Basic Econometric

Jackie Kim

Basic Econometric

This is my basic practice at working with Econometric I learned in the class. First, we will donwload packages *moments*, *PoEdata*, and *tidyverse*.

```
require(moments)
require(PoEdata)
require(tidyverse)
```

Extract Data and Analyze in terms of Wage and Education

Now, let's extract a data of CPS from PoEdata:

```
data(cps)
head(cps,6)
```

```
##   wage educ age exper female black white married union northeast midwest
## 1 1.05   12  37    19      0     0     1       0     1         0       0
## 2 1.05   13  42    23      0     0     1       0     0         1       0
## 3 1.23    8  54    40      0     0     1       0     0         0       0
## 4 1.28   10  59    43      1     0     1       1     1         1       0
## 5 1.34   18  28     4      1     0     1       0     0         0       0
## 6 1.47   12  40    22      1     0     1       1     1         0       1
##   south west fulltime metro
## 1     0    1        1     1
## 2     0    0        1     1
## 3     0    1        1     1
## 4     0    0        0     1
## 5     0    1        1     0
## 6     0    0        0     1
```

Notice that columns except wage, education, age, and experience are **Categorical Variables**.

Next, create a linear model to see the releationship between Wage and Education:

```
# Create a linear model
cpsm = lm(wage ~ educ, data = cps)
summary(cpsm)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = cps)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.282  -3.728  -1.188   2.382  63.088
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.20260    0.46549  -11.18   <2e-16 ***
## educ         1.15692    0.03446   33.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.585 on 4731 degrees of freedom
## Multiple R-squared:  0.1924, Adjusted R-squared:  0.1923
## F-statistic:  1127 on 1 and 4731 DF,  p-value: < 2.2e-16
```

Interpretation: For additional year of education, the average wage is expected to increase by $0.17. Hence, the linear model formula is:

$$Education = 11.608 + 0.166 Wage$$

However, note that $R^2 = 19.24\%$, so each variation of y does not seem to explain the model well.

For simplicity,

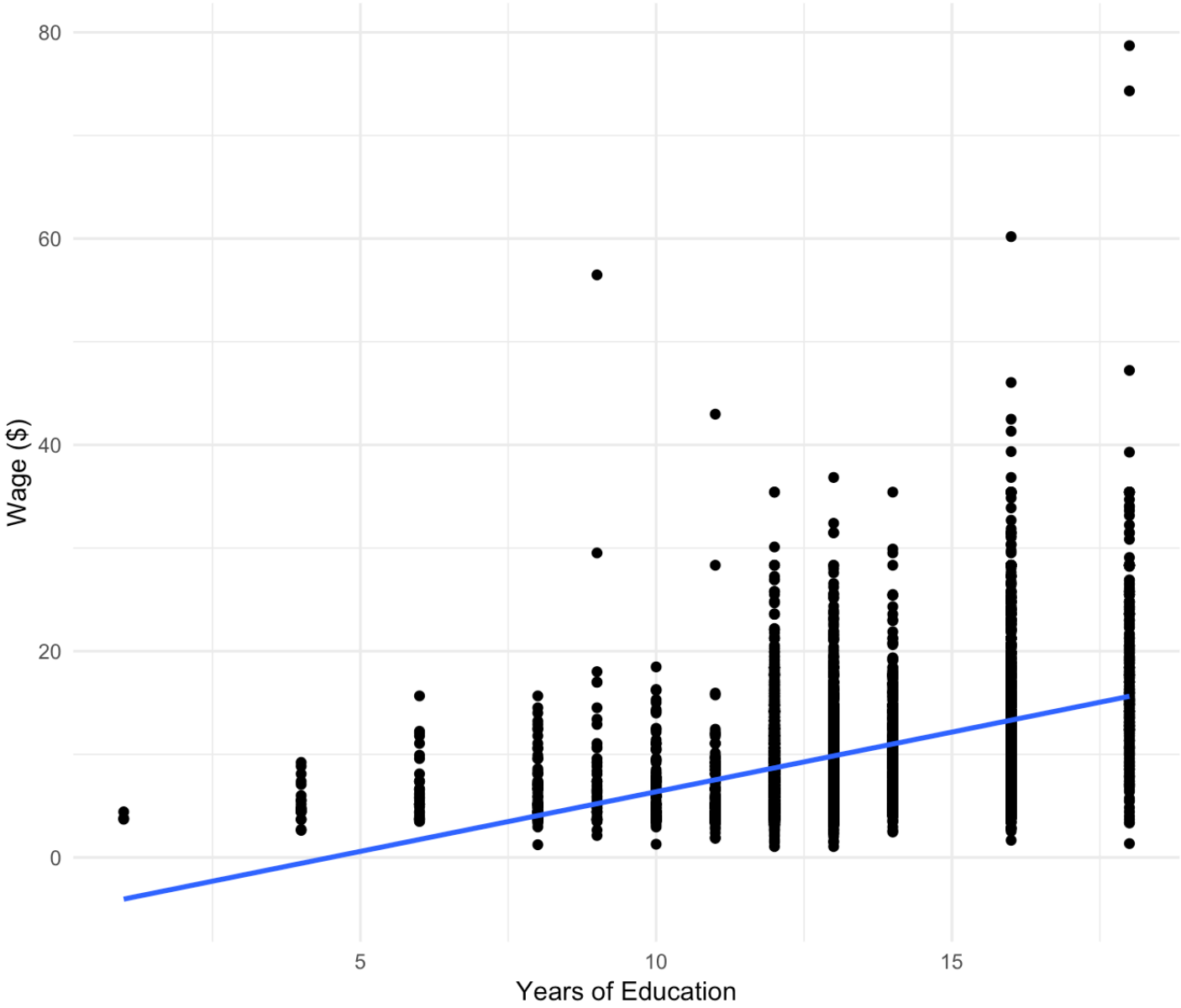
```
smod1 <- summary(cpsm)
smod1$coefficients
```

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) -5.202605 0.46548598 -11.17672  1.207751e-28
## educ         1.156924 0.03445659  33.57628 7.042095e-222
```

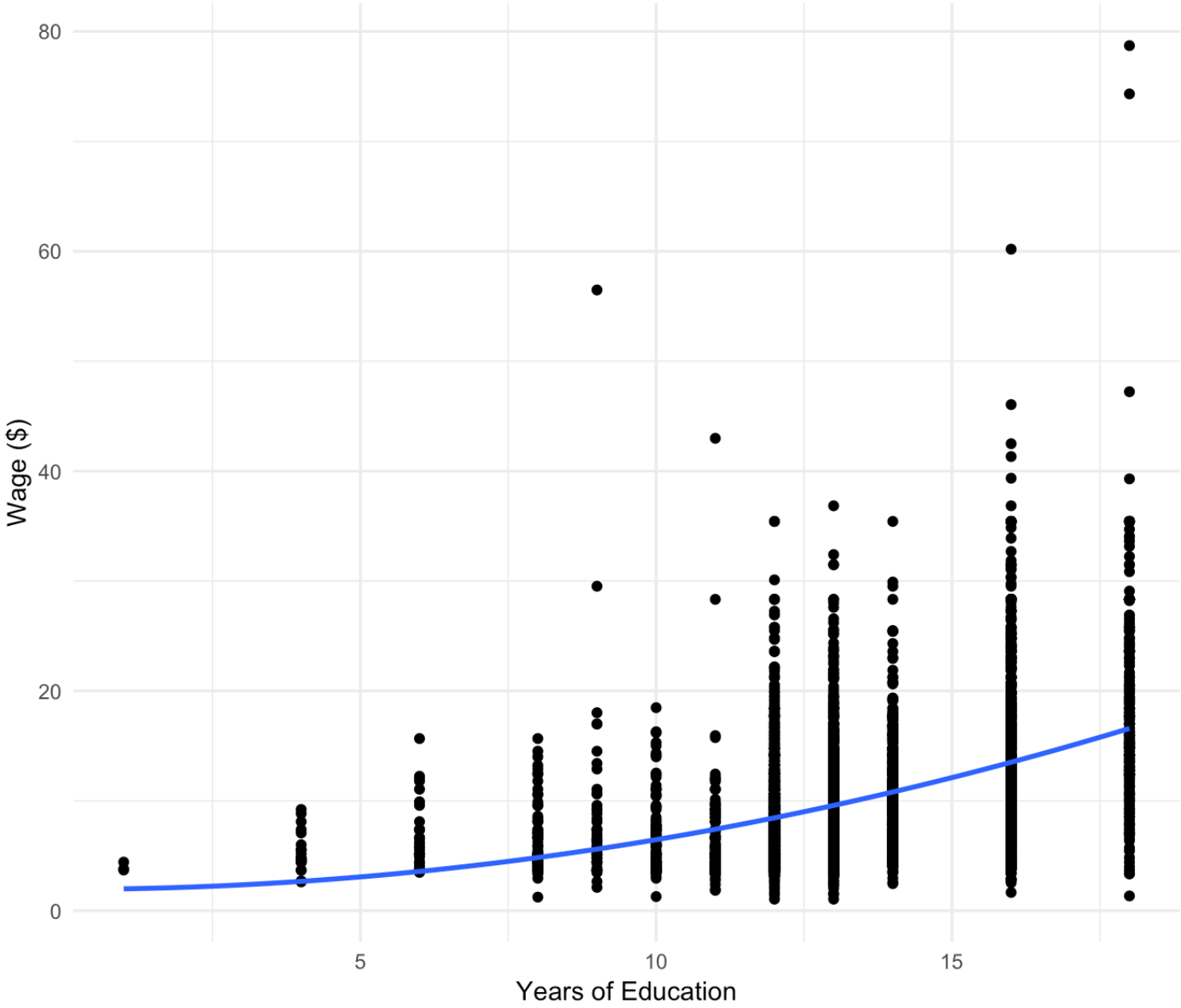
Graph the Linear Model

Next, we are going to plot a scatter plot to compare Wage and Education:

```
# Regular Linear Model
ggplot(cpsm, aes(x = educ, y = wage)) +
  geom_point() +
  labs(x = "Years of Education",
       y = "Wage ($)") +
  theme_minimal() +
  geom_smooth(method = "lm", se = FALSE)
```

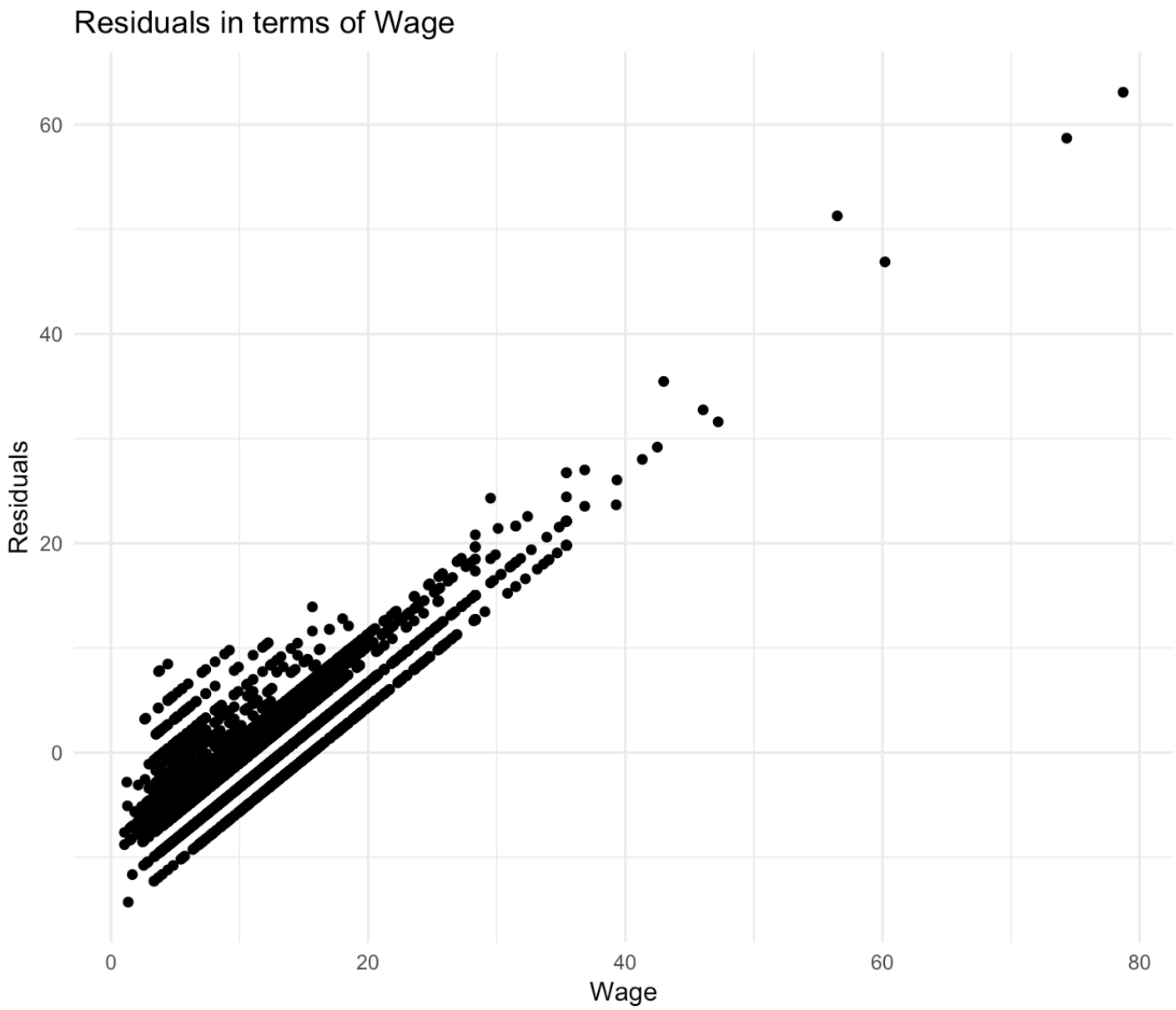



```
# Changing The line as Quadratic Curve
ggplot(cpsm, aes(x = educ, y = wage)) +
  geom_point() +
  labs(x = "Years of Education",
       y = "Wage ($)") +
  theme_minimal() +
  geom_smooth(method = "lm", formula = y ~ I(x^2), se = FALSE)
```



By seeing the graph, making the line quadratic instead of linear does not make the graph better. It seems there is a problem with residuals as we can notice **heteroskedasticity**. To confirm, check residuals graph.

```
# Residual Graph
cpsm.res = resid(cpsm)
ggplot(cpsm, aes(x = wage, y = cpsm.res)) +
  geom_point() +
  labs(x = "Wage",
       y = "Residuals",
       title = "Residuals in terms of Wage") +
  theme_minimal()
```



Notice that residuals are congested around 0 in terms of **Wages**. This result means residuals are not randomly distributed. Hence, residuals are not doing a good job at fitting the model with data.

Conclusion

Considering the fact that coefficient of determination is 19.23% and residuals are not randomly distributed, this data may not be an ideal data to create a linear regression model. The scatter points for the graph model did not seem to show a pattern, showing that Education is not really depended by Wages.