ECON 418 —— Assignment 4 (Problem 2.15)

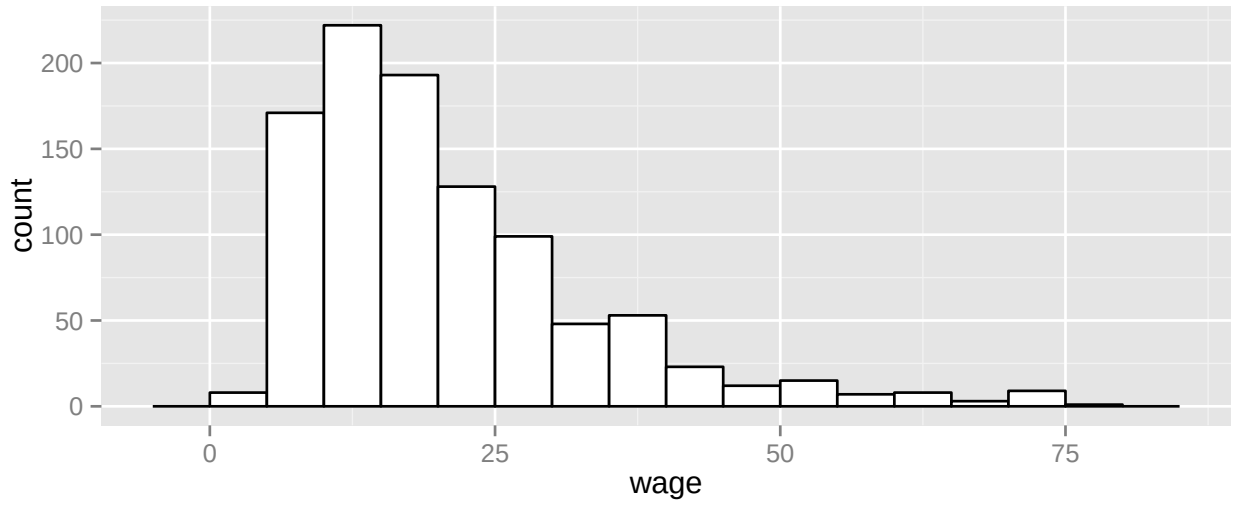
Subasish Das (sxd1684)

5th March 2014

1 Problem 2.15

(a)

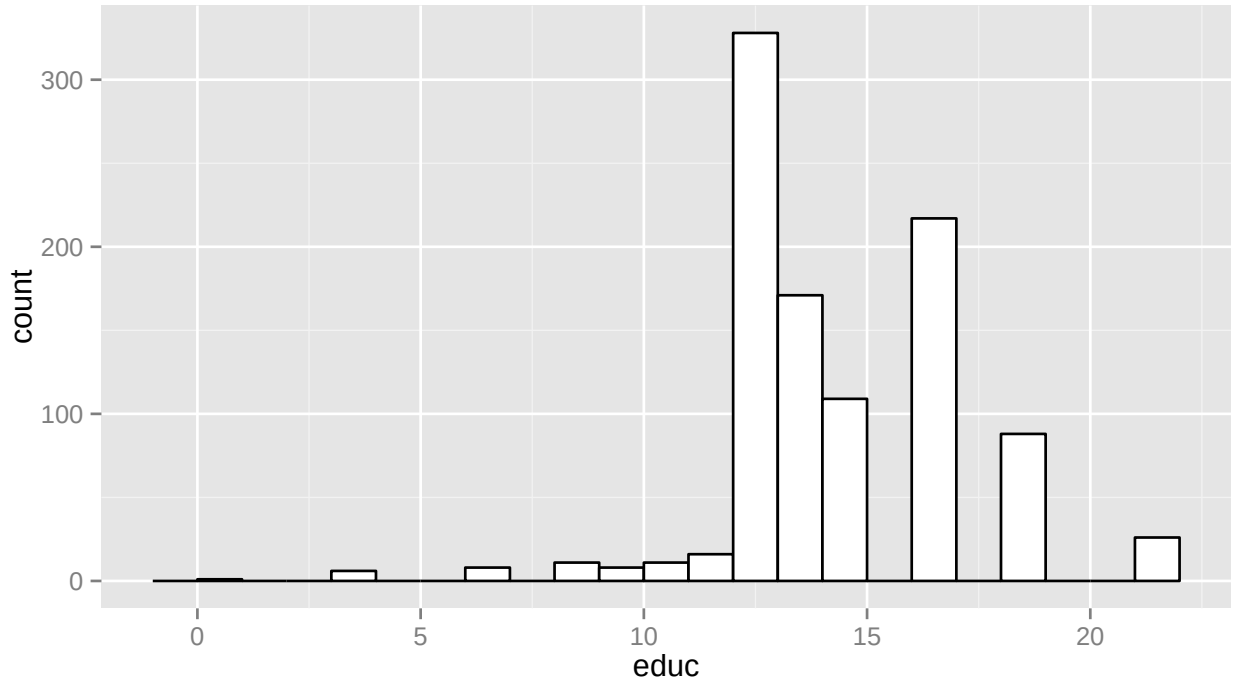
```
setwd("C:/Users/Subasish/Dropbox/A Spring 2014/Dr Sarah/HW")
cps4_small <- read.csv("cps4_small.csv")
head(cps4_small, n = 3)

##    wage educ exper hrswk married female metro midwest south west black
## 1 18.70   16    39    37       1      1     1       0     1    0     0
## 2 11.50   12    16    62       0      0     0       1     0    0     0
## 3 15.04   16    13    40       1      0     1       0     0    1     1
##   asian
## 1     0
## 2     0
## 3     0

cps4_small <- cps4_small[-c(3, 4, 5, 7)]
head(cps4_small, n = 3)

##    wage educ female midwest south west black asian
## 1 18.70   16      1       0     1    0     0     0
## 2 11.50   12      0       1     0    0     0     0
## 3 15.04   16      0       0     0    1     1     0


# Histogram of Wage
library(ggplot2)
ggplot(cps4_small, aes(x = wage)) + geom_histogram(binwidth = 5, colour = "black",
    fill = "white")
```



```
library(psych)

##
## Attaching package:  'psych'
## The following object is masked from 'package:ggplot2':
##
##    %+%

describe(cps4_small$wage)

##   var    n  mean    sd median trimmed   mad  min   max range skew kurtosis
## 1   1 1000 20.62 12.83   17.3   18.67 10.27 1.97 76.39 74.42 1.58     2.91
##     se
## 1 0.41
```

The above plot is skewed to the right. It indicates that majority of the observations are in between the hourly wages of 5 to 40. A smaller number of observations is seen with an hourly wage above 40. The average was is 20.62 dollars per hour. The maximum earned in this sample is 76.39 dollars per hour and the lowest wage is 1.97 dollars per hour. The descriptive statistics also give median, standard deviation, range, kutosis and skewness of the wage.

```
# Histogram of Education
library(ggplot2)
ggplot(cps4_small, aes(x = educ)) + geom_histogram(binwidth = 1, colour = "black",
    fill = "white")
```




```
library(psych)
describe(cps4_small$educ)

##   var    n mean   sd median trimmed  mad min max range  skew kurtosis   se
## 1   1 1000 13.8 2.71     13   13.68 1.48   0  21    21 -0.07     2.06 0.09
```

The peak at 12 years of education indicate that higher number of high school graduate. There are few observations which are less than 12 which represents high school dropout. The spike at 16 years describes the students with a four year college degree. The middle two values indicate two years of community college level degree.The rest of the histotrogram (above 16) indicate Masters and PhD holders.The descriptive statistics also give median, standard deviation, range, kutosis and skewness of the wage.

(b)

```
lm1 <- lm(wage ~ educ, data = cps4_small)
summary(lm1)

##
## Call:
## lm(formula = wage ~ educ, data = cps4_small)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28.63  -7.82  -2.62   5.02  55.38
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.710      1.914   -3.51  0.00048 ***
## educ           1.980      0.136   14.55  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.7 on 998 degrees of freedom
## Multiple R-squared:  0.175,Adjusted R-squared:  0.174
## F-statistic:  212 on 1 and 998 DF,  p-value: <2e-16
```

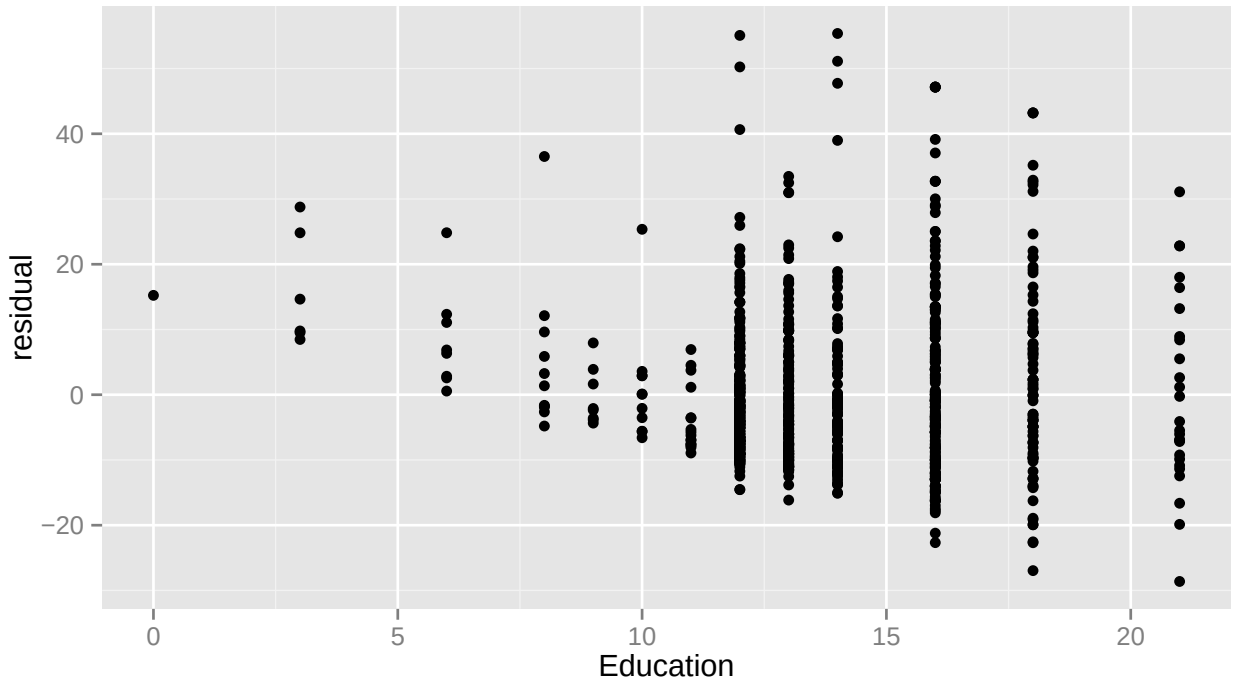
The estimated model:

WAGE= -6.7103+ 1.9803EDUC

The coefficient 1.9803 indicates the estimated increase in the expected hourly wage rate per year of education. The negative value of intercept represents the estimated wage rate of a worker with zero years of education. This value is meaningless as it is not possible to have a negative hourly wage rate.

(c)

```
residual = resid(lm1)
qplot(cps4_small$educ, residual, xlab = "Education")
```



The residuals are plotted against education above. There is a pattern evident; as EDUC increases, the magnitude of the residuals also increases. It implies that the error variance is greater for larger values of EDUC.

Violation of assumption SR3 is evident here. We know that if the assumptions (SR1-SR5) hold, no particular pattern should be evident in the residuals.

(d)

```
male <- cps4_small[cps4_small[, "female"] < 1, ]
head(male, n = 3)
```

```
##    wage educ female midwest south west black asian
## 2 11.50   12      0       1     0    0     0     0
## 3 15.04   16      0       0     0    1     1     0
## 5 24.03   12      0       0     0    0     0     0
```

```
female <- cps4_small[cps4_small[, "female"] > 0, ]
head(female, n = 3)
```

```
##     wage educ female midwest south west black asian
## 1  18.70   16      1       0     1    0     0     0
## 4  25.95   14      1       0     1    0     1     0
## 10 16.83   13      1       0     0    0     0     0
```

```
white1 <- cps4_small[cps4_small[, "black"] < 1, ]
white <- white1[white1[, "asian"] < 1, ]
head(white, n = 3)
```

```
##    wage educ female midwest south west black asian
## 1 18.70   16      1       0     1    0     0     0
## 2 11.50   12      0       1     0    0     0     0
## 5 24.03   12      0       0     0    0     0     0
```

```
black1 <- cps4_small[cps4_small[, "black"] > 0, ]
black <- black1[black1[, "asian"] < 1, ]
head(black, n = 3)
```

```
##     wage educ female midwest south west black asian
## 3  15.04   16      0       0     0    1     1     0
## 4  25.95   14      1       0     1    0     1     0
## 14 14.00   14      1       0     1    0     1     0
```

```
lm2 <- lm(male$wage ~ male$educ)
summary(lm2)
```

```
##
## Call:
## lm(formula = male$wage ~ male$educ)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24.65  -7.56  -2.50   5.55  47.85
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.054      2.494   -1.22     0.22
## male$educ      1.875      0.181   10.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.5 on 484 degrees of freedom
## Multiple R-squared:  0.181,Adjusted R-squared:  0.179
## F-statistic:  107 on 1 and 484 DF,  p-value: <2e-16
```

```r
lm3 <- lm(female$wage ~ female$educ)
summary(lm3)
```

```
##
## Call:
## lm(formula = female$wage ~ female$educ)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29.09  -6.71  -2.92   4.13  58.01
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -14.168      2.896   -4.89  1.3e-06 ***
## female$educ    2.358      0.202   11.69  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.4 on 512 degrees of freedom
## Multiple R-squared:  0.211,Adjusted R-squared:  0.209
## F-statistic:  137 on 1 and 512 DF,  p-value: <2e-16
```

```r
lm4 <- lm(white$wage ~ white$educ)
summary(lm4)
```

```
##
## Call:
## lm(formula = white$wage ~ white$educ)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27.33  -7.94  -2.35   5.15  55.05
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.551      2.076   -3.15   0.0017 **
## white$educ     1.992      0.148   13.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.7 on 843 degrees of freedom
## Multiple R-squared:  0.177,Adjusted R-squared:  0.176
## F-statistic:  181 on 1 and 843 DF,  p-value: <2e-16
```

```r
lm5 <- lm(black$wage ~ black$educ)
summary(lm5)
```

```
##
## Call:
## lm(formula = black$wage ~ black$educ)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16.10  -6.30  -3.56   1.76  48.03
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -15.086      6.169   -2.45    0.016 *
## black$educ     2.449      0.453    5.40  3.8e-07 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11 on 110 degrees of freedom
## Multiple R-squared:  0.21,Adjusted R-squared:  0.203
## F-statistic: 29.2 on 1 and 110 DF,  p-value: 3.8e-07
```

The estimated model for Male: **WAGE= -3.0545+ 1.8753EDUC**
The estimated model for Female: **WAGE= -14.1680+ 2.3575EDUC**
The estimated model for Black: **WAGE= -15.0859+ 2.4491EDUC**
The estimated model for White: **WAGE= -6.5507+ 1.9919EDUC**

In filtering white and black popular, we filtered out those who are asians. From the esimated models, it's visible that an extra year of education increases the wage rate for a black worker than for a white worker. Similar trend is visible in female workers case. Their wage rate increases with an extra year of education more than male worker.

(e)

```
lmfit <- lm(wage ~ I(educ^2), data = cps4_small)
summary(lmfit)

##
## Call:
## lm(formula = wage ~ I(educ^2), data = cps4_small)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.24  -7.67  -2.44   4.98  55.90
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.08283    1.02316    5.95  3.8e-09 ***
## I(educ^2)    0.07349    0.00483   15.21  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.6 on 998 degrees of freedom
## Multiple R-squared:  0.188,Adjusted R-squared:  0.187
## F-statistic:  231 on 1 and 998 DF,  p-value: <2e-16
```

The estimated quadratic model:

WAGE= 6.082831+ 0.073489EDUC^2

Marginal effect:

SLOPE= 2(0.073489)EDUC

A person with 12 years of education has the estimated marginal effect of an additional year of education on expected wage is:

SLOPE= 2*0.073489*12= 1.7637

A person with 14 years of education has the estimated marginal effect of an additional year of education on expected wage is:

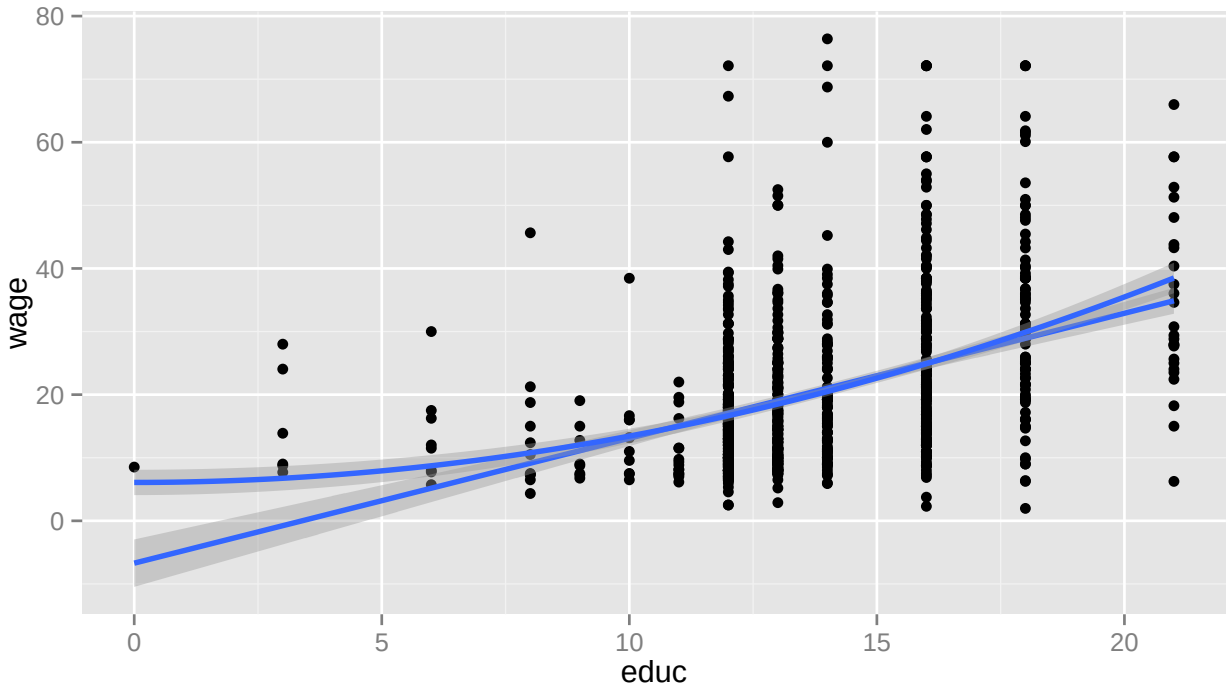
SLOPE= 2*0.073489*14= 2.0577

The linear estimated model indicated that an additional year of education is expected to increase wage by dollar 1.98 regardless of the number of years of education attained. So, the rate of change is pretty much

constant. On the other hand, the quadratic model implies that the effect of an additional year of education on wage increases with the level of education.

(f)

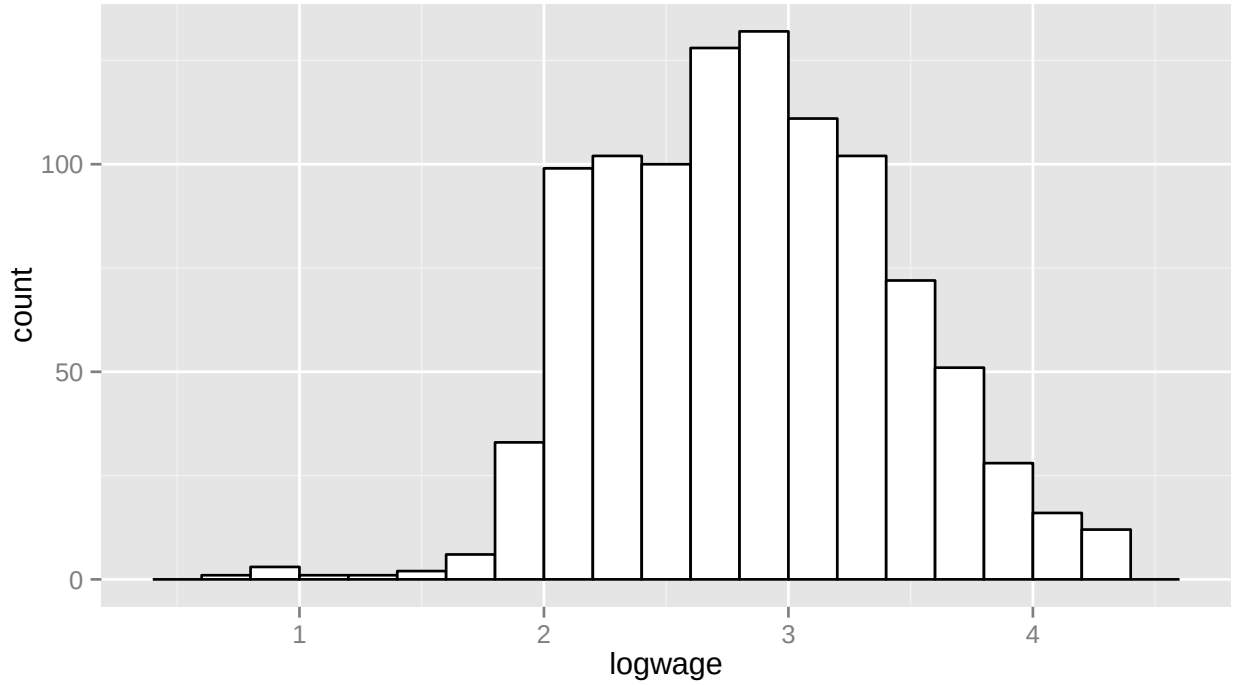
```
p <- qplot(educ, wage, data = cps4_small)
p + stat_smooth(method = "lm", formula = y ~ x, size = 1) + stat_smooth(method = "lm",
    formula = y ~ I(x^2), size = 1)
```



The quadratic model seems to fit the data better than the linear model equation.

(g)

```
m <- log(cps4_small$wage)
cps4_small$logwage <- log(cps4_small$wage)
head(cps4_small)

##    wage educ female midwest south west black asian logwage
## 1 18.70   16      1       0     1    0     0     0   2.929
## 2 11.50   12      0       1     0    0     0     0   2.442
## 3 15.04   16      0       0     0    1     1     0   2.711
## 4 25.95   14      1       0     1    0     1     0   3.256
## 5 24.03   12      0       0     0    0     0     0   3.179
## 6 20.00   12      0       0     0    0     0     0   2.996

ggplot(cps4_small, aes(x = logwage)) + geom_histogram(binwidth = 0.2, colour = "black",
    fill = "white")
```



From the above plot, the histogram of log(WAGE) is more symmetrical and bell-shaped than the histogram of WAGE.

(h)

```
lm6 <- lm(logwage ~ educ, data = cps4_small)
summary(lm6)

##
## Call:
## lm(formula = logwage ~ educ, data = cps4_small)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.559 -0.392  0.007  0.361  1.584
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.60944    0.08642    18.6   <2e-16 ***
## educ         0.09041    0.00615    14.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.527 on 998 degrees of freedom
## Multiple R-squared:  0.178,Adjusted R-squared:  0.177
## F-statistic:  216 on 1 and 998 DF,  p-value: <2e-16
```

The estimated model:

log(WAGE)= 1.609444+ 0.090408EDUC

Estimated marginal effect of education on WAGE is:

dWAGE/dEDUC=const* WAGE

For workers with 12 years of education the predicted wage rate:

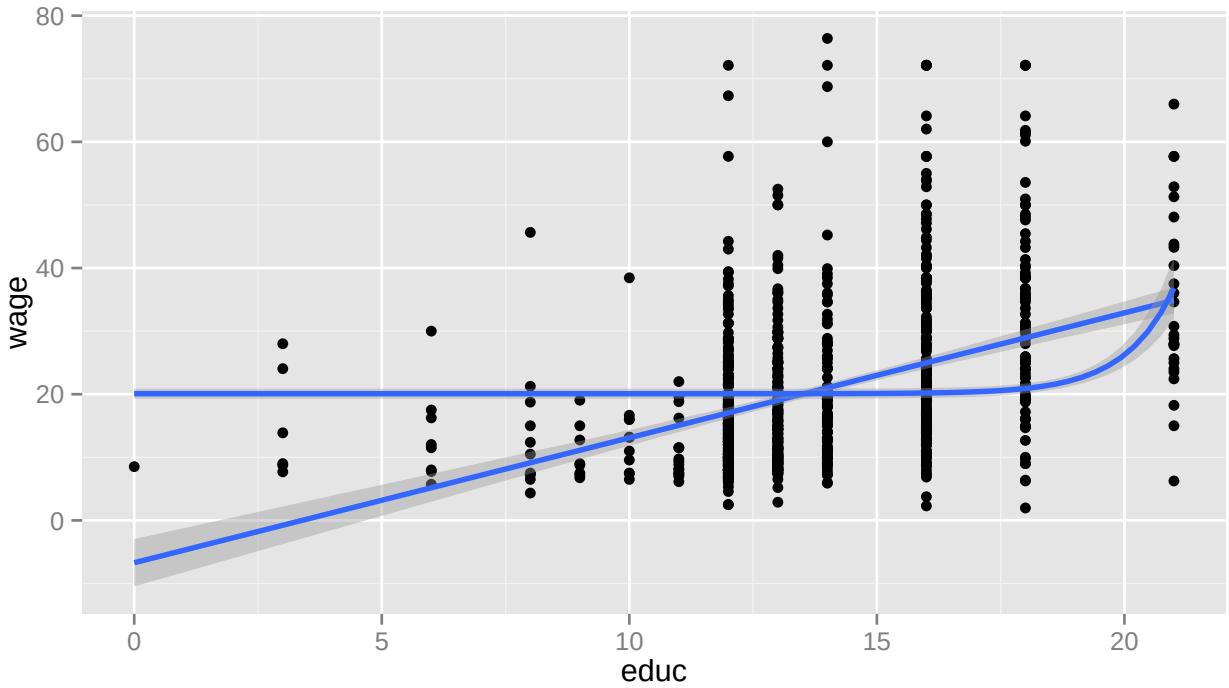
$\text{WAGE}_{12} = exp(1.60944 + 0.090408 * 12) = 14.796$

For workers with 14 years of education the predicted wage rate:

$\text{WAGE}_{14} = exp(1.60944 + 0.090408 * 14) = 17.728$

The marginal effects at these values are 1.34 and 1.60 respectively.

```
p <- qplot(educ, wage, data = cps4_small)
p + stat_smooth(method = "lm", formula = y ~ x, size = 1) + stat_smooth(method = "lm",
    formula = y ~ exp(x), size = 1)
```



For the linear model, effect of education was estimated to be dollar 1.98. For the quadratic model, the corresponding marginal effect estimates are dollar 1.76 and dollar 2.06 respectively. The marginal effects of the log-linear model are lower.