Econ 103L - Week 1

Ryan Martin

January 8, 2018

Brief Intro to R and Rmarkdown

Free (and very good!) Books

- A very short intro to R https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf
- A not-so-short intro to R: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- Hadley Wickham and Garrett Grolemund's **R for Data Science**: http://r4ds.had.co.nz/ (very useful!)
- **Advanced R**: http://adv-r.had.co.nz/ (whole book and code downloadable!)

These books are all you really need to get start, but many other useful and free-cheap books are available! For example, I found no starch press's **The ARt of R Programming** helpful, and it is selling for less than $20 on Amazon: https://www.amazon.com/Art-Programming-Statistical-Software-Design/dp/1593273843/ref=sr_1_1?srs=6327961011&ie=UTF8&qid=1515469349&sr=8-1&keywords=R. Many, many other books have been written.

Should I Learn R?

Maybe. It will take some investment to start, but the investment will be well worth it if you plan on doing statistics or working for a tech firm or a firm that employs highly-qualified researchers or data science.

Note, the professor allows you to use any language you want for the work. However, the exams will likely ask you to interpret stata code. While both R and Stata produce very similar output, it would be safest to make sure

If I were you, I would try to do the homework in R (or python if you are a CS student), and then compare the answers with the professor's Stata output answers. Save all the code in a well-organized file. That way you learn both. If you refer back to one thing from your undergraduate career, it will almost certainly be these code solutions you wrote.

What you need to get up and running with R

- R can be installed (for free) from here: https://www.r-project.org/
- Most R coders use Rstudio as their (interactive) development environment (IDE). It's also free for individuals: https://www.rstudio.com/
- You will want to use Rmarkdown to write code. To do so, execute the following code in an R console (it's that easy)!

```r
install.packages("rmarkdown")
library(rmarkdown)
```

- Now you will have the option to create File -> New File -> R markdown… This let's you section off code chunks and markdown text (Markdown is the language of computer science documents. For markdown intro see second page and left hand side of here: https://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf)

- To insert a code block, hit ctrl + alt + i in R studio (for windows) or type three forward ticks then brackets around an r, then skip a few lines and 3 forward ticks again.
- For help, type about `blah`, type `?blah`. If you can't remember `blah` exactly but think it was something like `bla` or `blar`, type `??bla` or `??blar` into the console
- Rmarkdown documents can have code chunks in other languages too (Rcpp, Python, C, C++, more). However, only R and Rcpp code can be used interactively in the Rstudio IDE (so far). If you are interested in developing in these other languages, there are many other great IDEs. Or you can use a fancy text editor with packages like Sublime Text (https://www.sublimetext.com/) or Atom (https://atom.io).
- Note that R is the language of choice for all statisticians and is under constant expansion. Statisticians write packages themselves to expand R's capabilities. The statisticians are rewarded for new code with journal publications in JSS (Journal of Statistical Software. (Journal Publications are the currency of academia.)

Important Advice

1. Always comment your code. This is just being a good citizen. Whatever language you are writing in. If you get into data science, you may spend a lot more time reading code than writing it. It is much easier when it is properly commented. Moreover, you will often want to refer back to your code! After a year, you will not remember what you were doing easily without clear comments and a good code structure. There are plenty of style guides that you can check out to help you write cleaner code. One famous one is: https://www.amazon.com/Pragmatic-Programmer-Journeyman-Master/dp/020161622X.
2. If you don't know how to already, learn to touch type. It is a very useful skill that no one will sit down and teach you if you haven't already learned. Just find an internet typing tutor and grind through yourself. Treat it like a game.
3. Learn how to use the shortcut keys and autocomplete of your favorite programs. You would be surprised how much you can do with, for example, Google Chrome or in Gmail without ever touching the mouse.
4. Organize your code so you will be able to use it (perhaps years) later. Your job may (will) ask if you can code in R or Stata and it is important to be able to review this work quickly so you don't look foolish on your first day or during the interview!
5. If you have coding experience, take careful note that R indexes all start at 1 rather than 0! A major source of bugs when going between R and python, for example.

Comparison of R and Stata

Item	R	Stata
Can I see the internal code?	Yes! Open Source	No
Price?	Free	Very Expensive
Ease of Use	Difficult at first then Easy	Easy then Limited
When are new features available?	When Statisticians, Econometricians or whoever write new packages	When the business decides its profitable to add new features from papers written by statisticians or econometricians
Who can help me?	Vibrant open-source community, stack exchange, cheap and well written books	Paid technician, thick instruction manuals

Item	R	Stata
Who uses it?	Data Scientists, all statisticians, economists at tech companies, more	Economists at government, low-tech firms, more

Discussion Problems

Aside, Models and Interpretation.

In this class, there will be several models that will pop up repeatedly.

1. Linear Model:
$$y = \beta_1 + \beta_2 x + e$$
Predictive form: $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$ and $\frac{d\hat{y}}{dx} = \hat{\beta}_2$. So, here $\hat{\beta}_2$ is exactly the slope.
2. log-linear Model:
$$\log y = \beta_1 + \beta_2 x + e$$
Predictive Form: $\log \hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$. Note here that $\frac{\%\Delta \hat{y}}{\Delta x} \approx 100 \times \frac{d\hat{y}}{dx}\frac{1}{\hat{y}} = 100 \times \hat{\beta}_2$ and thus $100 \times \hat{\beta}_2$ can be interpreted as the percent change in predicted y for a 1 unit change in x (or $\hat{\beta}_2$ can be interpreted as the fraction change in y for a 1 unit change in x). In this case, $\hat{\beta}_2$ is called the semi-elasticity.
3. log transformed Model:
$$y = \beta_1 + \beta_2 \log x + e$$
Predictive Form: $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \log x$. Note here that $\frac{d\hat{y}}{dx} = \hat{\beta}_2 / x$. Thus $\hat{\beta}_2 = x\frac{d\hat{y}}{dx} \approx \frac{\Delta \hat{y}}{\frac{1}{100}\%\Delta x} = 100 \times \frac{\Delta \hat{y}}{\%\Delta x}$. That is, $\frac{1}{100}\hat{\beta}_2$ can be interpreted as the change in predicted y for a 1 percent change in x.
4. log-log:
$$\log y = \beta_1 + \beta_2 \log x + e$$
Predictive Form: $\log \hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \log x$. Thus, $\frac{dy}{dx}\frac{1}{\hat{y}} = \hat{\beta}_2 \frac{1}{x}$ which implies $\hat{\beta}_2 = \frac{d\hat{y}}{dx}\frac{x}{y} \approx \frac{100}{100}\frac{\%\Delta y}{\%\Delta x} = \frac{\%\Delta y}{\%\Delta x}$ is (interpreted as) the elasticity of the predicted y with respect to the x variable. That is, it is the expected percent change in y for a 1 percent change in x.

2.6

A soda vendor at Louisiana State University football games observes that more sodas are sold the warmer the temperature at game time is. Based on 32 home games covering five years, the vendor estimates the relationship between soda sales and temperature to be

$$\hat{y} = -240 + 8x$$

where y is the number of sodas she sells and x is the temperature in degrees Fahrenheit

a

Interpret the estimated slope and intercept. Do the estimates make sense? Why or why not?

Solution Since it's a line, the slope is just 8. (It is worth remembering for later that this could be done with calculus as well, especially in more complicated - nonlinear - relationships.) This says that for every single degree fahrenheit increase in temperature, our best estimate is that the vendor sells 8 more sodas. This has soda sales increasing with temperature, which is what was described above. The number does not seem crazy, although I have little experience with soda sales to know how reasonable it is.

b

On a day when the temperature at game time is forecast to be 80° F, predict how many sodas the vender will sell.

Solution:

```
-240 + 8*80
```

```
## [1] 400
```

c

Below what temperature are the predicted sales zero?

Solution: 30, because

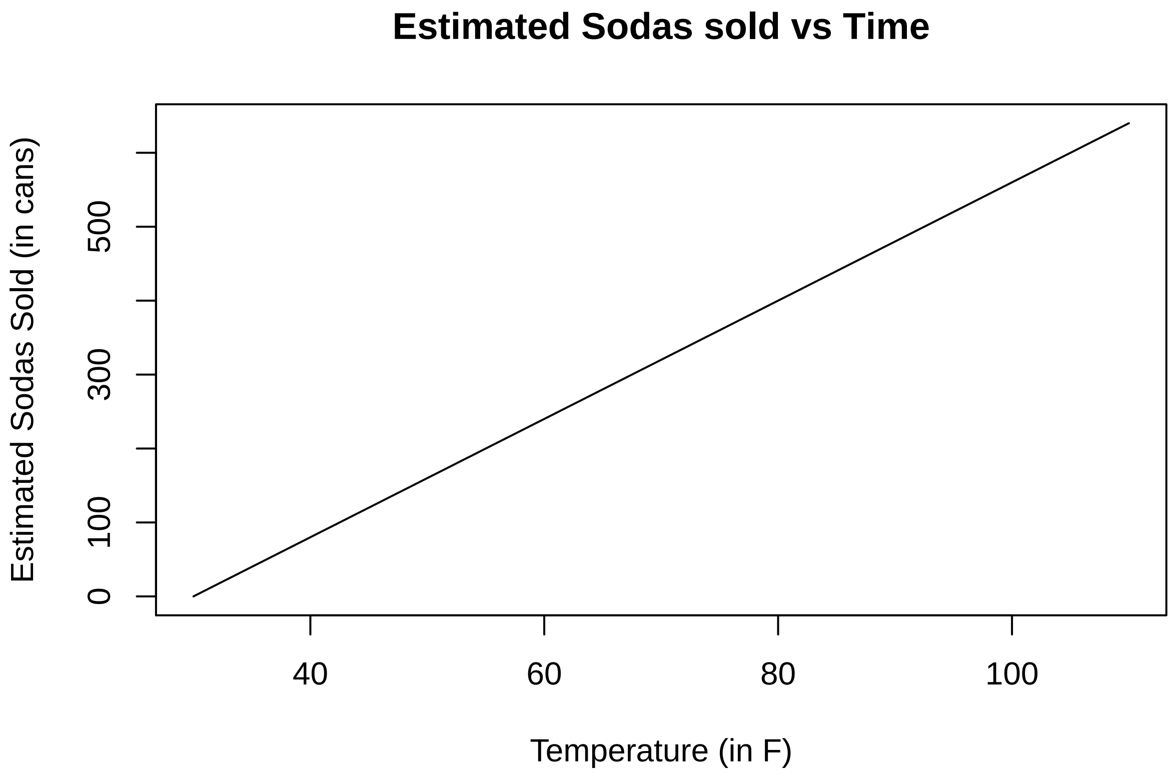
$$0 = -240 + 8 \times \text{zero\_sale\_temp}$$

and solve

d

Sketch a graph of the estimated regression line

```
x <-  30:110
y <-  -240 + 8*x
plot(x,y, xlab = "Temperature (in F)", ylab = "Estimated Sodas Sold (in cans)",
     main = "Estimated Sodas sold vs Time",type = 'l')
```



Note that very few of the actual data points (maybe none!) will fall on the line for most real-world datasets!

Stata Solution

Quick note on running Stata

Code graciously provided by Conor Foley. All mistakes are my own.

Note, without downloading (and paying for) Stata or using a school computer, you may be able to run it remotely from here:

https://software.library.ucla.edu/Citrix/SoftwareWeb/

You have to play with it a little, but you can upload data files to the remote server and move code on your clipboard (i.e. that you have copied/ctrl +c) into the remote server. Use the navigation buttons on the top.

There are other options, but this one seems to be the easiest. It takes some effort to explore UCLA's computing options, since different departments offer different resources. If you find a better solution, let me know.

Solution

Note that stata is a much more rigid language than R. This contributes to its ease of use (just memorize 1 pattern), but also restricts what can be done with it.

```
log using wk1_section_log, replace

//Demonstration STATA code for week 1
//Principles of Econometrics 4th Edition

/////////////////////////// Question 2.6 \\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\

//Part A (no STATA content)

//Part B
// use display (disp for short) and artimetic symbols to use STATA like a calculator
// will not work without!
disp -240 + 8*80

//Pact C
disp 240/8

//Part D
//Manually create a dataset using the formula in the problem
//Want to set up an x variable with values 30:110 (counting by 1)
clear all
scalar setobsnum = 110-30+1 // +1 because 110-30 doesn't count 30 itself
//to pass the scalar setobsnum to the command set obs we need to use this funny
// `=setobsnum' syntax. The issue is that STATA is expecting an integer value,
// and will not automatically interpret the name "setobsnum" to its numeric
// value, so we use `=setobsnum' to evaluate the scalar name back to a number
// and create a valid input for set obs
set obs `=setobsnum'
gen x = . // creates a column of empty values
replace x = _n + 29 // _n is the obseration number (1 to 81) and +29 shifts that
                    // this vector up to 30 to 110
```

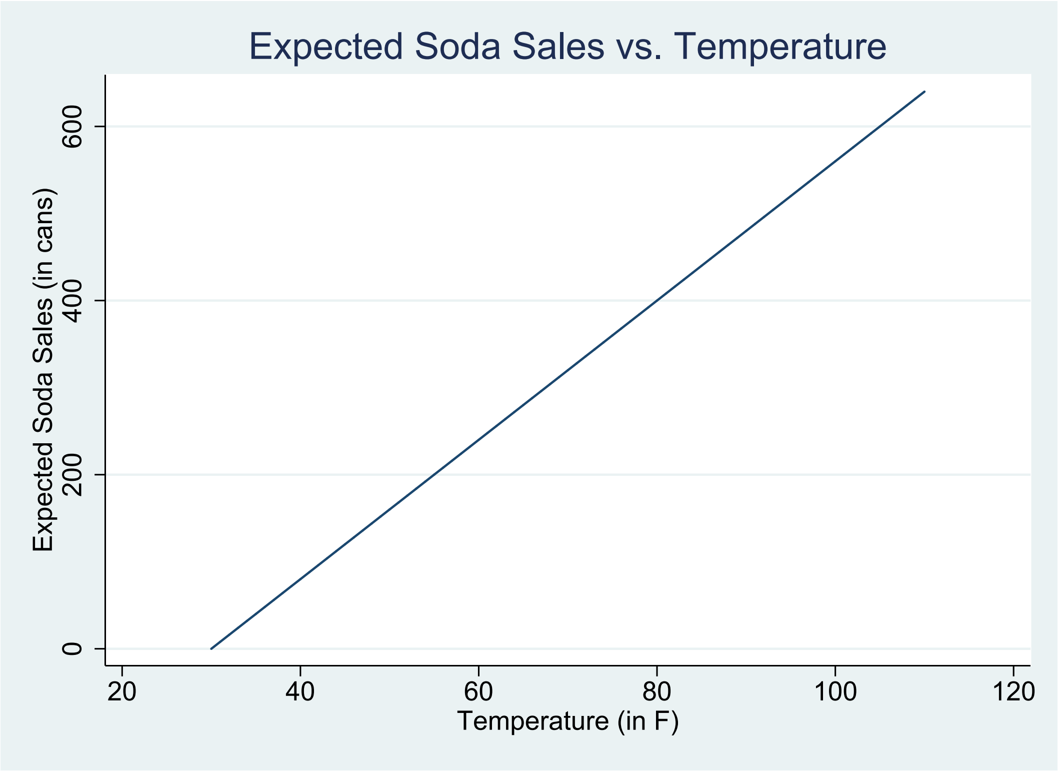


Figure 1: Stata graph output

```
gen y = -240 + 8*x // generates y data using the equation from the class

label variable x "Temperature (in F)"
label variable y "Expected Soda Sales (in cans)"
twoway line y x, title("Expected Soda Sales vs. Temperature")
graph export "Question 2-6 Figure.pdf", replace
```

2.15

R Solution

How much does education affect wage rates? The data file `cps4_small.dat` contains 1000 observations on hourly wage rates, education and other variables from the 2008 Current Population Survey (CPS).

a

Obtain the summary statistics and histograms for the variables $WAGE$ and $EDUC$. Discuss the data characteristics.

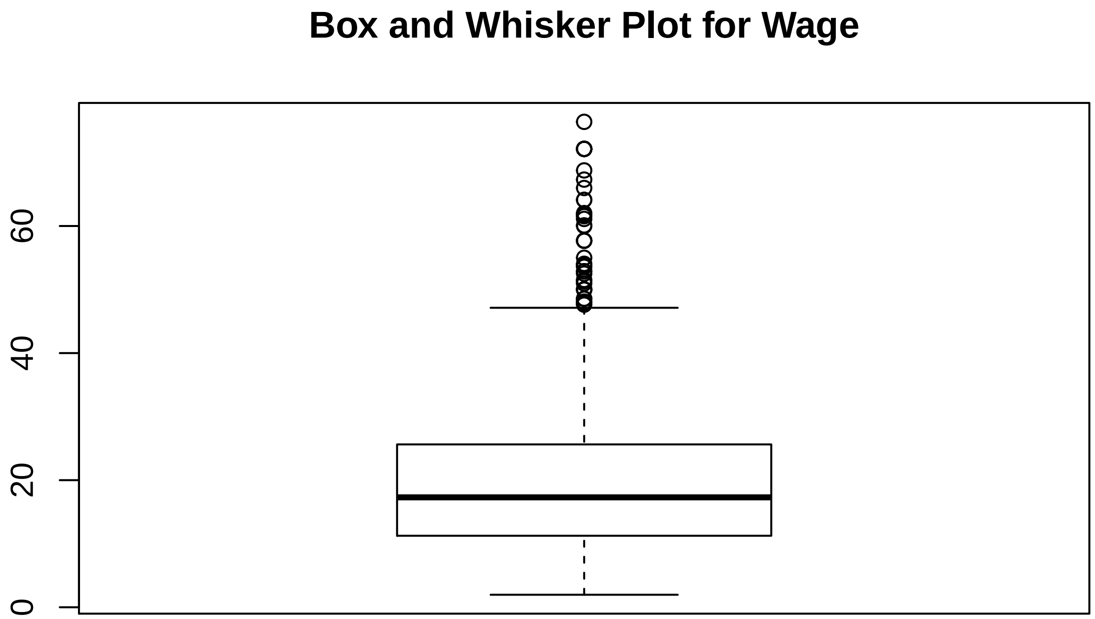
Solution

Note that R can easily read in Stata files (`.dta`) with the right package (e.g. `haven`). Note that the # sign comments out a line.

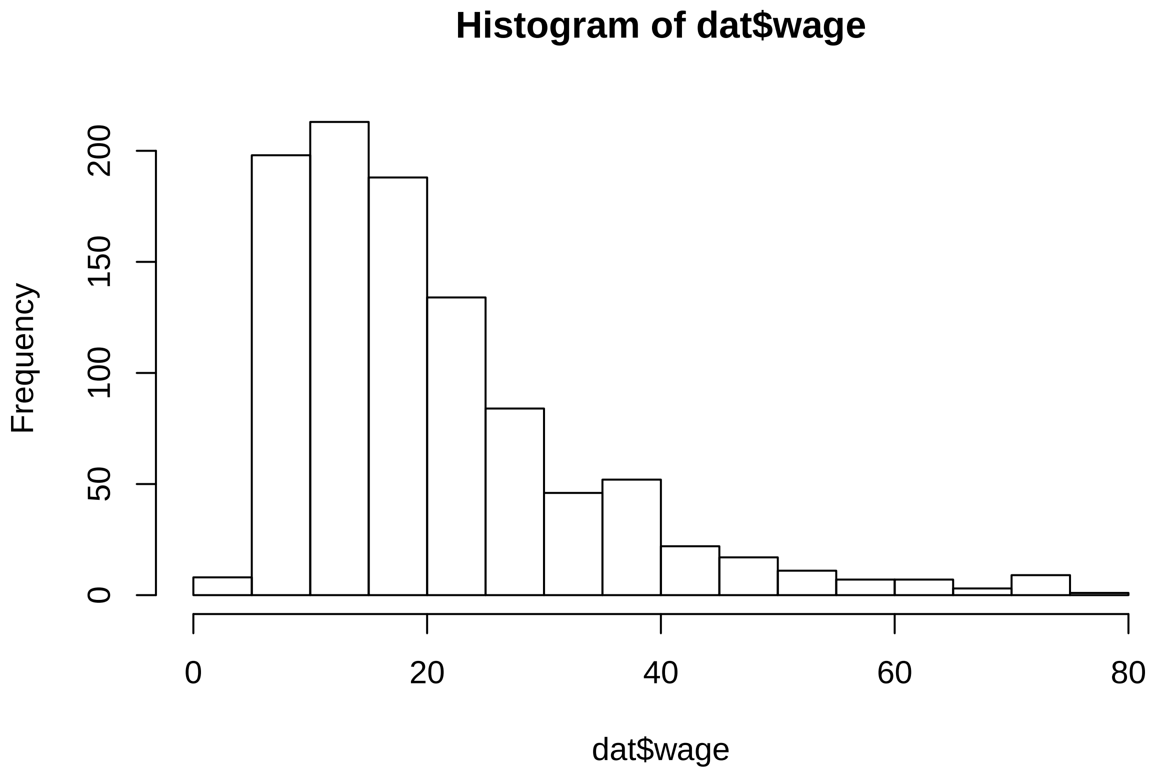
```
#You should set it to your own working directory
#Note that the \ slashes are the default, but you have to change these to
#either double backslashes or / as I did
# file location from computer: C:\Users\ryanj\Dropbox\TA\Econ 103\Winter 2018\Data
my_wd <- "C:/Users/ryanj/Dropbox/TA/Econ 103/Winter 2018/Data/s4poe_statadata"
my_file <- paste(my_wd, "cps4_small.dta", sep = "/")

#to read in stata files, need special package (many exist)
#install.packages('haven') #must run if haven't installed it yet
library(haven)
dat <- read_stata(my_file)
    #could also use
    #haven::read_stata(my_file)
    #for questions, type ?read_stata or ?haven::read_stata

#Take a look
View(dat)



##################################################
#First looking at wage
##################################################

#refer to items with dollar sign
#these autocomplete, type the dollar sign then hit tab and choose from the
#dropdown menu. much easier than remembering the names
#these are the built in plots
summary(dat$wage) #5 number summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.97   11.25   17.30   20.62   25.63   76.39
```

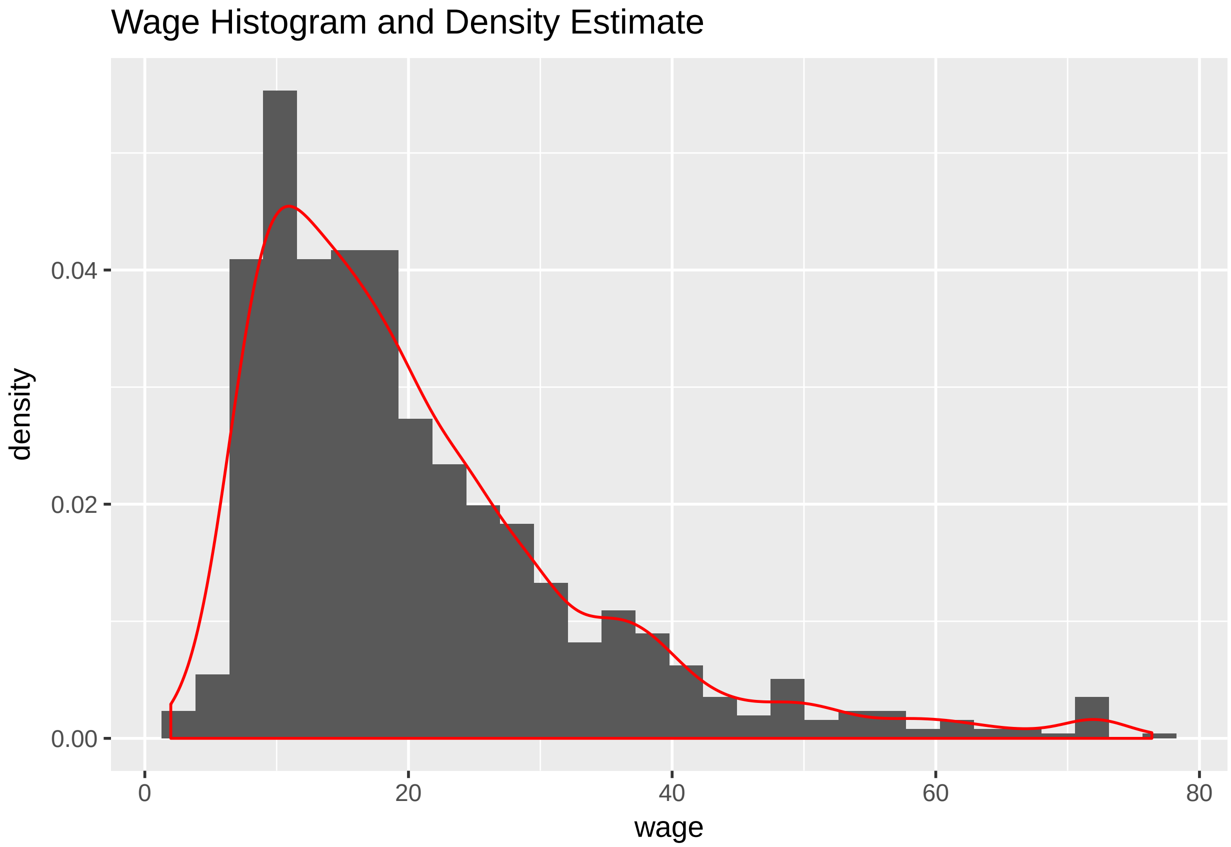
```
boxplot(dat$wage, main = "Box and Whisker Plot for Wage")
```



```
hist(dat$wage)

#Note, ggplot2 has some fancier plots
#install.packages("ggplot2")
library(ggplot2)
```



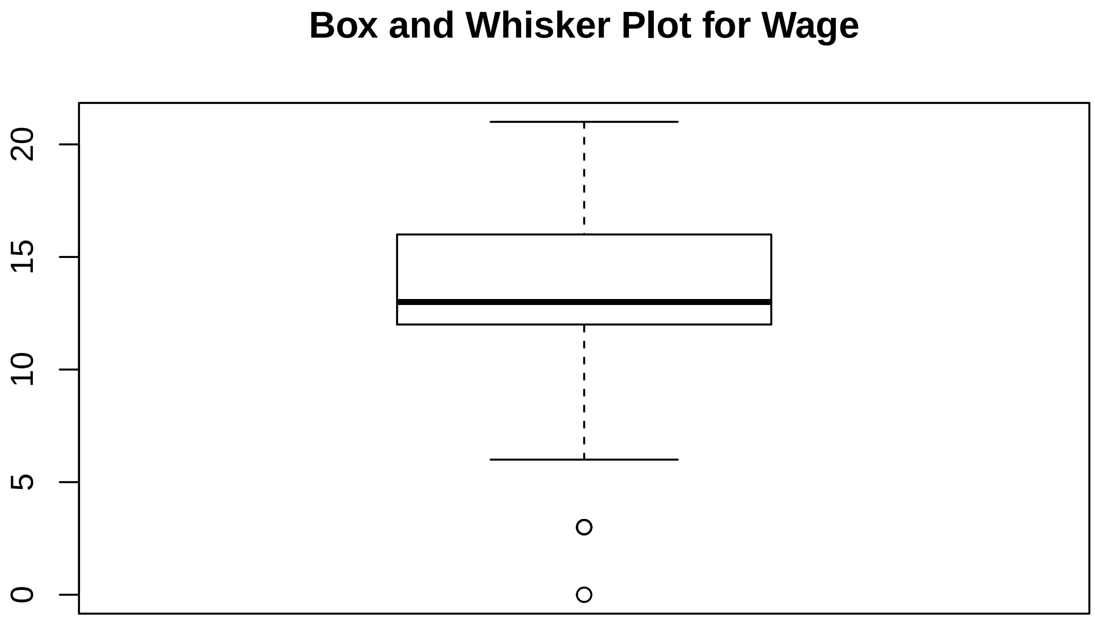
```r
ggplot(dat = dat, aes(wage, ..density..)) +
  geom_histogram() +
  geom_density(col = 'red') +
  ggtitle("Wage Histogram and Density Estimate")
```



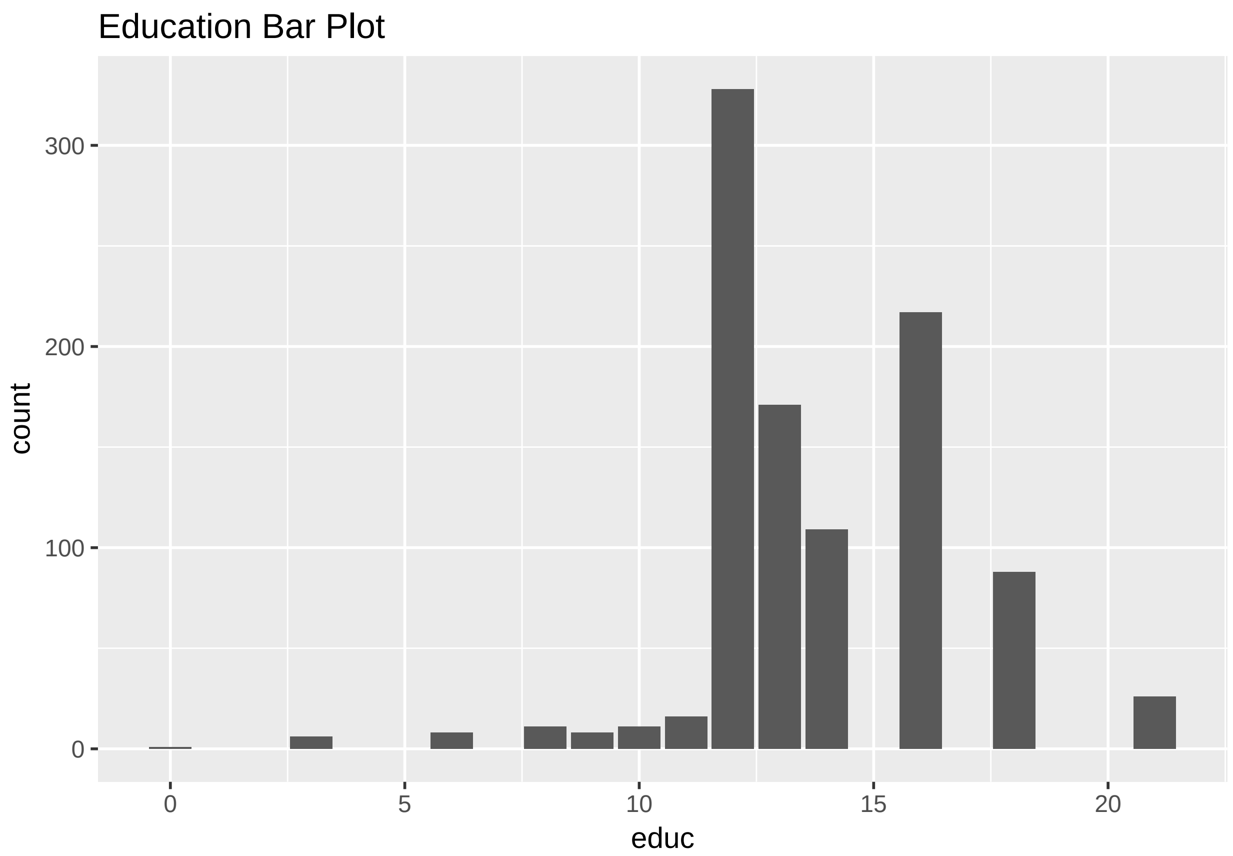
```
##########################################################
#Education
#########################################################



summary(dat$educ) #5 number summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    12.0    13.0    13.8    16.0    21.0
```

```
boxplot(dat$educ, main = "Box and Whisker Plot for Wage")
```



```
## Education is discrete, so we go with a barplot
ggplot(dat = dat, aes(educ)) +
  geom_bar() +
  ggtitle("Education Bar Plot")
```



Discusion Wages seem to have a long right tail (skew left). Education has two peeks at 12 (high school graduate) and 16 (college graduate). The outliers for education are in the lower education levels.

b

Estimate the linear regression $WAGE = \beta_1 + \beta_2 EDUC + e$ and discuss the results

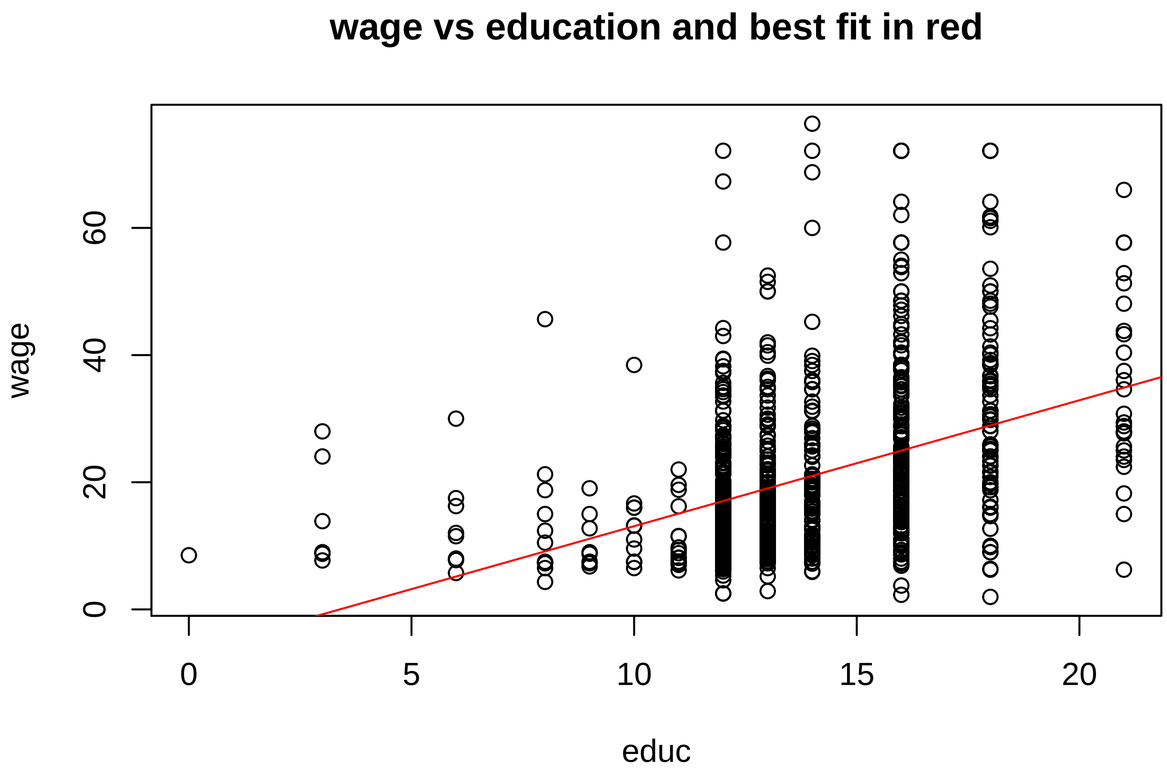
Solution

```r
attach(dat)
reg_out <- lm(wage ~ educ + 1)
summary(reg_out)
```

```
##
## Call:
## lm(formula = wage ~ educ + 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.626  -7.816  -2.623   5.019  55.376
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.7103     1.9142  -3.506 0.000476 ***
## educ          1.9803     0.1361  14.548  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 11.66 on 998 degrees of freedom
## Multiple R-squared:  0.175,  Adjusted R-squared:  0.1741
## F-statistic: 211.7 on 1 and 998 DF,  p-value: < 2.2e-16
```

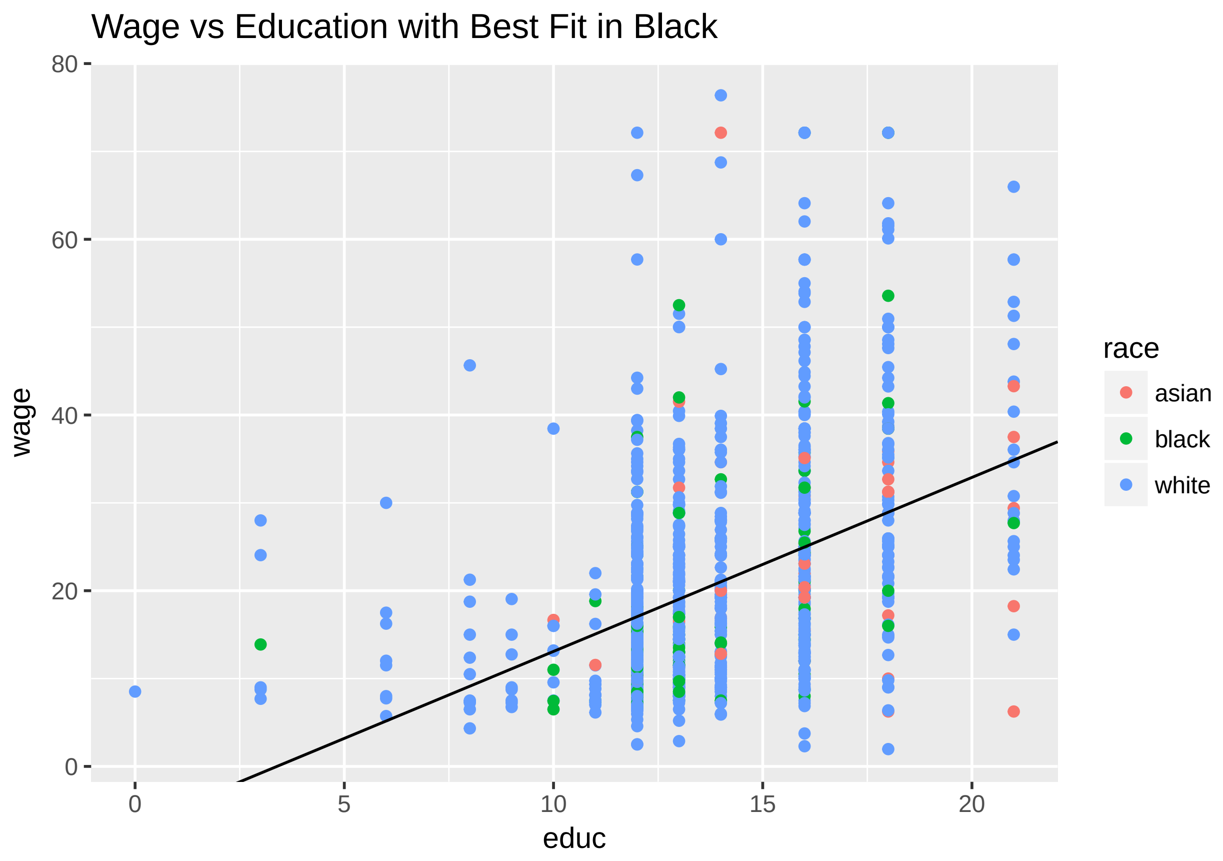
```
#Plot using base functions
plot(educ, wage, main = "wage vs education and best fit in red")
abline(reg_out, col = "red")
```



```
#make a little fancier with ggplot
race  <- rep("white", nrow(dat))
race[black == 1] <- "black"
race[asian == 1] <- "asian"

my_coef <- reg_out$coefficients
my_coef
```

```
## (Intercept)        educ
##   -6.710328    1.980288
```

```
dat2 <-  cbind(dat, race) #adds extra column to dat
ggplot(data = dat2, aes(x= educ, y = wage, color = race)) +
  geom_point() +
  geom_abline(aes(intercept = my_coef[1],
                  slope = my_coef[2] )) +
  ggtitle("Wage vs Education with Best Fit in Black")
```



The results say that the best *linear* predictor for wage in terms of race is

$$WAGE = -6.71 + 1.98 \times EDUC.$$

This predicts that 1 extra year of education gets an additional 1.98 dollars per hour in wages.

c

Calculate the least squares residuals and plot them against $EDUC$. Are any patterns evident? If assumptions $SR1 - SR5$ hold, should any patterns be evident in the least squares residuals?

Solution

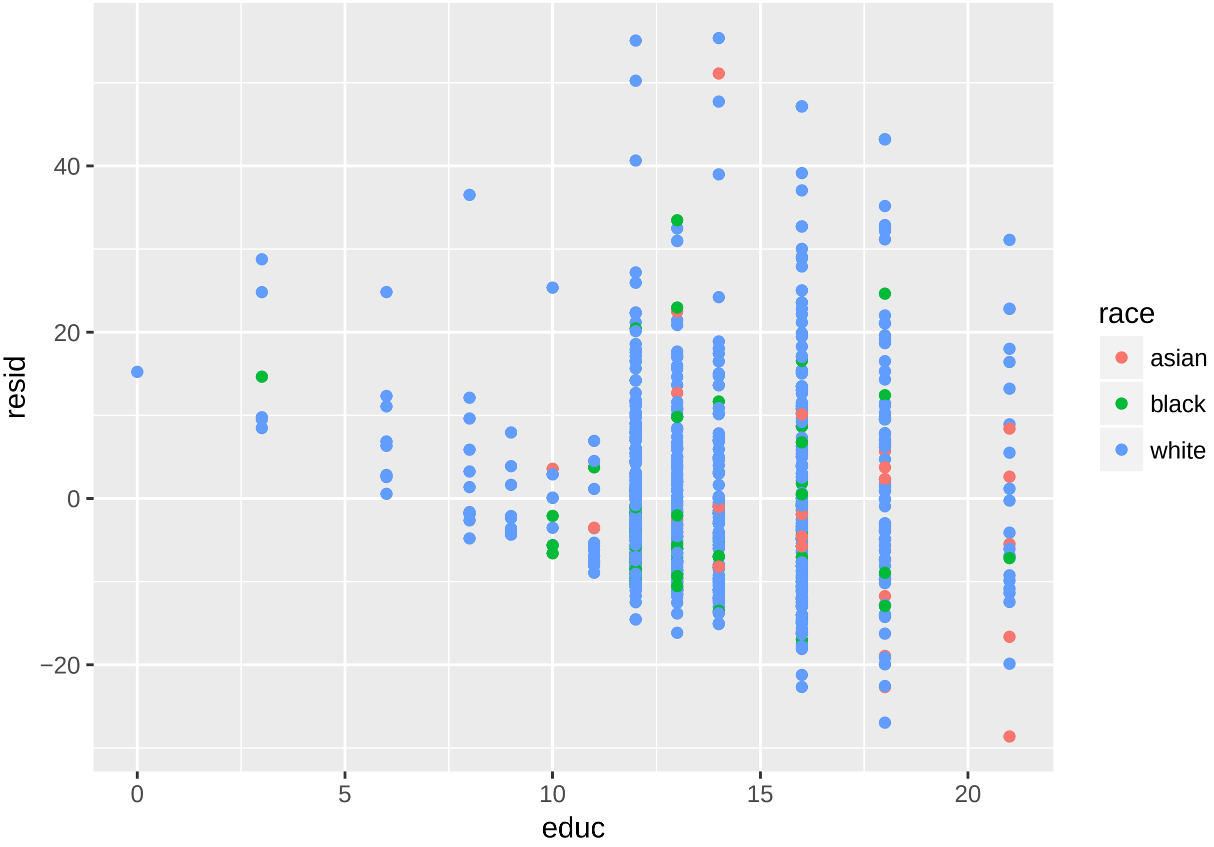
Recall the assumptions (treating dependent variables as given, a.k.a. conditioning on it):

- SR1: The linear model is correct
- SR2: The errors are mean 0
- SR3: The errors are homoskedastic
- SR4: the errors are uncorrelated
- SR5: dependent variable (e.g. x) is treated as given but is not a constant (i.e. takes at least two values)
- SR6: (optional) errors are normally distributed

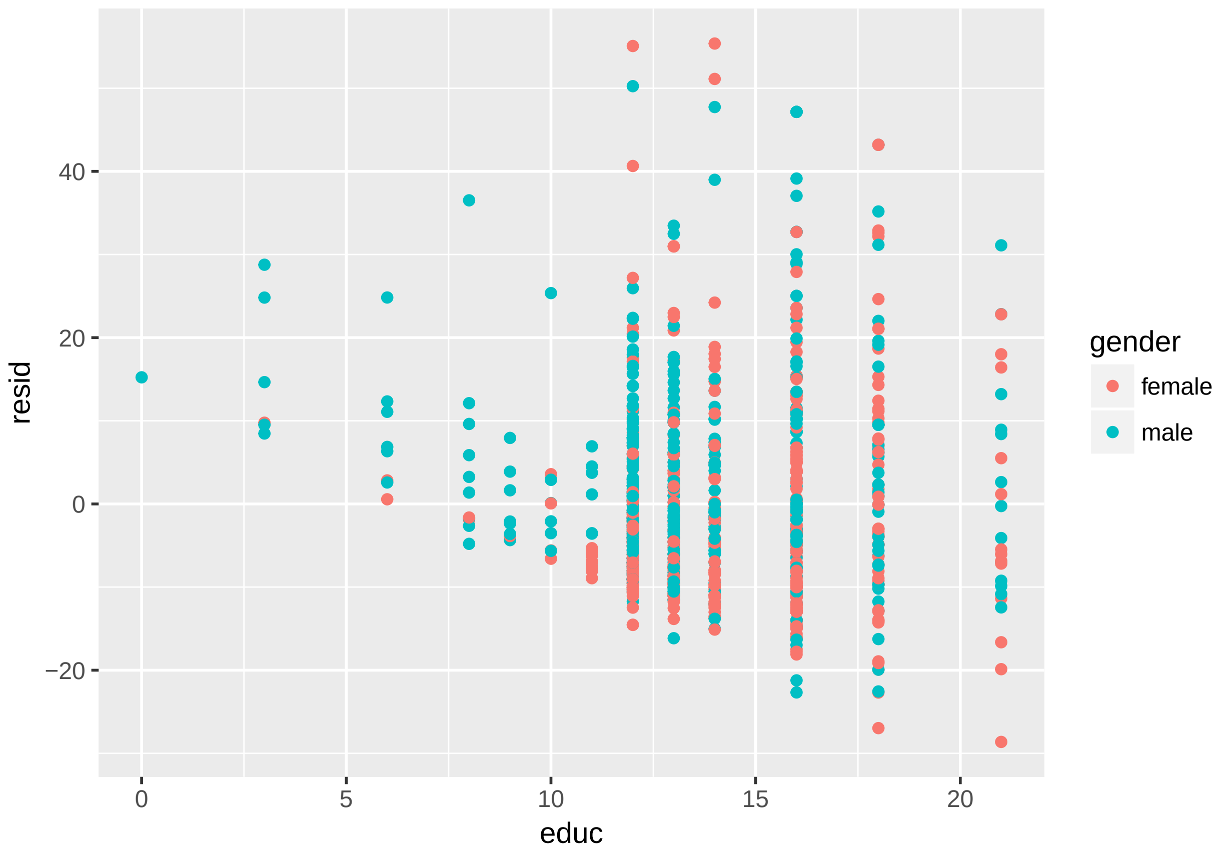
If the assumptions hold, there should be no change in the residual point spread as a function of education. (This is SR2 and SR3, in particular) It seems here that the residual variance does have some dependence on education. This suggests that our (linear) model may be the incorrect dependency.

```
resid <- reg_out$residuals
gender <- rep("male", nrow(dat))
gender[dat$female==1] <- "female"
```

```
dat3 <- cbind(dat2, resid,gender)
ggplot(data=dat3,aes(x = educ, y = resid, color = race) ) +
  geom_point()
```



```
#repeat by gender
ggplot(data=dat3,aes(x = educ, y = resid, color = gender) ) +
  geom_point()
```



d

Estimate separate regressions for males, females, blacks, and whites. Compare the results.

Solution There is a little ambiguity here. Is it just males, just females, just blacks and just whites. Or is it males and black, males and white, females and black, females and white? I will assume the first but I think both interpretations are reasonable

One point of comparison is the returns to education, which is the estimated slope. Note that males get lower returns to education than women and whites have lower returns to education than blacks. At the same time, we see that the intercept term, which corresponds to predicted wage with 0 education. Here, male is larger than female and white is larger than black.

Since males have higher initial wages but lower returns to education than females, do males or females have higher wages as high school graduates and as college graduates? The calculations are done below. Our regressions predict that males make more than females as high school graduates, with two years of college, and with 4 years of college. Similarly, our regressions predict whites make more than blacks at these four points in educational attainment as well.

Note We will learn how to do these later much more quickly and easily in one regression, with indicator variables and interaction effects.

```
dat_male <- dat3[dat$female!=1,]
dat_female <- dat3[dat$female==1,]
dat_black <- dat3[dat$black==1,]
dat_white <- dat3[dat3$race=="white",]
  #could also use (dat$black!=1) & (dat$asian!=1)
```

```
reg_male <- lm(data = dat_male, wage ~ educ)
reg_female <- lm(data = dat_female, wage ~ educ)
reg_white <- lm(data = dat_white, wage ~ educ)
reg_black <- lm(data = dat_black, wage ~ educ)
summary(reg_male)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = dat_male)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.650  -7.558  -2.499   5.551  47.851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.0545     2.4935  -1.225    0.221
## educ          1.8753     0.1814  10.336   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.55 on 484 degrees of freedom
## Multiple R-squared:  0.1808, Adjusted R-squared:  0.1791
## F-statistic: 106.8 on 1 and 484 DF,  p-value: < 2.2e-16
```

```
summary(reg_female)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = dat_female)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.090  -6.710  -2.918   4.132  58.008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.1680     2.8957  -4.893 1.33e-06 ***
## educ          2.3575     0.2017  11.690  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.35 on 512 degrees of freedom
## Multiple R-squared:  0.2107, Adjusted R-squared:  0.2091
## F-statistic: 136.7 on 1 and 512 DF,  p-value: < 2.2e-16
```

```
summary(reg_black)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = dat_black)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.099  -6.303  -3.557   1.760  48.031
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.0859     6.1693  -2.445   0.0161 *
## educ          2.4491     0.4531   5.405  3.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.02 on 110 degrees of freedom
## Multiple R-squared:  0.2098, Adjusted R-squared:  0.2027
## F-statistic: 29.21 on 1 and 110 DF,  p-value: 3.803e-07
```

```
summary(reg_white)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = dat_white)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.334  -7.944  -2.353   5.147  55.054
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.5507     2.0764  -3.155  0.00166 **
## educ          1.9919     0.1482  13.444  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.67 on 843 degrees of freedom
## Multiple R-squared:  0.1766, Adjusted R-squared:  0.1756
## F-statistic: 180.8 on 1 and 843 DF,  p-value: < 2.2e-16
```

```
new.dat <- data.frame(educ = c(12,14,16))
predict.lm(reg_male, newdata =new.dat)
```

```
##        1        2        3
## 19.44912 23.19972 26.95032
```

```
predict.lm(reg_female, newdata =new.dat)
```

```
##        1        2        3
## 14.12218 18.83722 23.55226
```

```
predict.lm(reg_white, newdata =new.dat)
```

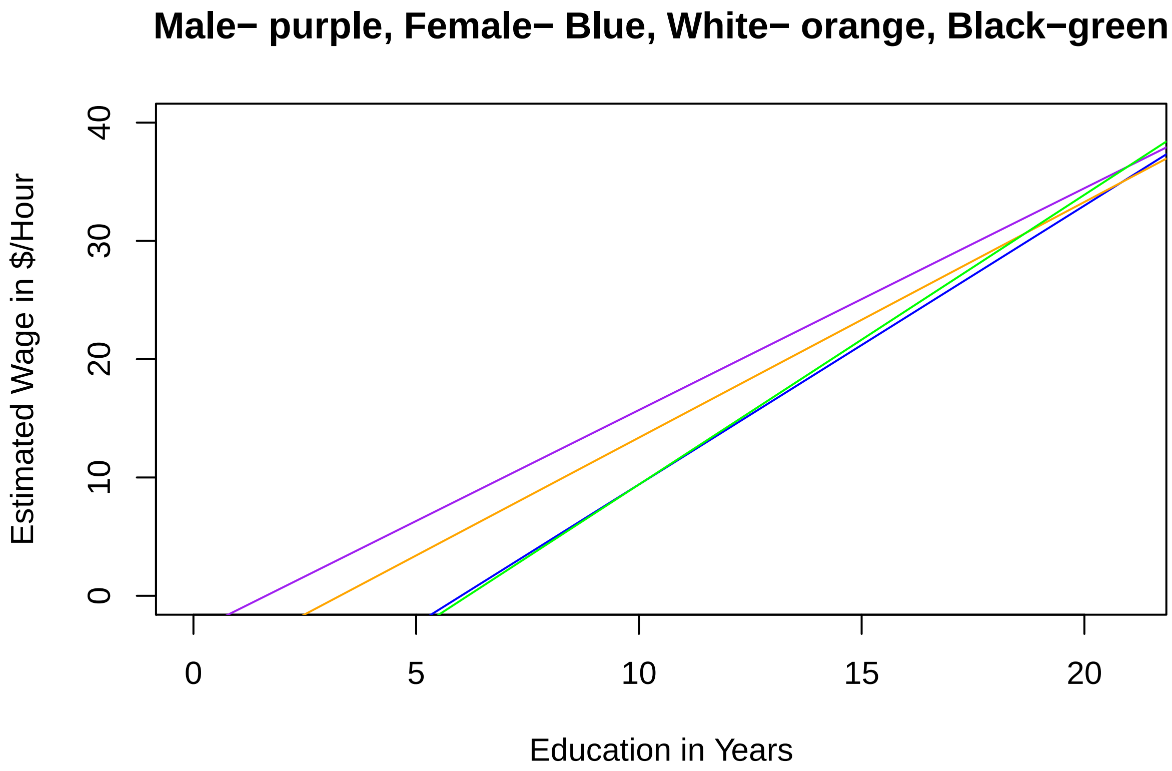
```
##        1        2        3
## 17.35255 21.33641 25.32028
```

```
predict.lm(reg_black, newdata =new.dat)
```

```
##        1        2        3
## 14.30277 19.20088 24.09900
```

```
plot(c(0,21),c(0,40), col = "white", main =
       "Male- purple, Female- Blue, White- orange, Black-green",
     xlab = "Education in Years",
     ylab = "Estimated Wage in $/Hour")
```

```
abline(reg_male, col = "purple")
abline(reg_female, col = "blue")
abline(reg_white, col = "orange")
abline(reg_black, col = "green")
```



```
#needs fixed
#ggplot(aes(x = 1:20, y = seq(1,80, length.out = 20))) +
#  geom_abline(intercept = coef(reg_male)[1], slope = coef(reg_male)[2]) +
#  geom_abline(intercept = coef(reg_female)[1], slope = coef(reg_female)[2])
```

e

Estimate the quadratic regression $WAGE = \alpha_1 + \alpha_2 EDUC^2 + e$ and discuss the results. Estimate the marginal effect of another year of education on wage for a person with 12 years of education and for a person with 14 years of education. Compare these values to the estimated marginal effect of education from the linear regression in part (b).

Solution We see the coefficient on $educ^2$ is estimated to be .0735. The marginal effect of education on wage is the derivative of wage with respect to education, which is $2 \times educ$. Thus, in the quadratic model, the returns are .88 dollars per hour and 1.02 dollars per hour for 12 and 14 years respectively. This is in comparison with the original linear model. For the linear model, the returns to education are constant (slope, take the derivative if confused) and estimated to be 1.98. Thus, our new model has smaller marginal returns to education for the 12th to 13th year and the 14th to 15th year.

```
attach(dat3)
quad_reg <- lm(wage ~ 1 + I(educ^2)) #note, must use I()!
```

```
    #because it's a formula object
summary(quad_reg)
```

```
##
## Call:
## lm(formula = wage ~ 1 + I(educ^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.242  -7.665  -2.437   4.977  55.903
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.082831   1.023161   5.945 3.82e-09 ***
## I(educ^2)   0.073489   0.004832  15.210  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.57 on 998 degrees of freedom
## Multiple R-squared:  0.1882, Adjusted R-squared:  0.1874
## F-statistic: 231.3 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
quad_reg_coef <- quad_reg$coefficients
quad_reg_coef
```

```
## (Intercept)   I(educ^2)
##  6.08283101  0.07348906
```

```
12*quad_reg_coef[2]
```

```
## I(educ^2)
## 0.8818687
```

```
14*quad_reg_coef[2]
```

```
## I(educ^2)
##  1.028847
```

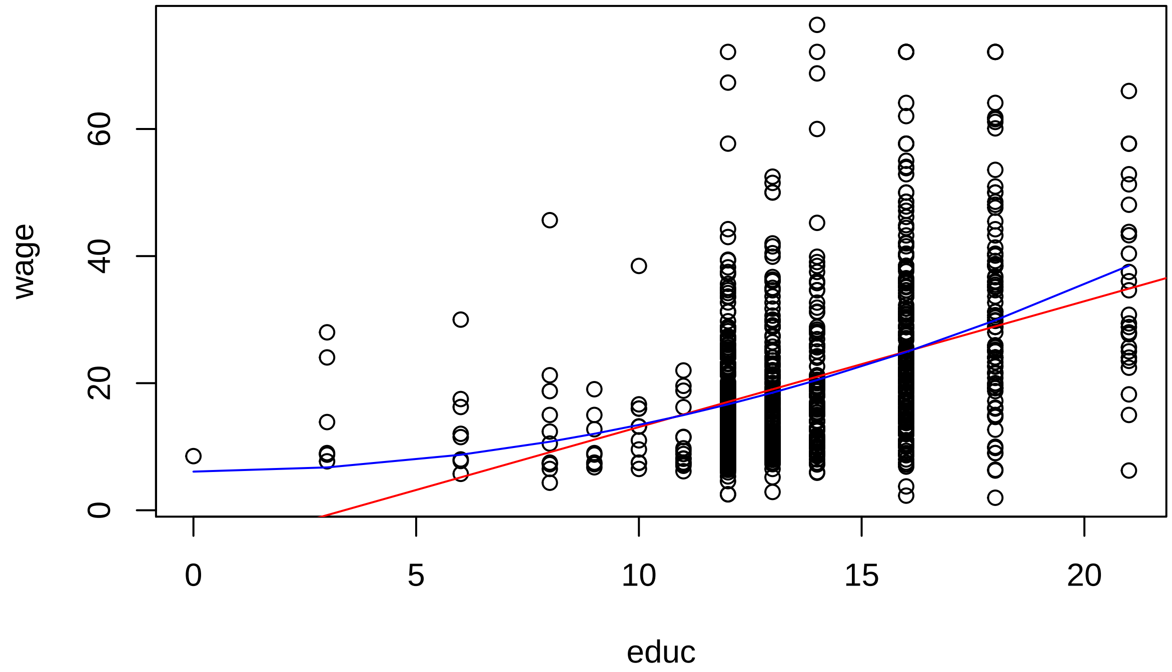
f

Plot the fitted linear model from part (b) and the fitted values from the quadratic model from part (e) in the same graph with the data on $WAGE$ and $EDUC$. Which model appears to fit the data better?

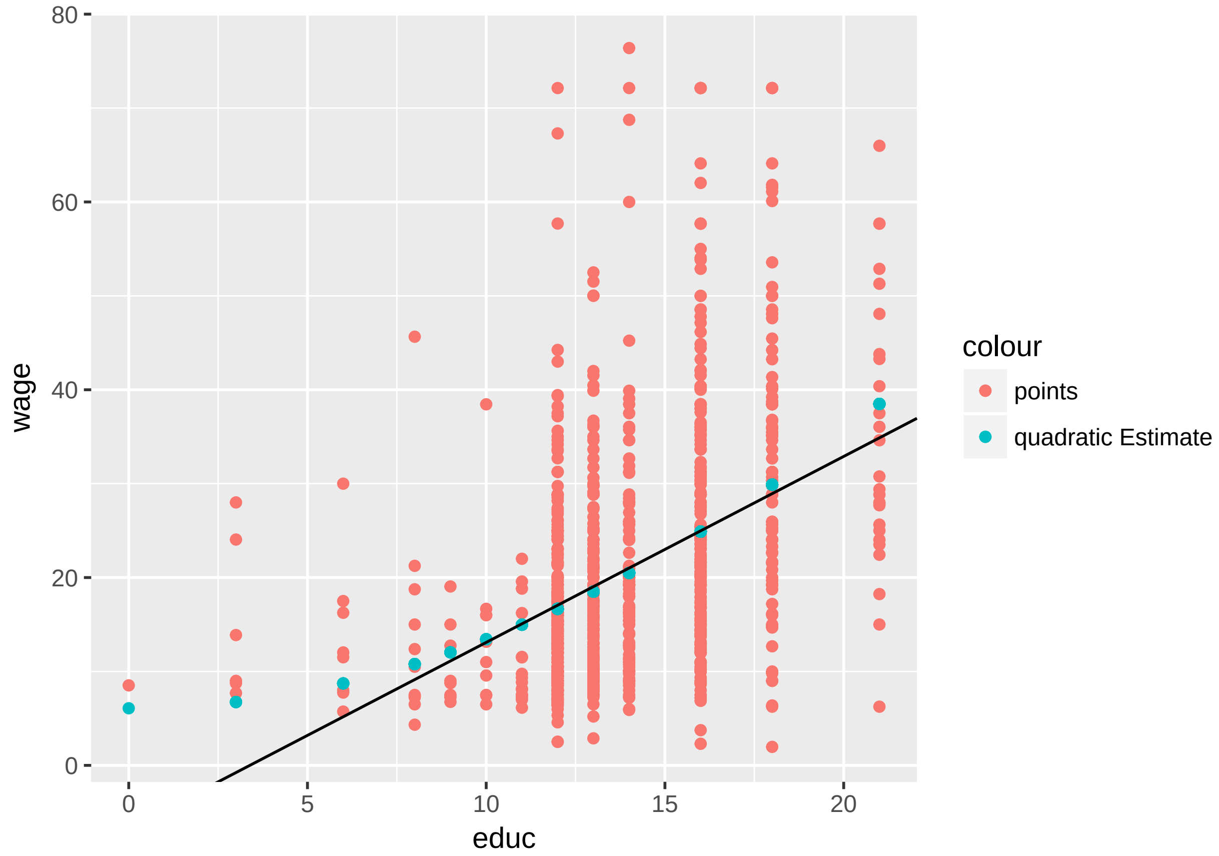
Solution The plots are below. The quadratic fit appears better.

Note, there are actually many, many formal ways of testing which model fits better. The problem is called model selection

```
quad_wage_est = quad_reg_coef[1] + sort(educ,decreasing = F)^2*quad_reg_coef[2]

plot(educ, wage)
abline(reg_out, col = "red")
lines(sort(educ,decreasing = F), quad_wage_est,col="blue")
```



```
#abline(quad_reg, col = "blue") #doesn't work

dat4 <- cbind(dat3,quad_wage_est)
ggplot(data = dat4) +
  geom_point(aes(x = educ, y = wage, color = "points")) +
  geom_point(aes(x = sort(educ,decreasing = F), y = quad_wage_est,color = 'quadratic Estimate')) +
  geom_abline(intercept = my_coef[1], slope = my_coef[2])
```

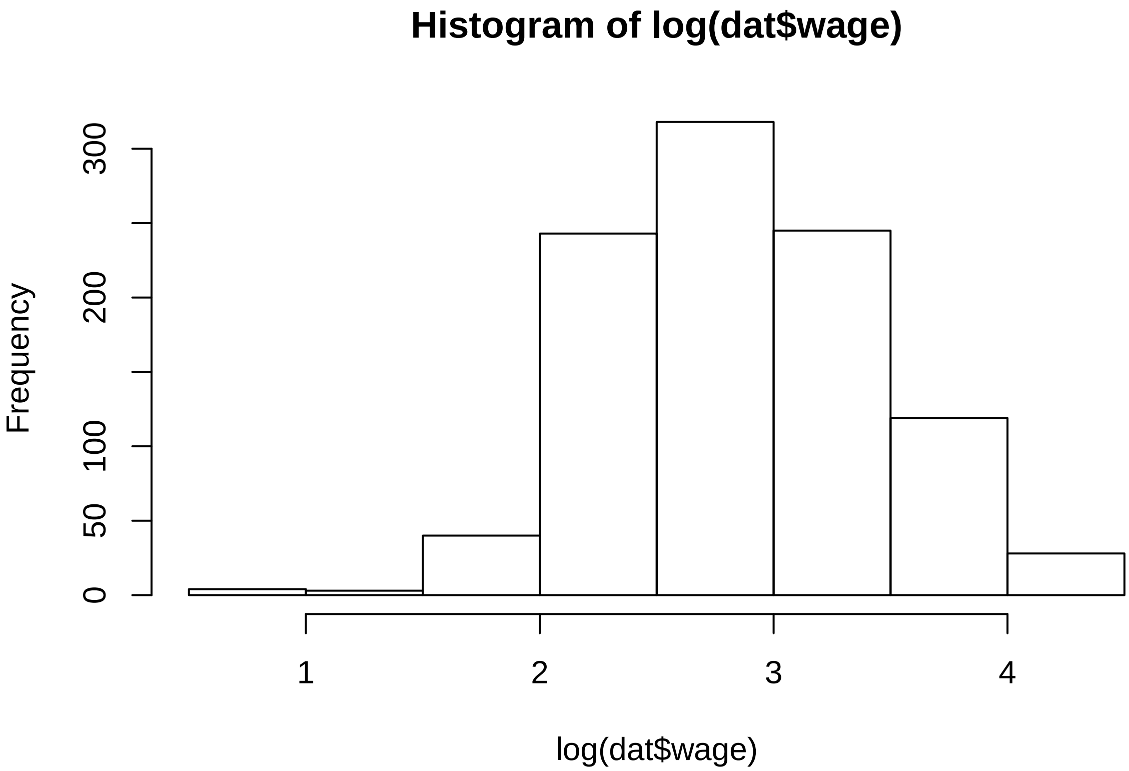


g

Construct a histogram of $ln(WAGE)$ (Note, log is usually natural log now too, even though in grade school it was the abbreviation for the base-10 log). Compare the shape of this histogram to that for $WAGE$ from part (a). Which appears more symmetric and bell-shaped?

Solution Histogram is below. This one is much more symmetric and bell-shaped.

```r
hist(log(dat$wage))
```



h

Estimate the log-linear regression $\log(WAGE) = \gamma_1 + \gamma_2 EDUC + e$. Estimate the marginal effect of another year of education on wage for a person with 12 years of education and for a person with 15 years of education. Compare these values to the estimated marginal effects of education from the linear regression in part (b) and the quadratic equation in part (e).

Solution Using implicit differentation, taking the derivative of both sides with respect to education, we have

$$\frac{1}{\widehat{WAGE}} \frac{d\widehat{WAGE}}{dEDUC} = \hat{\gamma}_2$$

and thus the marginal effect of wage on education is $\widehat{WAGE} \times \hat{\gamma}_2$ Thus, our estimate is

$$ga\hat{m}ma_2\widehat{WAGE} = \gamma_2 \times \exp(\gamma_1 + \gamma_2 EDUC)$$

[1] So the marginal effect of 1 more year of schooling for a person with 12 years is or 14 years is 1.338 dollars per hour and 1.603. These numbers are a little higher than the earlier quadratic estimates but a little lower than the linear estimates.

```r
log_wage <- log(wage)
log_dat <- cbind(dat3,log_wage)
log_reg <- lm(dat = log_dat, log_wage ~ educ)
summary(log_reg)
```

[1] In chapter 4, we will learn a better predictor of $\widehat{WAGE}$

```
##
## Call:
## lm(formula = log_wage ~ educ, data = log_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.55876 -0.39176  0.00699  0.36057  1.58413
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.609444   0.086423   18.62   <2e-16 ***
## educ        0.090408   0.006146   14.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5266 on 998 degrees of freedom
## Multiple R-squared:  0.1782, Adjusted R-squared:  0.1774
## F-statistic: 216.4 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
log_coef <- log_reg$coefficients

#marginal estimates at 12 and 14 years of education
log_coef[2] *exp(log_coef[1] + log_coef[2]* 12)
```

```
##     educ
## 1.337662
```

```
log_coef[2] *exp(log_coef[1] + log_coef[2]* 14)
```

```
##     educ
## 1.602781
```

Stata Solution

Code graciously adapted from that provided by Conor Foley. All errors are my own

```
/////////////////////////// Question 2.15 \\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\

//Navigate to your preferred working directory
cd "C:\Users\ryanj\Dropbox\TA\Econ 103\Winter 2018\Data\s4poe_statadata"

//Read in data file (already in this directory)
use cps4_small.dta, clear

//the clear option removes all data in the workspace (in this case, variables
// x and y) prior to reading in cps4_small.dta however, non-workspace data
// such as the scalar setobsnum will not be deleted by this command

//When importing data, I could also choose to supply the path to the file.
// For example, if I am in directory STATA Work, and the file is in a subfolder
// STATA Work > Week 1 then I could give the command as:
// use "./Week 1/cps4_small.dta", clear

// part (a)
sum wage educ // summarize wage and educ variables
```

```
hist wage // plot histogram for wage
graph export "Wage Hist.pdf", replace //Export figure as a PDF
hist educ // plot histogram for educ
graph export "Educ Hist.pdf", replace

// part (b)
reg wage educ // regress wage on educ, including a constant (i.e. wage = b1 + b2*educ)
scalar marg_eff_lin = _b[educ] // store beta for eduction (= marginal effect), useful for parts (e) and
predict yhat_lin // calculate fitted values for this regression, useful for part (f)

// part (c)
predict ehat_lin, residual // calculate residuals from regression in part (b)
twoway scatter ehat_lin educ // plot scatter of residuals against educ
graph export "Linear Residual Plot.pdf", replace

// part (d)
reg wage educ if female == 0                // regression, using males only
reg wage educ if female == 1                // female only
reg wage educ if black == 1                 // black only
reg wage educ if (black == 0 & asian == 0)  // white (non-black, non-asian)

//Another option for the male vs female case would be to use STATAs "by"

// Example -
// sort female
// by female: reg wage educ
// --> This tells stata to run reg wage educ if female == ?
//     where ? is each unique value the variable female has in the data
// --> if data is not already sorted, you could also say:
// bysort female: reg wage educ

// part (e)
gen educ_sqr = educ^2 // generate educ squared
reg wage educ_sqr // run regression
predict yhat_quad // calculate fitted values, useful for part (f)

// in quadratic case, marginal effect is beta * 2 * educ
scalar marg_eff_quad_12 = _b[educ_sqr]*2*12 // calculate marginal effects at educ = 12
scalar marg_eff_quad_14 = _b[educ_sqr]*2*14 // calculate marginal effects at educ = 14

//Report the linear and quadratic marginal effects
disp marg_eff_lin
disp marg_eff_quad_12
disp marg_eff_quad_14

// part (f)
sort educ // need to sort data to get a sensible line for fitted values
          // when stata draws the line, it is literally "connecting the dots"
          // going in the order observations appear in rows of the workspace

// Variable labels become legend identifiers on figure
label variable wage "Raw data"
label variable yhat_lin yhat_lin
label variable yhat_quad yhat_quad

twoway scatter wage educ || line yhat_lin educ || line yhat_quad educ, ytitle(Earnings per hour)
// alternative syntax for plotting multiple twoway data plots:
// twoway (scatter wage educ) (line yhat_lin educ) (line yhat_quad educ)
// the ytitle option sets the Y axis label
graph export "Linear and Quad Fitted Values.pdf", replace

// part (g)
gen ln_wage = ln(wage) // generate natural log of wages

hist ln_wage //give histogram of log wages
graph export "Log Wage Hist.pdf", replace

// part (h)
reg ln_wage educ // Log-Linear regression using log wages
predict ln_yhat_log // fitted values (note - this fitted value is for ln_wage)
gen yhat_log = exp(ln_yhat_log) // transform log-level fitted value to regular-level

//For adjustment to log-level estimate (section 4.5.3 of textbook), calculate
// sigma_hat_sqr using rss/(N-rank) and values stored in e()
scalar define sigma_hat_sqr = e(rss)/(e(N)-e(rank))

//Implement adjustment from section 4.5.3
gen yhat_log_adju = yhat_log*exp(sigma_hat_sqr/2)

//Calculate marginal effects
scalar marg_eff_log_12 = exp(_b[_cons]+_b[educ]*12)*_b[educ] // = y_hat(educ = 12) * _b[educ]
scalar marg_eff_log_14 = exp(_b[_cons]+_b[educ]*14)*_b[educ] // = y_hat(educ = 14) * _b[educ]
scalar marg_eff_log_adju_12 = exp(_b[_cons]+_b[educ]*12+sigma_hat_sqr/2)*_b[educ] // = y_hat_adju(educ =
scalar marg_eff_log_adju_14 = exp(_b[_cons]+_b[educ]*14+sigma_hat_sqr/2)*_b[educ] // = y_hat_adju(educ =

//Report the linear, quadratic, log, and adjusted-log marginal effects on wages
disp marg_eff_lin
disp marg_eff_quad_12
disp marg_eff_quad_14
disp marg_eff_log_12
disp marg_eff_log_14
disp marg_eff_log_adju_12
disp marg_eff_log_adju_14

//Update labels for each fitted value - will show up on figure legend
label variable wage "raw data"
label variable yhat_lin yhat_lin
label variable yhat_quad yhat_quad
label variable yhat_log yhat_log
label variable yhat_log_adju yhat_log_adju

//Plot scatter and 4 fitted value series
twoway scatter wage educ || line yhat_lin educ || line yhat_quad educ || line yhat_log educ || line yhat
graph export "All Fitted Values.pdf", replace

//Additional Comments
```

```
//(1) Can generate a box and whisker plot for wages with command:
// graph box wage
//
//(2) Can generate a histogram with a density curve with option density
// Example: hist wage, density


//Convert log file (smcl) to pdf
translate wk1_section_log.smcl "Week 1 TA Section STATA.pdf"

log close
```

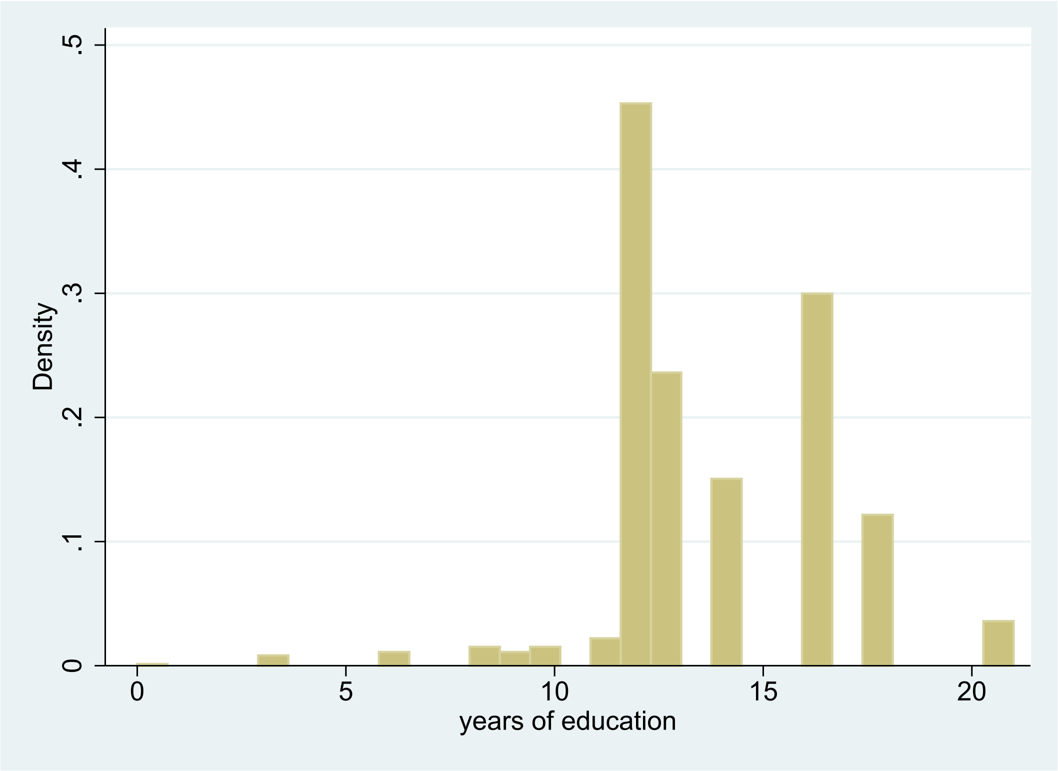


Figure 2: Educ Histogram

Outputed Figures

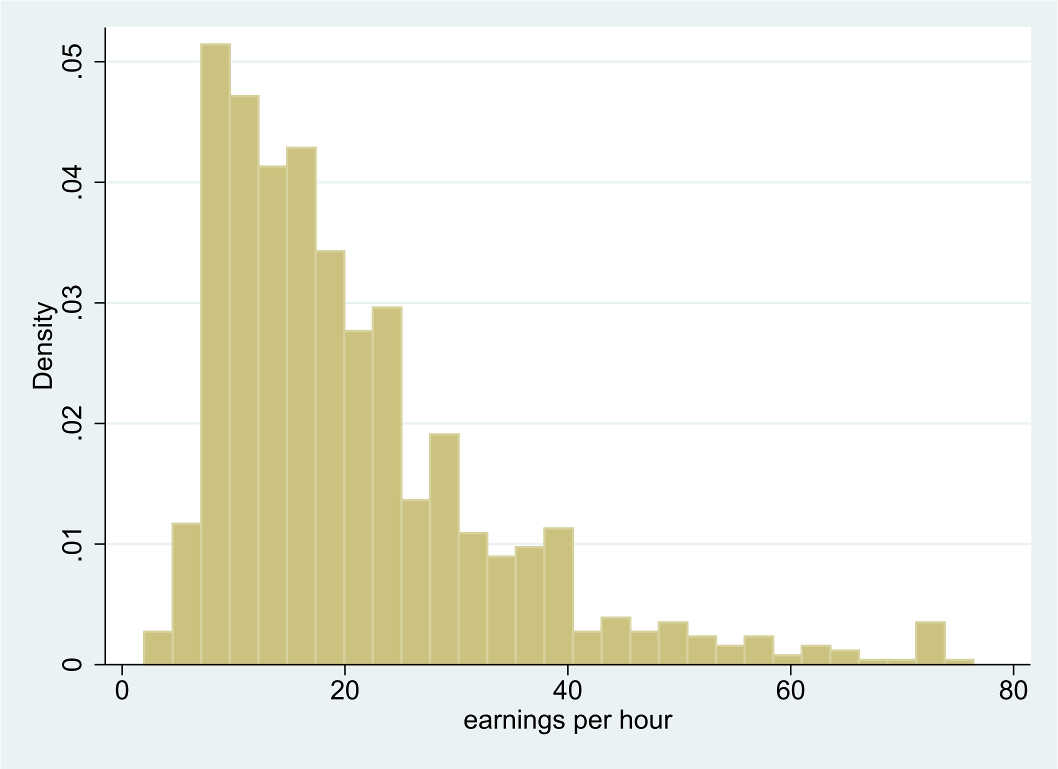


Figure 3: Wage Histogram

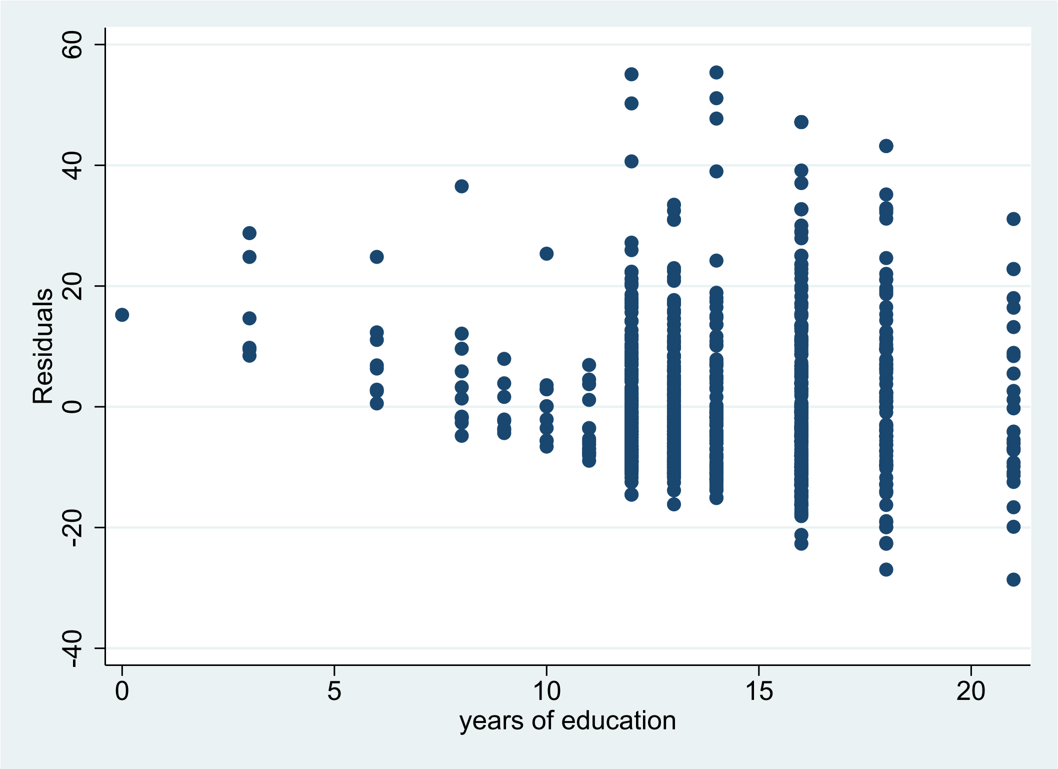


Figure 4: Linear Residual Plot

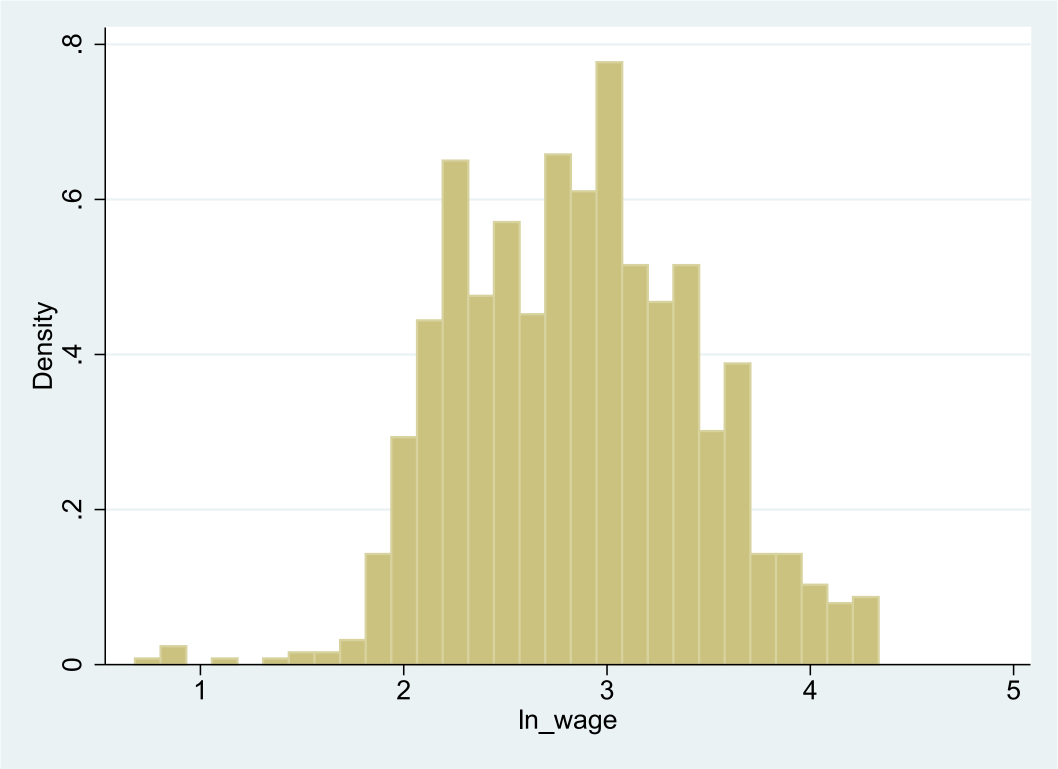


Figure 5: Log Wage Hist

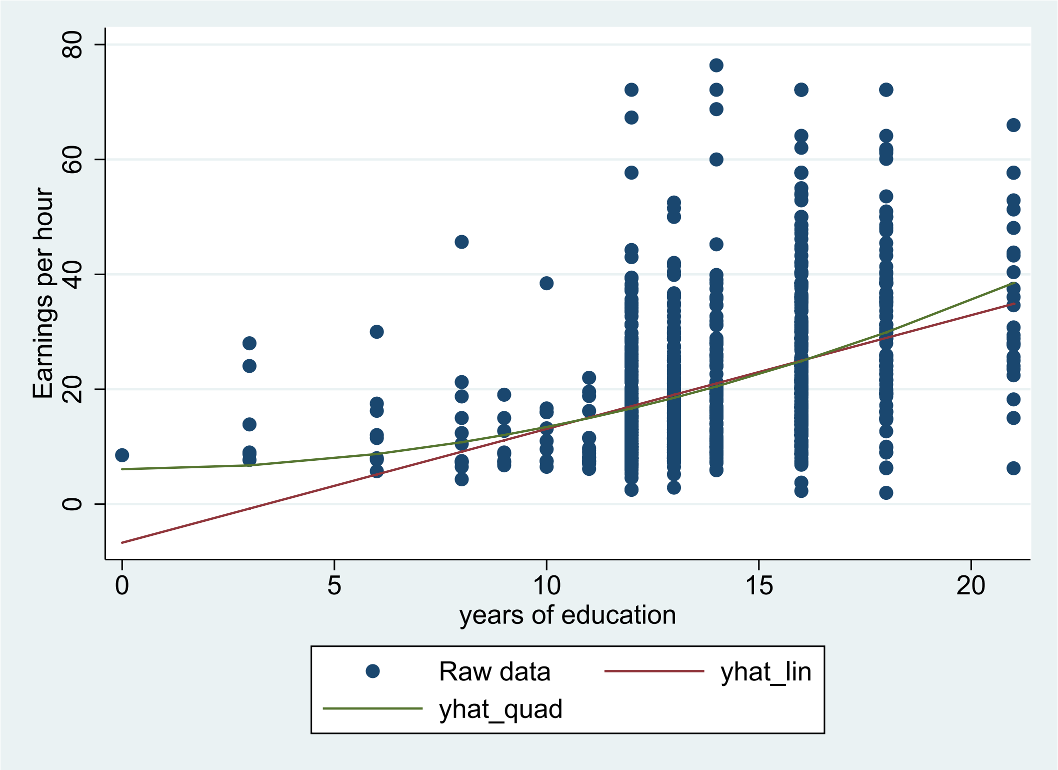


Figure 6: Linear and Quadratic Fitted Values

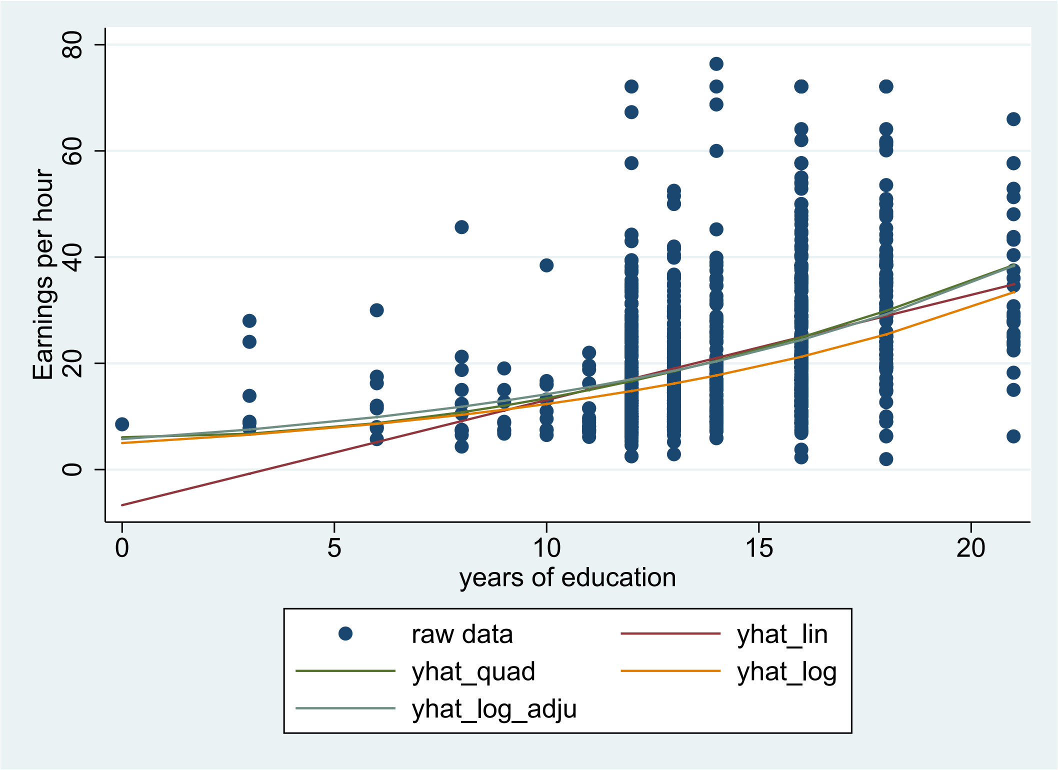


Figure 7: All Fitted Values