Labor Market Analysis

Exploratory Data Analysis

Marjorie Blanco, Joe Thomson, Haodi Tu

Data

We used data from the 2016 ACS for Puerto Rico to examine wage gaps between individuals with different education levels. Our research questions are: 1) How do earnings vary by education level? 2) How does the premium for education vary by gender? The 2016 ACS is a nationally representative sample of 5194. The household survey includes questions pertaining to each household member's demographic characteristics and labor market activity.

We restrict our sample to these three racial groups: White, Black and Other. In addition, given our goal of examining earning differences by gender and marital status and the reporting of earnings in the ACS on an annual basis (wages, salary, commissions, bonuses, tips, and self-employment income during the past 12 months), we restrict our sample to full-time year-round (FTYR) workers. We define FTYR workers as individuals who report positive earnings over the past year, who worked at least 40 of the past 52 weeks, and who worked at least 35 hours per week in a usual work week over this period.

EDA Insights:

For our exploratory analysis we looked at population breakdowns by education, age, marital status, gender, race, earnings, and work hours. We applied filters on education (HS diploma or above), age (18-64), and work hours (>35/week).

An earnings histogram identified a default maximum amount of earnings (189k) which we also filtered out of the data. The earning distribution is progressive above the median, but drops off sharply below the median, likely indicating the presence of a minimum wage. The correlation between age and earnings is very weak (.23). Likewise, earnings is very weakly correlated with hours worked among those who work more than 35 hours per week. However, white individuals appear to have an earnings premium over other races, and both married and divorced individuals appear to have an earnings premium over those who have never been married. Given that the correlation between age and earnings was weak, this may be due to other qualitative factors possessed by those who get married. Married was recategoried to married, divorced and never married. Men also appear to earn a small premium over women.

The age distribution of full time workers is skewed towards older adults, possibly indicating that younger workers have trouble finding full-time work, wait to enter the workforce, or are leaving the territory.

Preliminary Econometric Estimates

First Model

$$Earning = \beta_0 + Widowed * \beta_1 + Divorced * \beta_2 + Separated * \beta_3 + NeverMarried * \beta_4 + RaceBlack * \beta_5 + RaceOther * \beta_6 + SomeCollege * \beta_7 + Associate * \beta_8 + Bachelor * \beta_9 + Master * \beta_10 + Professional * \beta_11 + Doctoral * \beta_12 + Age * \beta_13 + Age * Age\beta_14$$

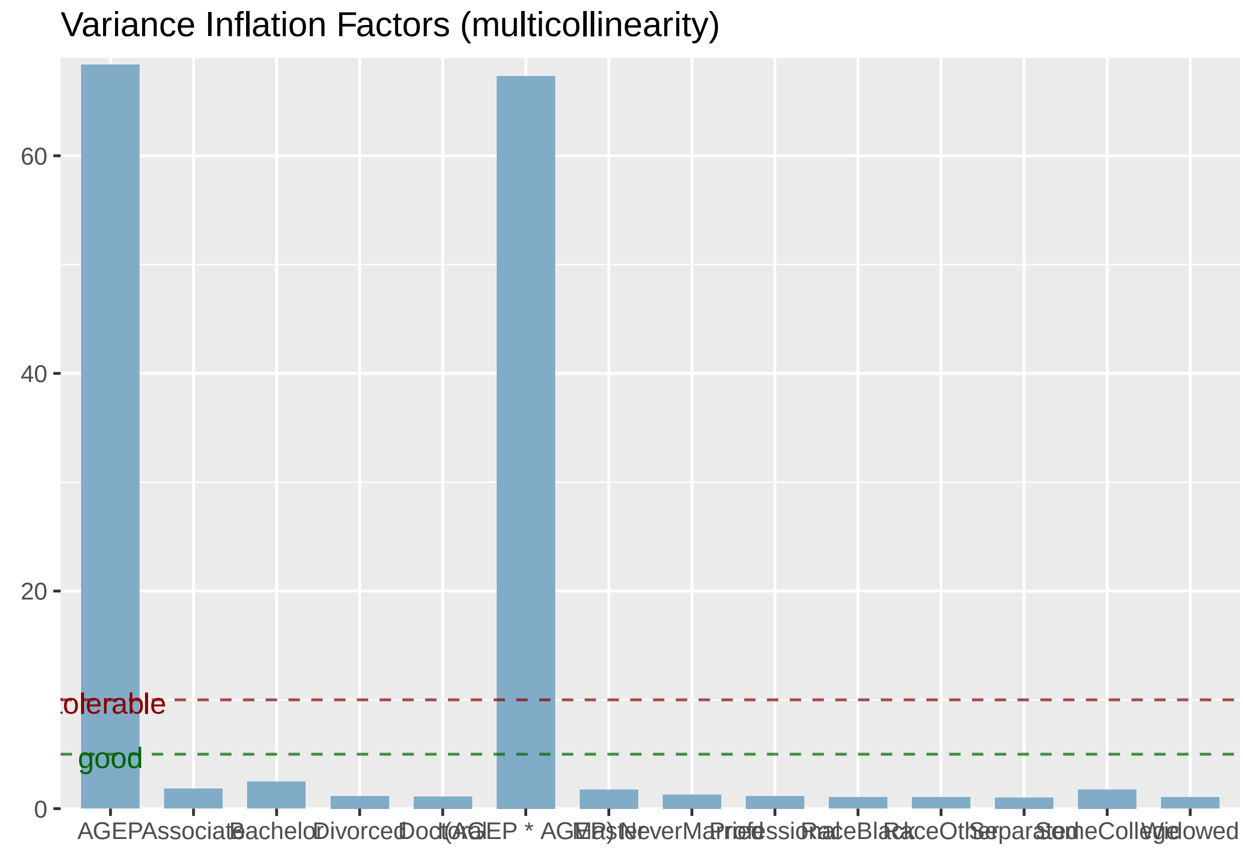
Stratified Model Gender

Gender: Female

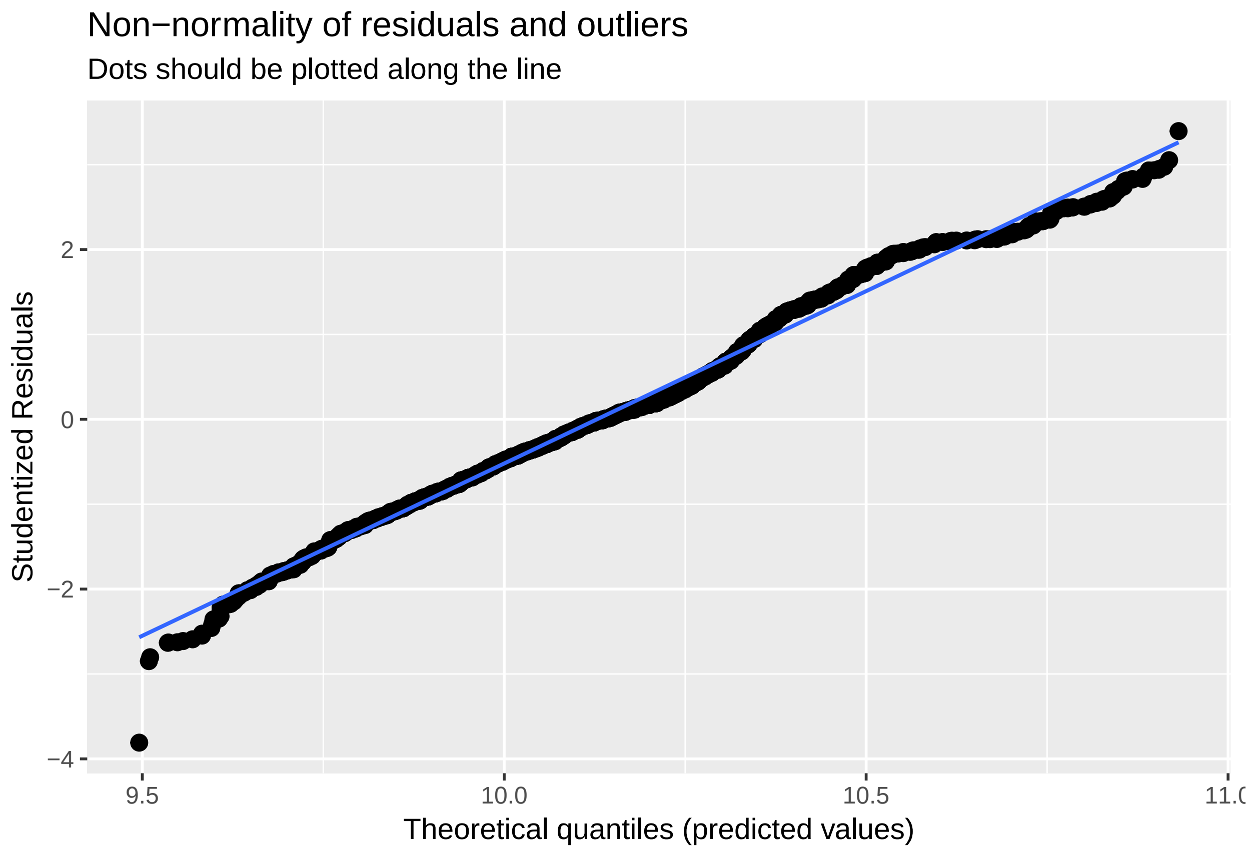
##

```
## Call:
## lm(formula = log(PERNP) ~ Widowed + Divorced + Separated + NeverMarried +
##     RaceBlack + RaceOther + SomeCollege + Associate + Bachelor +
##     Master + Professional + Doctoral + AGEP + I(AGEP * AGEP),
##     data = ss16ppr_female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60623 -0.29348 -0.02351  0.24300  1.44216
##
## Coefficients:
##                   Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)     9.29081116  0.14197142  65.441 < 0.0000000000000002 ***
## Widowed         0.12617920  0.06182067   2.041             0.041348 *
## Divorced       -0.00769942  0.02211596  -0.348             0.727765
## Separated      -0.03799682  0.05852598  -0.649             0.516248
## NeverMarried   -0.03120139  0.02141248  -1.457             0.145196
## RaceBlack      -0.04721312  0.02432301  -1.941             0.052357 .
## RaceOther      -0.08455957  0.02373290  -3.563             0.000373 ***
## SomeCollege     0.12652738  0.03293833   3.841             0.000125 ***
## Associate       0.13857783  0.03200975   4.329            0.0000155 ***
## Bachelor        0.38951105  0.02684627  14.509 < 0.0000000000000002 ***
## Master          0.55873934  0.03352996  16.664 < 0.0000000000000002 ***
## Professional    0.87022193  0.05954679  14.614 < 0.0000000000000002 ***
## Doctoral        0.91227935  0.06429409  14.189 < 0.0000000000000002 ***
## AGEP            0.01640263  0.00666118   2.462             0.013866 *
## I(AGEP * AGEP) -0.00007842  0.00007625  -1.028             0.303816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4267 on 2552 degrees of freedom
## Multiple R-squared:  0.2492, Adjusted R-squared:  0.245
## F-statistic: 60.49 on 14 and 2552 DF,  p-value: < 0.00000000000000022
```

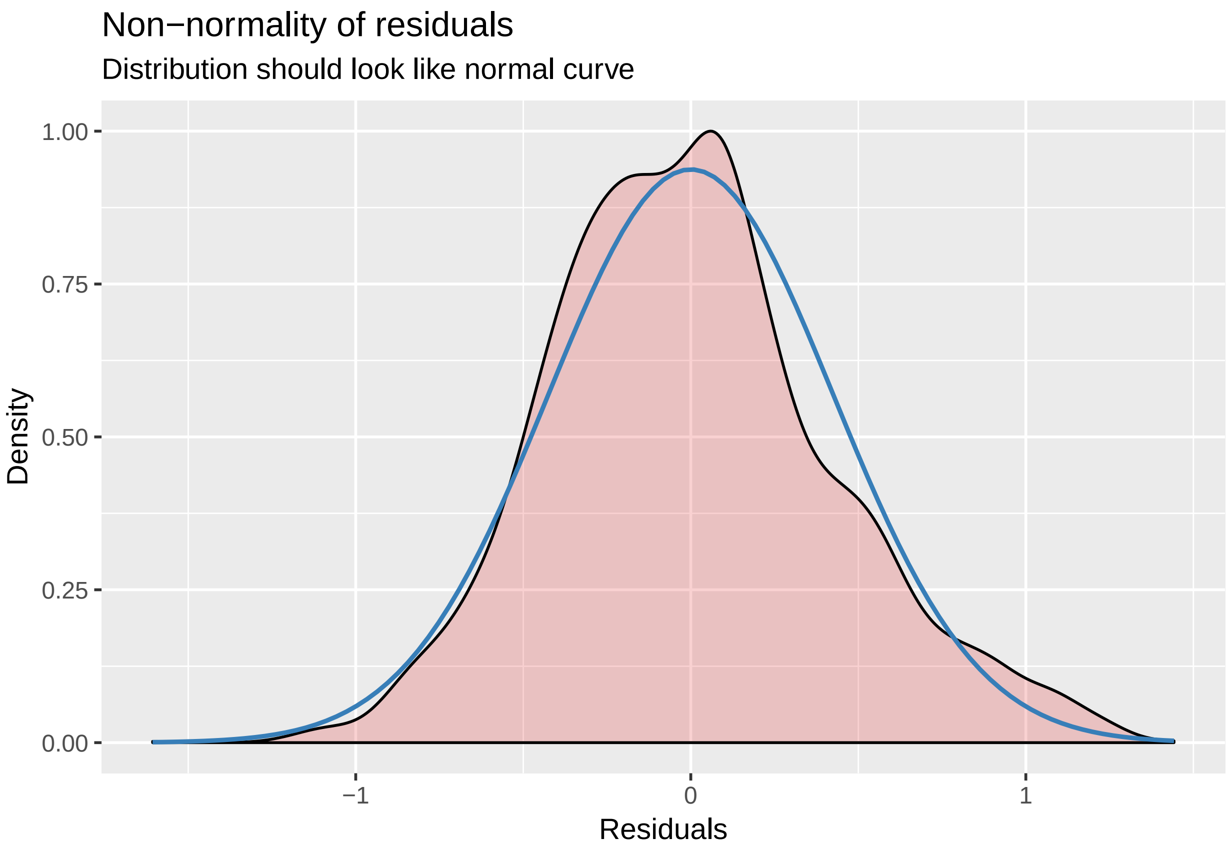
```
## [[1]]
```


Variance Inflation Factors (multicollinearity)
60
40
20
0
olerable
good
AGEP
Associate
Bachelor
Divorced
Doctoral
I(AGEP * AGEP)
Master
NeverMarried
Professional
RaceBlack
RaceOther
Separated
SomeCollege
Widowed

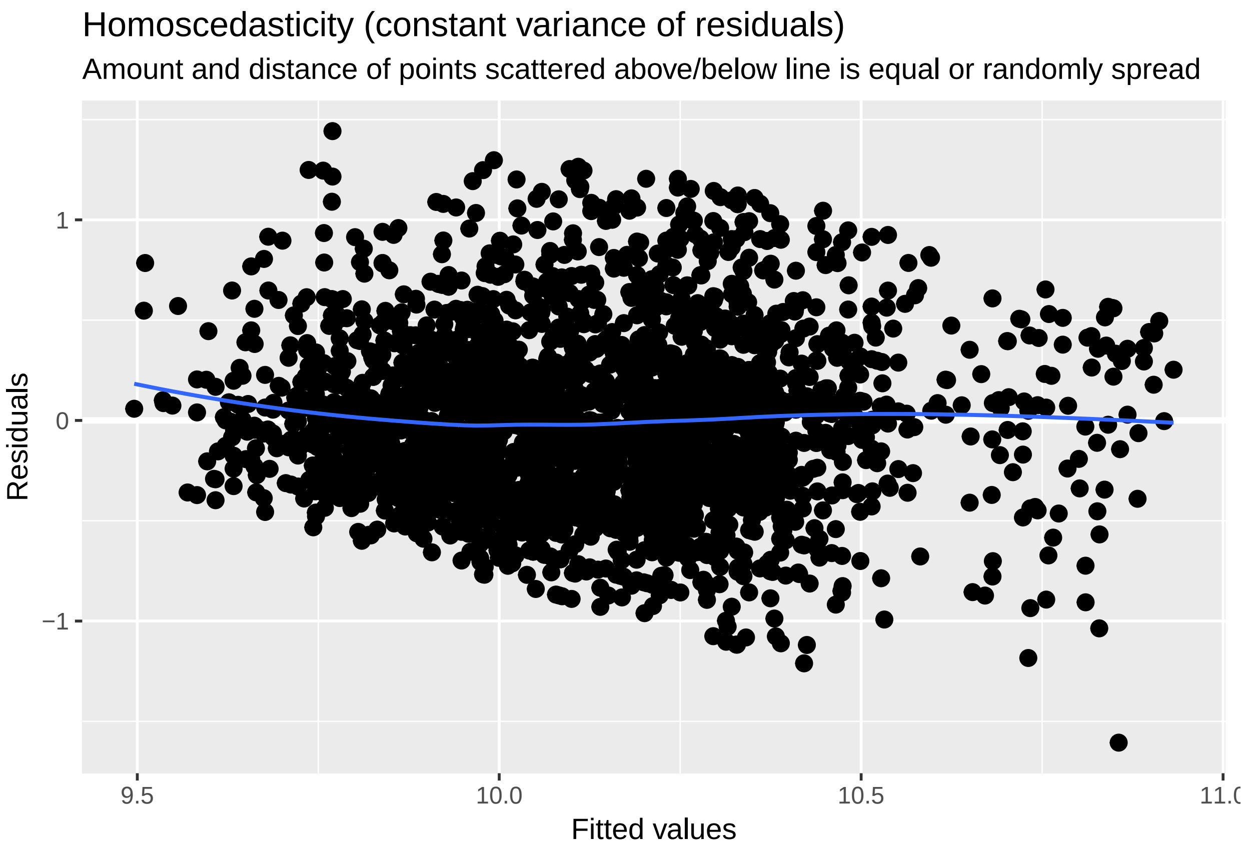
```
##
## [[2]]
```


Non−normality of residuals and outliers
Dots should be plotted along the line
Studentized Residuals
2
0
−2
−4
9.5
10.0
10.5
11.0
Theoretical quantiles (predicted values)

```
##
## [[3]]
```


Non−normality of residuals
Distribution should look like normal curve
Density
1.00
0.75
0.50
0.25
0.00
−1
0
1
Residuals

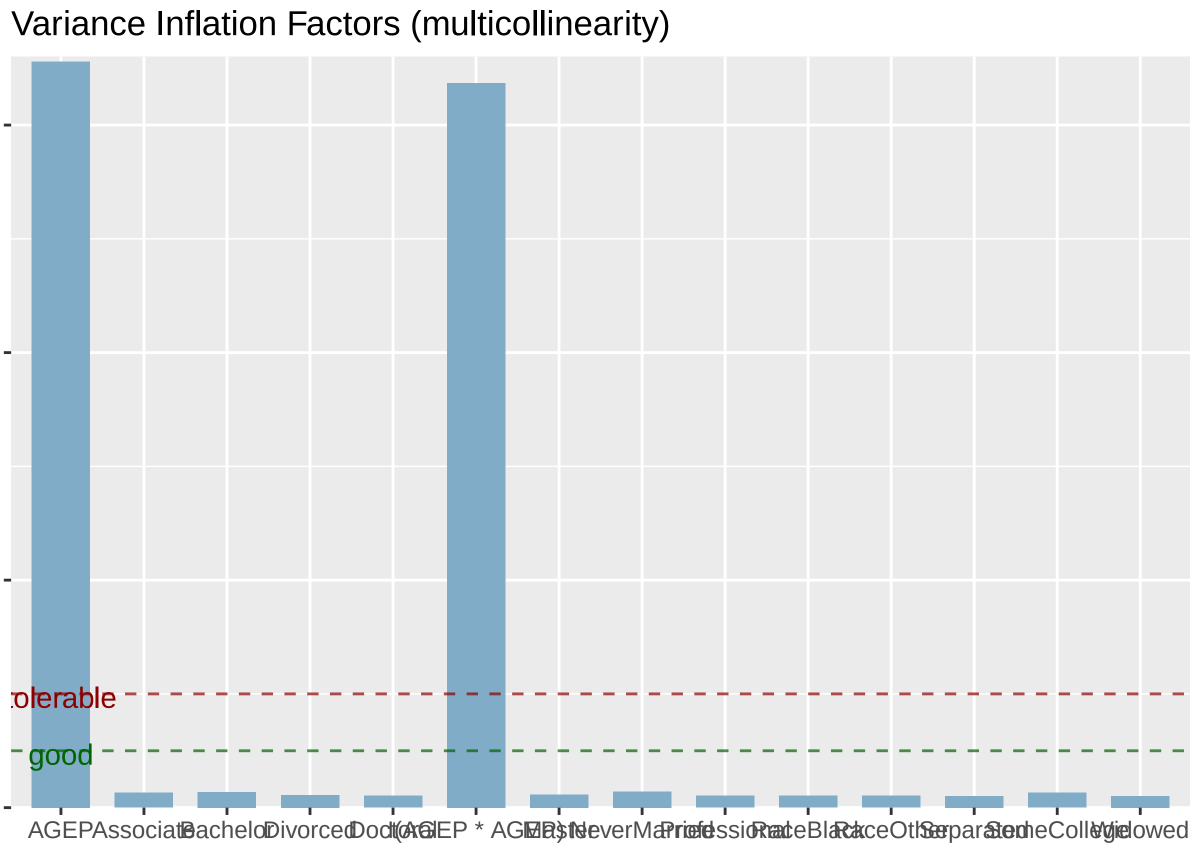
```
##
## [[4]]
```



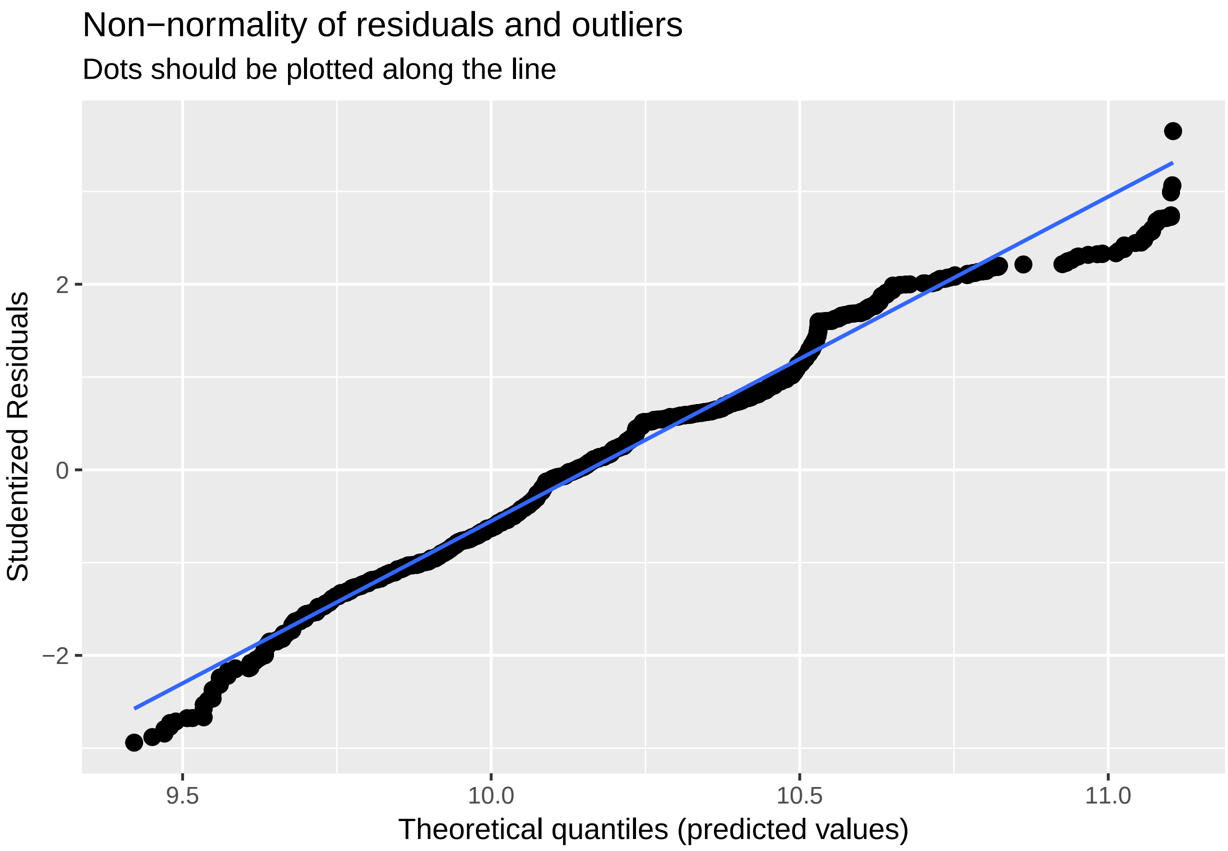
Gender: Male

```
##
## Call:
## lm(formula = log(PERNP) ~ Widowed + Divorced + Separated + NeverMarried +
##     RaceBlack + RaceOther + SomeCollege + Associate + Bachelor +
##     Master + Professional + Doctoral + AGEP + I(AGEP * AGEP),
##     data = ss16ppr_male)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33073 -0.32575 -0.02096  0.30213  1.66547
##
## Coefficients:
##                  Estimate Std. Error t value             Pr(>|t|)
## (Intercept)     8.8897325  0.1389042  63.999 < 0.0000000000000002 ***
## Widowed        -0.2436797  0.1237557  -1.969              0.04905 *
## Divorced       -0.0777776  0.0264188  -2.944              0.00327 **
## Separated      -0.0481621  0.0667929  -0.721              0.47093
## NeverMarried   -0.1449325  0.0233668  -6.203     0.00000000064405 ***
## RaceBlack      -0.0182384  0.0253545  -0.719              0.47200
## RaceOther      -0.0367754  0.0250006  -1.471              0.14142
## SomeCollege     0.1563020  0.0270663   5.775     0.00000000861667 ***
## Associate       0.1459627  0.0273833   5.330     0.00000010640493 ***
## Bachelor        0.4414121  0.0241685  18.264 < 0.0000000000000002 ***
## Master          0.5617214  0.0379129  14.816 < 0.0000000000000002 ***
## Professional    0.7341881  0.0656264  11.187 < 0.0000000000000002 ***
## Doctoral        1.0162476  0.0729256  13.935 < 0.0000000000000002 ***
## AGEP             0.0452870  0.0065397   6.925     0.00000000000547 ***
## I(AGEP * AGEP) -0.0004275  0.0000751  -5.691     0.00000001400332 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4582 on 2612 degrees of freedom
## Multiple R-squared:  0.2788, Adjusted R-squared:  0.275
## F-statistic: 72.13 on 14 and 2612 DF,  p-value: < 0.0000000000000022
```

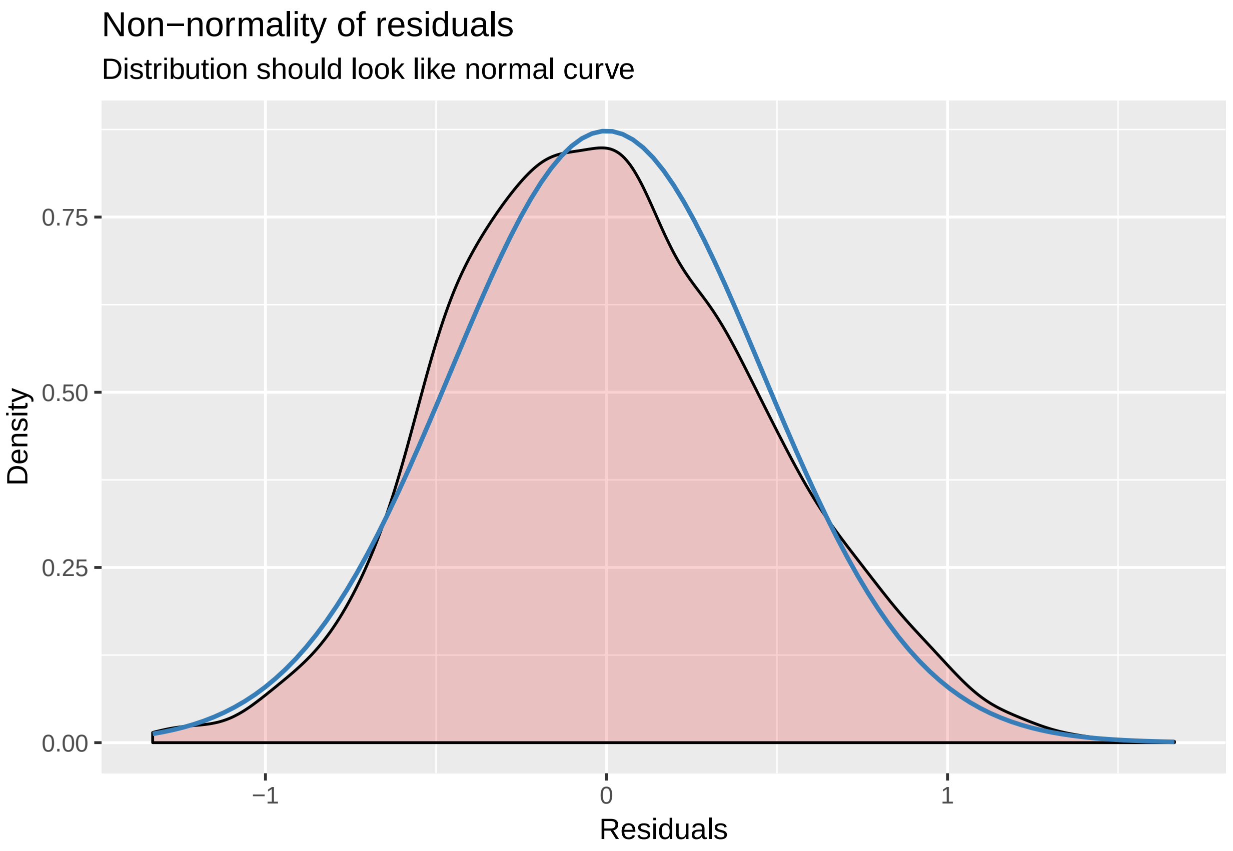
```
## [[1]]
```



```
##
## [[2]]
```


Non−normality of residuals and outliers
Dots should be plotted along the line
Studentized Residuals
2
0
−2
9.5
10.0
10.5
11.0
Theoretical quantiles (predicted values)

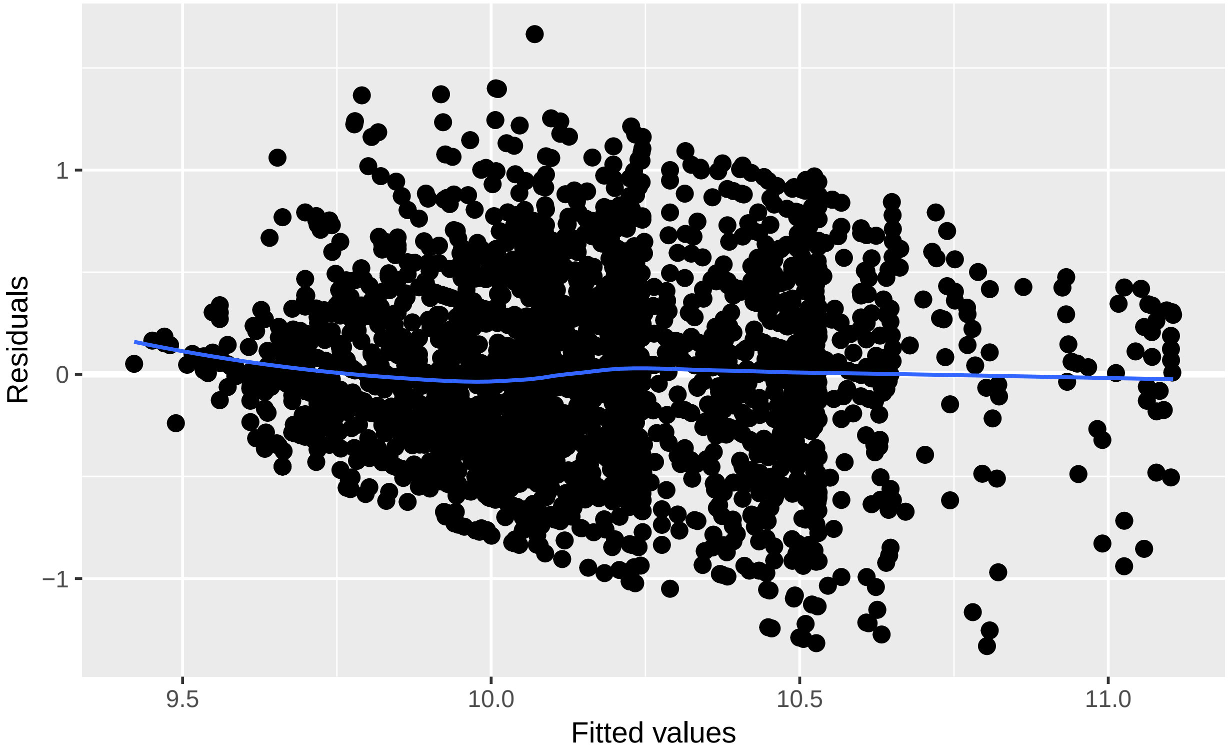
```
##
## [[3]]
```


Non−normality of residuals
Distribution should look like normal curve
Density
0.75
0.50
0.25
0.00
−1
0
1
Residuals

```
##
## [[4]]
```

Homoscedasticity (constant variance of residuals)

Amount and distance of points scattered above/below line is equal or randomly spread


Residuals
Fitted values