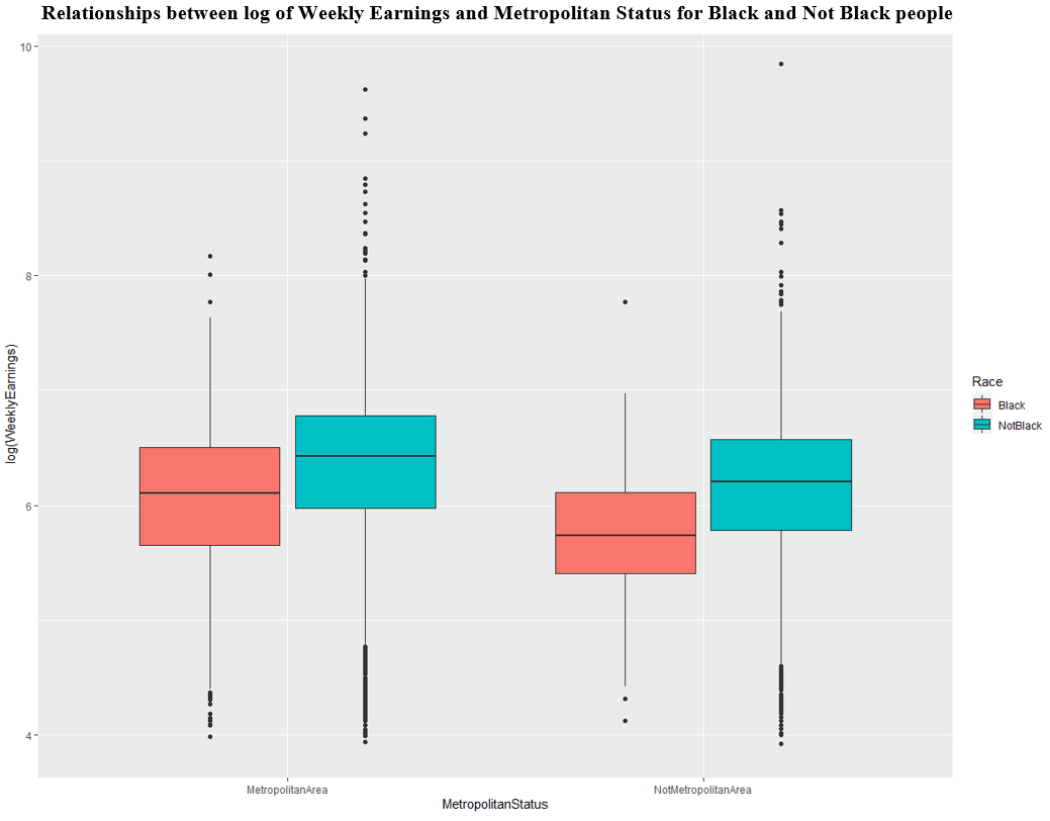


Figure 4 shows in the metropolitan area and not metropolitan area, the log of weekly earning of black and not black people have overlap of mean, which mean there is no sufficient different between black and not black people in metropolitan status and not metropolitan status.

Figure 5: Ggplot for the log of weekly earning, education and metropolitan status of black and not black people

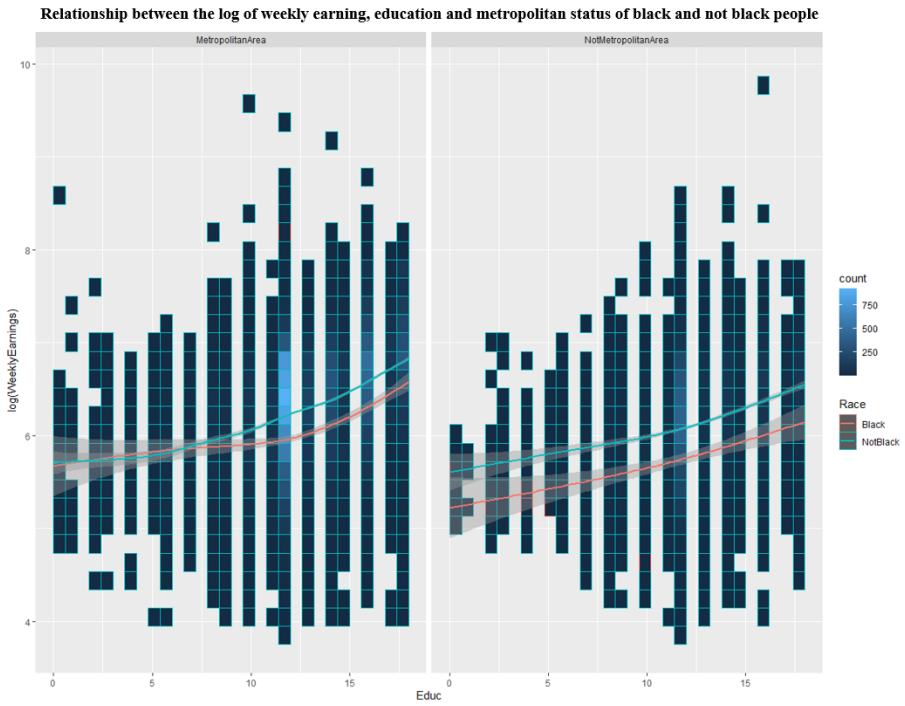


Figure 5 shows in the metropolitan area, with less than seven years of education, black and not black people have the similar weekly earning, with more than seven years of education, not black people have more weekly earning than black people. In the not metropolitan area, with the same education, not black people have more weekly earnings than not black people.

Figure 6: Ggplot for the log of weekly earning, experience and metropolitan status of black and not black people

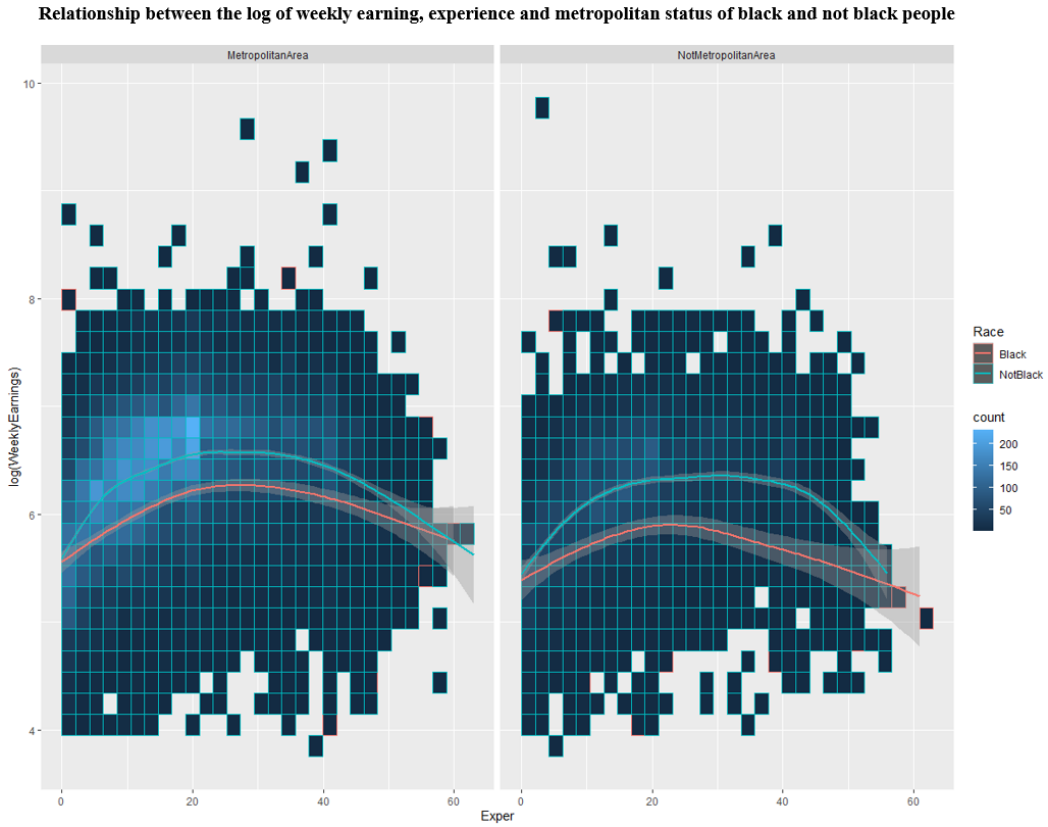


Figure 6 shows in the metropolitan area and not in the metropolitan area, black and not black people have the similar weekly earning when they have no experience and more than 55 years of experience. With other years of experience, not black people have more weekly earnings than black people, in the not metropolitan area, the gap is bigger.

Table 3: Summary of log regression on comparison of weekly earning, race, experience, education, I(Exper&2), metropolitan status, region for the dataset ex1029

Summary of log regression on comparison of weekly earning, race, experience, education, I(Exper&2), metropolitan status, region for the dataset ex1029

```
Call:
lm(formula = log(WeeklyEarnings) ~ Race + Exper + Educ + I(Exper^2) +
    MetropolitanStatus + Region, data = ex1029)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7136 -0.2850  0.0349  0.3254  3.9057

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        4.350e+00  2.176e-02 199.957  < 2e-16 ***
RaceNotBlack                       2.352e-01  1.219e-02  19.288  < 2e-16 ***
Exper                              5.496e-02  9.112e-04  60.315  < 2e-16 ***
Educ                               8.862e-02  1.172e-03  75.597  < 2e-16 ***
I(Exper^2)                        -8.356e-04  1.958e-05 -42.681  < 2e-16 ***
MetropolitanStatusNotMetropolitanArea -1.648e-01  7.433e-03 -22.167  < 2e-16 ***
RegionNortheast                    4.297e-02  9.374e-03   4.584 4.58e-06 ***
RegionSouth                       -6.147e-02  8.746e-03  -7.029 2.14e-12 ***
RegionWest                        -1.136e-02  9.507e-03  -1.195    0.232
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5106 on 25428 degrees of freedom
Multiple R-squared:  0.3307,    Adjusted R-squared:  0.3305
F-statistic:  1570 on 8 and 25428 DF,  p-value: < 2.2e-16
```

Table 3 shows that the Pr(>|t|) is <2e-16, which means the null hypothesis that the intercept term is zero can be rejected as there is sufficient evidence. The Pr(>|t|) of not black people, experience, education, I(Exper&2), metropolitan status, region of Northeast and South are <2e-16, which means the null hypothesis that the coefficient of not black people, experience, education, I(Exper&2), metropolitan status, region of Northeast and South are zero and can be rejected as there is strong evidence. Therefore, the null hypothesis that there is no relationship can be rejected. There is a low adjusted R-squared, which is 0.3305.

Figure 7: Residual plot for Regression in Table 3

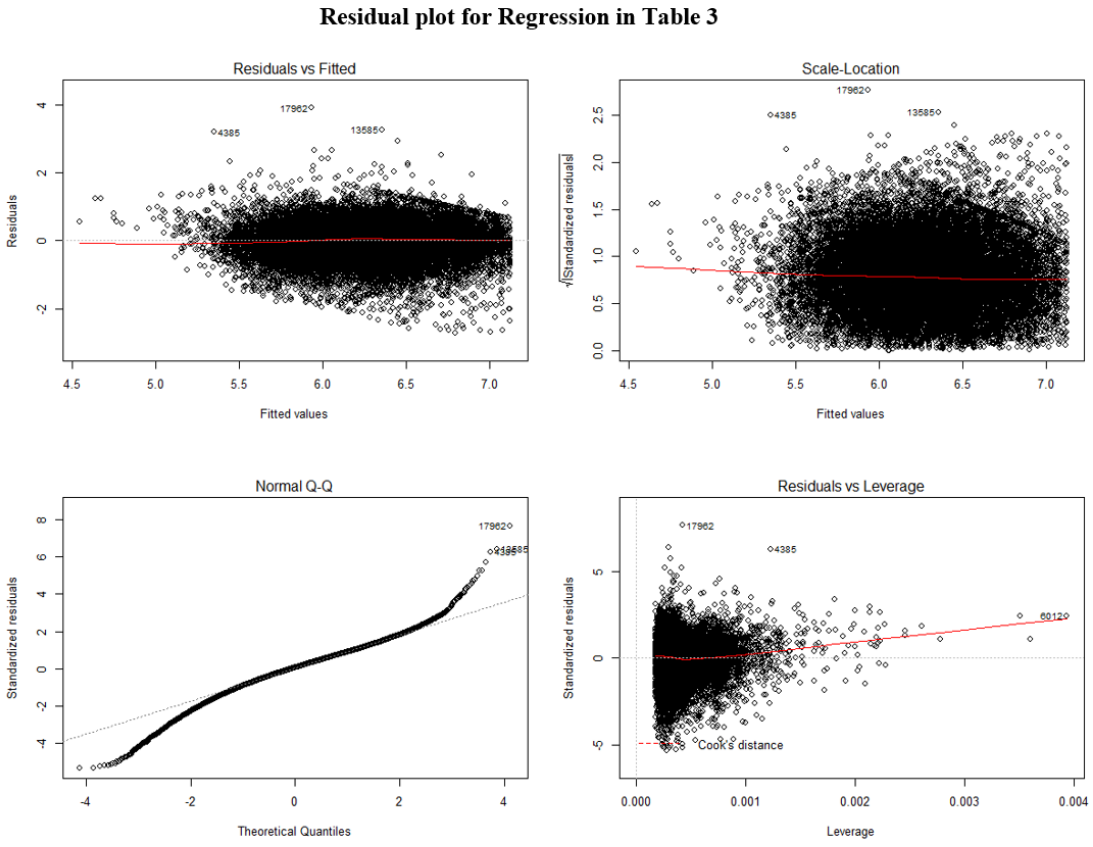


Figure 7 shows there is particular pattern (homoscedasticity) in the top-left figure (Residuals vs Fitted) and top-right figure (Scale-Location). The residuals are distributed normally as the bottom-left figure (Normal Q-Q) almost show a straight line. No influential outlier is shown in the bottom-right figure (Residuals vs Leverage).

Table 4: Confidence interval of polynomial regression model

Confidence interval of polynomial regression model

	2.5 %	97.5 %
(Intercept)	74.2788437	80.8921066
RaceNotBlack	1.2352900	1.2957746
Exper	1.0546146	1.0583886
Educ	1.0901550	1.0951762
I(Exper^2)	0.9991264	0.9992030
MetropolitanStatusNotMetropolitanArea	0.8358251	0.8605380
RegionNortheast	1.0249053	1.0632702
RegionSouth	0.9243947	0.9566378
RegionWest	0.9704471	1.0072985

Table 4 shows all the confidence interval do not include zero so they are all significant.