STAT 223: Project 1

Nrepesh Joshi

Professor Kevin Hastings

Applied Analytics

4/9/20

The Effect of Educational Level on Income

By how much will another year of schooling most likely raise one's income?

==========

ACCOMPANYING DATA PROVIDED BY: Guido Imbens, PhD

UCLA, Department of Economics

==========

==========

BRIEF DESCRIPTION OF THE DATA

==========

The data were collected by a team of five interviewers at the 16th Annual Twins Day Festival in Twinsburg, Ohio, in August 1991. A booth was set up at the festival's main entrance, and an ad inviting all adult twins to participate in the survey was placed in the festival program. In addition, the interviews roamed the festival grounds, approaching all adult twins for an interview, and almost every pair of twins accepted. In total, 495 individuals over the age of 18 were interviewed.

==========

HOW TO USE THE DATA FILE

==========

The data file is comma delimited text. Each row contains information on a pair of twins for sixteen variables, which are explained below. Note that everyone in a pair of twins was randomly assigned a number: twin 1 or twin 2.

There is missing data. Examine the data file and decide how to identify it.

NOTE: For most data analysis purposes, the logarithm of the hourly wage is used instead of the hourly wage itself.

DLHRWAGE.....the difference (twin 1 minus twin 2) in the logarithm of hourly wage, given in dollars.
DEDUC1.....the difference (twin 1 minus twin 2) in self-reported education, given in years.
AGE.....Age in years of twin 1.
AGESQ.....AGE squared.
HRWAGEH.....Hourly wage of twin 2.
WHITEH.....1 if twin 2 is white, 0 otherwise.
MALEH.....1 if twin 2 is male, 0 otherwise.
EDUCH.....Self-reported education (in years) of twin 2.
HRWAGEL.....Hourly wage of twin 1.
WHITEL.....1 if twin 1 is white, 0 otherwise.
MALEL.....1 if twin 1 is male, 0 otherwise.
EDUCL.....Self-reported education (in years) of twin 1.
DEDUC2.....the difference (twin 1 minus twin 2) in cross-reported education. Twin 1's cross-reported education, for
example, is the number of years of schooling completed by twin 1 as reported by twin 2.
DTEN.....the difference (twin 1 minus twin 2) in tenure, or number of years at current job.
DMARRIED.....the difference (twin 1 minus twin 2) in marital status, where 1 signifies "married" and 0 signifies
"unmarried".
DUNCOV.....the difference (twin 1 minus twin 2) in union coverage, where 1 signifies "covered" and 0 "uncovered".

==========
DATA ANALYSIS
==========

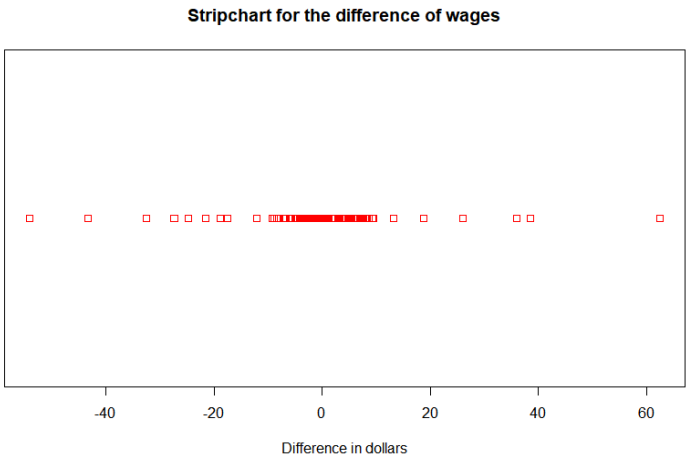
1. First, we need a convenient way to eliminate the missing data. After downloading the data, open the file in Excel or Google Sheets and manually delete all rows in which there is a column with missing data. Save a copy for import into R in your R home directory (probably MyDocuments in Windows systems).

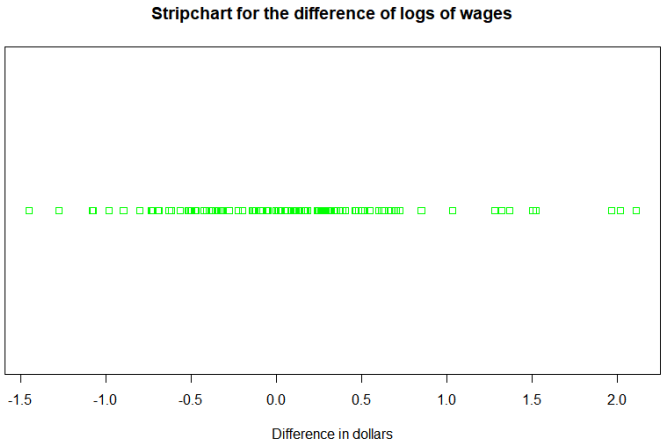
= Manually deleted all rows in which there is a column with missing data. File attached "Nrepesh Joshi - Twins (Missing Removed).csv"

2. The authors chose to analyze the difference of logs of wages (the first variable). Why? Produce a stripchart of these logs. Also create a new variable which is the difference between the hourly wage of twin1 and twin 2 and examine a stripchart of that. What features stand out that might explain the authors' choice?

= The reason the author chose to analyze the difference of logs of wages (DLHRWAGE) is because log transformation makes the data somewhat symmetric and approximately normal. The reason we want the data to be approximately normal is because magnitude is equidistance on the log axis and highly skewed distribution will be less skewed. Another advantage is that, analysis conclusions in the log scale can draw the same conclusion in the regular scale. Hence, the author chose to analyze the difference of logs of wages.

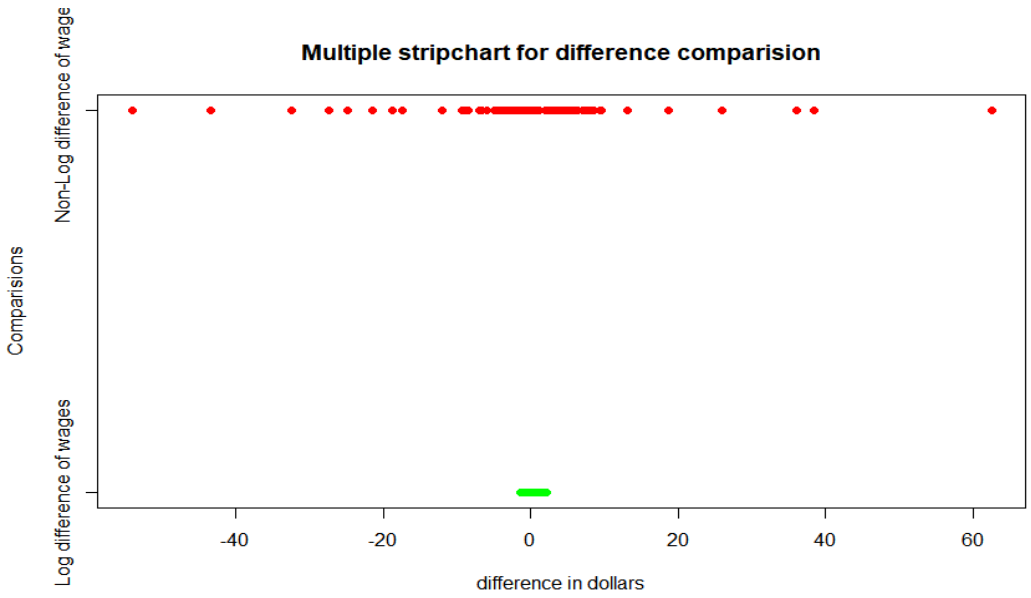
We can show this by using strip charts in R.





In the stripchart for difference in wages, we see a range from around 80 to -60. The data points are all scattered.

In the stripchart for difference in log wages, we see that the data is now scaled to a range of -1.5 to 2.0. The analysis that we do in logs will be equivalent to what we do in with the regular data. We can see the difference more clearly when we juxtapose the graphs in the same axis below.



---

3. Now run a regression of difference in log hourly wage as a function of difference in self-reported education. Is increased education a significant predictor of increased wages? What is the expected increase in log hourly wage per additional year of education? How does that translate to additional wages per additional year of education, over a full year of earnings at 40 hours/week for 52 weeks?

= After implementing a difference in log hourly wages as a function of difference in self-reported education, the summary and graph of the model is as follows.

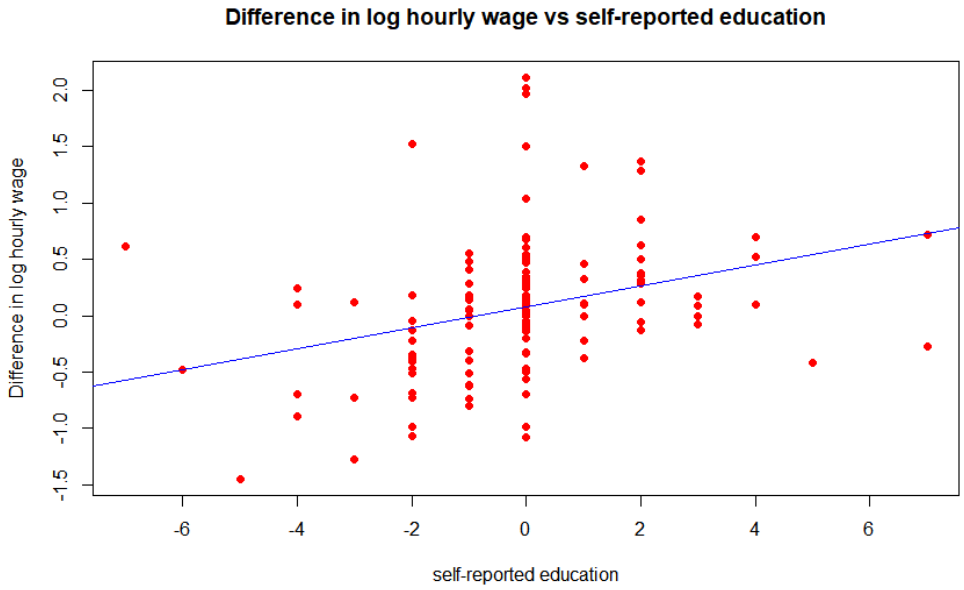
```
Call:
lm(formula = dset$DLHRWAGE ~ dset$DEDUC1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.15740 -0.34395 -0.00722  0.20909  2.03115

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07859    0.04547   1.728 0.086022 .
dset$DEDUC1  0.09157    0.02371   3.862 0.000168 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5542 on 147 degrees of freedom
Multiple R-squared:  0.09211,   Adjusted R-squared:  0.08593
F-statistic: 14.91 on 1 and 147 DF,  p-value: 0.0001682
```

Question 3



We see from the linear increase in graph that increased education is a predictor of increased wages but only a weak predictor. The R squared value is very close to 0 but the slope coefficients are statistically significant from the model summary. With the linear equation of y = 0.09157 x + 0.07859 it does follow a linear regression. That said, the expected increase in log hourly wage per additional year of education is given by the slope of the graph which is 0.09157. So, for each additional year of education, our model predicts the expected increase in log hourly wage to be 0.09157.

Here, wages in Log hours does not help us estimate real values. The log transformation made it very easy to understand and analyze data, but we need to convert is back to get the actual increase in log hours. Now when we take the exponential of the slope which is exp(0.09157) we get 1.097 which is the actual increase in wages for each additional year of education. So, for each week the increase is 40 * 1.097 = $43.88 per week. For 52 weeks the increase is 43.88 * 52 = $ 2,281.76. Hence, over a full year of earnings at 40 hours/week for 52 weeks, additional wages per additional year of education is $2281.76. We conclude that the yearly increase in wages is around 200 times the current hourly level. If the latter was $15/hour for instance, that would come to about $3000 per year.

4. Repeat 3 for log hourly wage as a function of difference in cross-reported education. Which slope coefficient is greater?

Do your results bear out what the authors say about the underestimation effect of inaccurate reporting?

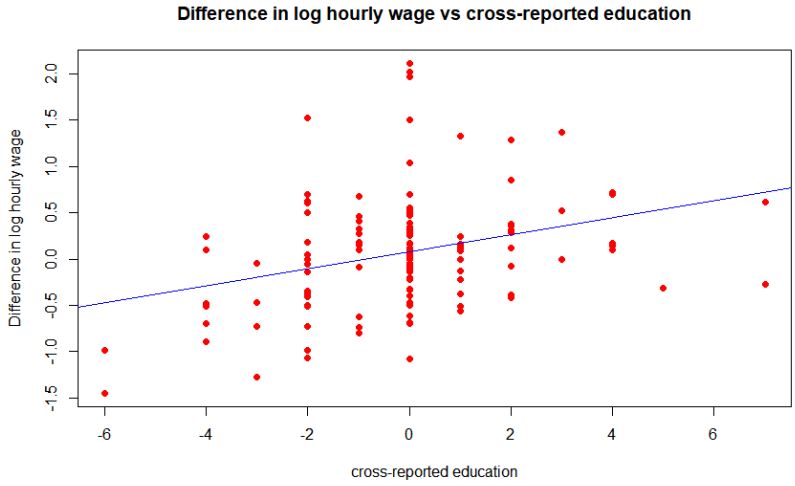
```
Call:
lm(formula = dset$DLHRWAGE ~ dset$DEDUC2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.15687 -0.30913 -0.03687  0.20963  2.03169

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07806    0.04529   1.724   0.0869 .
dset$DEDUC2  0.09234    0.02297   4.020 9.28e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5521 on 147 degrees of freedom
Multiple R-squared:  0.09903,   Adjusted R-squared:  0.0929
F-statistic: 16.16 on 1 and 147 DF,  p-value: 9.28e-05
```

Question 4



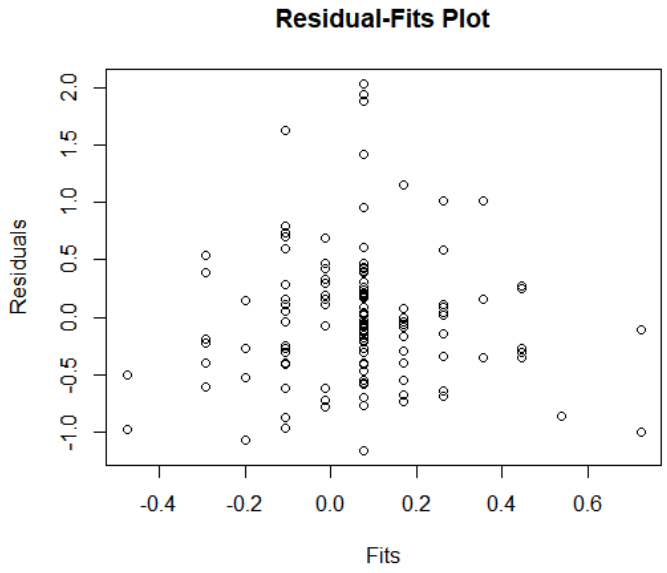
The slope coefficient is slightly greater in cross reported education by 0.001 which is a very slight difference. The model summary is also very close to the previous one with slope coefficients being statistically significant. Hence the results do not bear our what the author says about the underestimation effect of inaccurate reporting.

5. Finally, do diagnostic checking. Produce the residuals and fitted values from your model, and have R graph residuals vs. fits, and a histogram of residuals. Do the assumptions of regression seem to hold?

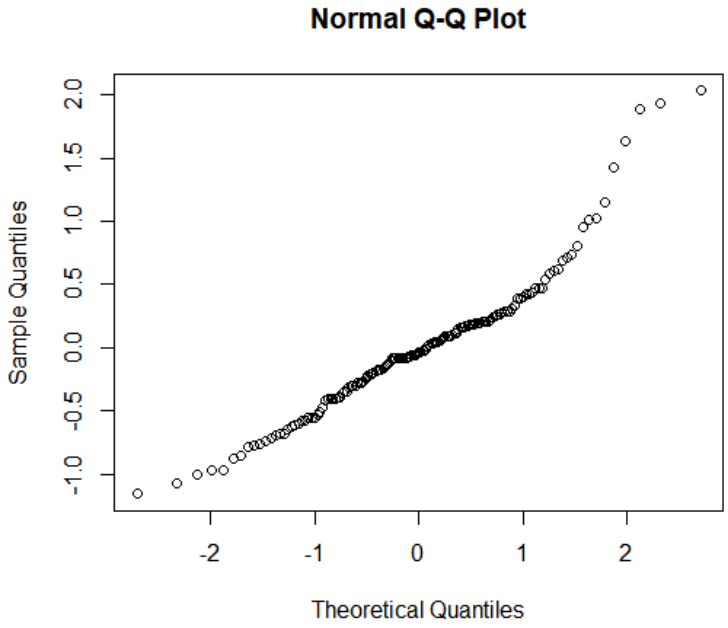
Here I will be doing model checking for difference in log hourly wage vs cross-reported wage because the author is leaning towards this hypothesis.

Assumption 1: The variables Xi are the only predictor variables of interest.
This assumption could be false because our R^2 value is quite low. This means that the education level that we accounted for might not be the only predictor variable of interest to figure out increase in wages. More knowledge about the study can help us more with this assumption.

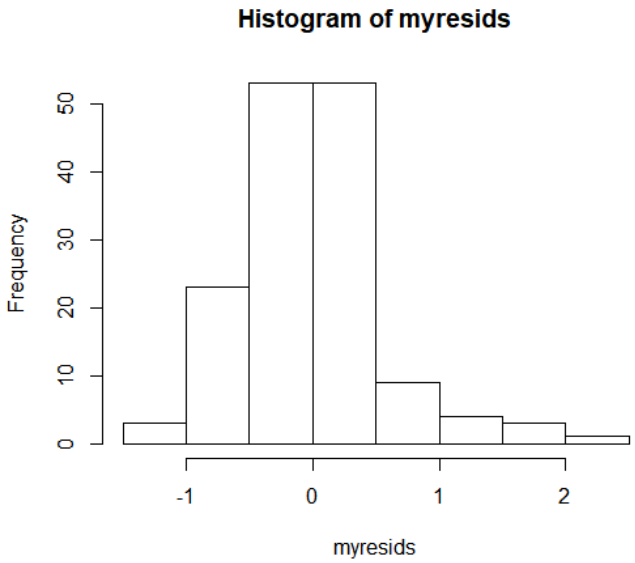
Assumption 2: The function f which determines the way the mean of Y depends on the predictors is correctly specified.
= This might also be false because if the R^2 value is close to 0 then the linear regression model does not fit quite well. We can see this from our “Residual vs Fitts” as well. We see a concave down quadratic pattern from which we can conclude that this assumption is false.



Assumption 3: The errors are normally distributed with mean 0.
= From the Q-Q plot below we do not see a straight line (not linear) as theoretical quantiles increases. Most of the plot looks linear so the errors are normally distributed with mean 0.

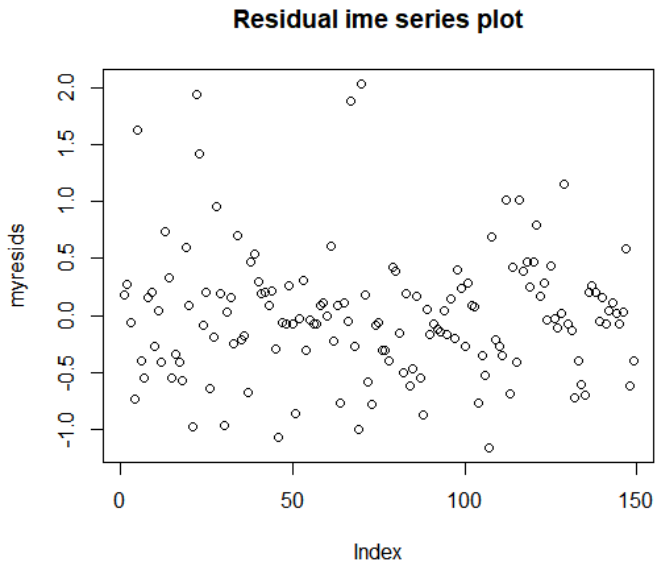


From our histogram plot we see that it is skewed right.



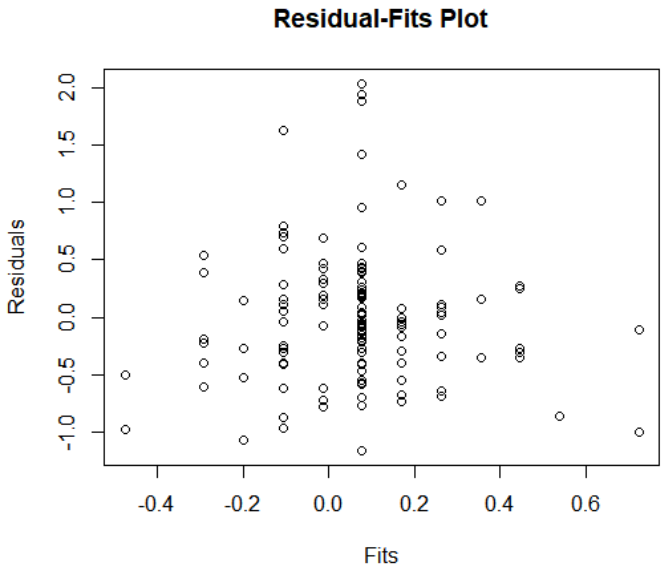
Assumption 4: The errors are mutually independent.

When plotting time series graph to check independence, we see a patten which falsifies this assumption.

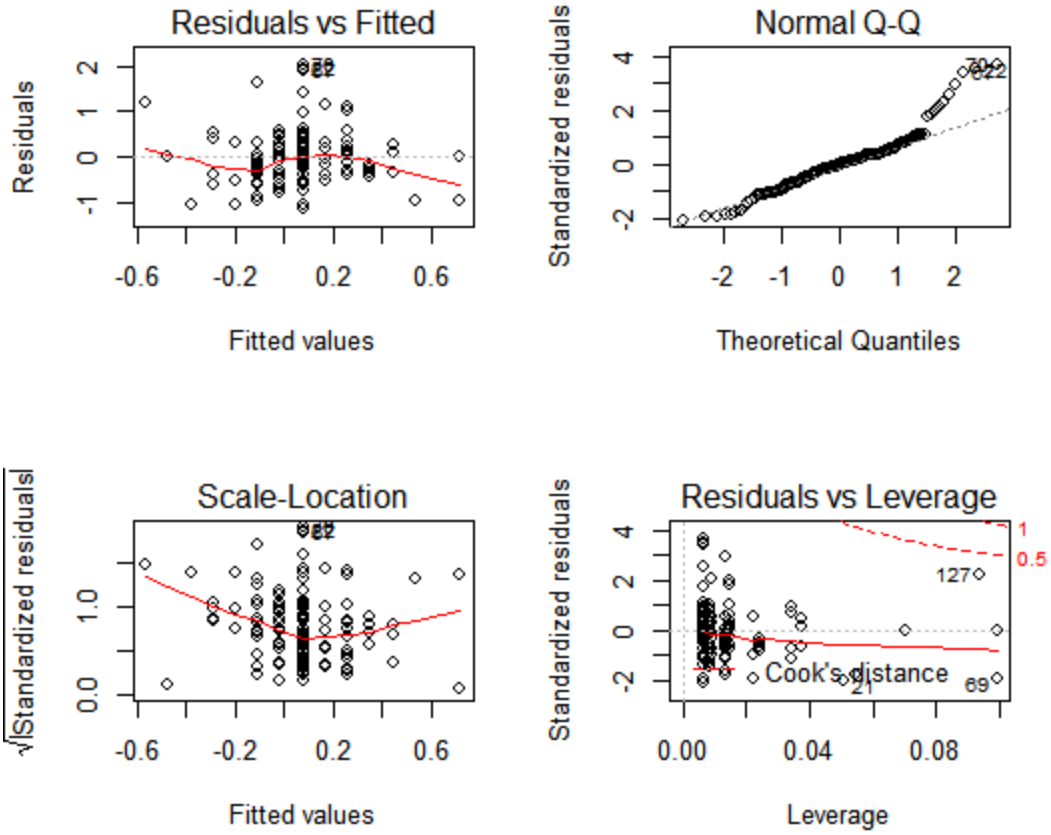


Assumption 3: Homoscedasticity

= To check for homoscedasticity, we see that the variance is equal. It does increase in the middle of the plot.



Other helpful plots:



FULL R code script:

#----------

Part 1 Importing data

dset <- read.csv("Nrepesh Joshi - Twins (Missing Removed).csv", header = TRUE);
head(dset)

#----------

Part 2 Strip Charts

Stripchart for the difference of logs of wages

```
stripchart(dset$DLHRWAGE, main="Stripchart for the difference of logs of wages",
xlab="Difference in dollars",
        col="green")

# New variable for difference in hourly wages of twin 1 and twin 2

dset$DHRWAGE <- dset$HRWAGEL - dset$HRWAGEH; View(dset)

# Stripchart for the difference of wages

stripchart(dset$DHRWAGE, main= "Stripchart for the difference of wages", xlab="Difference in
dollars",
        col="red")

# Multiple strip charts

x <- list("Log difference of wages"=dset$DLHRWAGE, "Non-Log difference of
wages"=dset$DHRWAGE)
stripchart(x,
        main="Multiple stripchart for difference comparision",
        xlab="difference in dollars",
        ylab="Comparisions",
        col=c("green","red"),
        pch=16
)

#-------------------------------------------------------------------------------------

# Part 3 Linear Regression

model1 <- lm(dset$DLHRWAGE ~ dset$DEDUC1); summary(model1)
plot(dset$DEDUC1,dset$DLHRWAGE, main = "Difference in log hourly wage vs self-reported
education",
    xlab = "self-reported education",
```

```
    ylab = "Difference in log hourly wage",
    pch = 16,
    col = "red")
cor(dset$DEDUC1,dset$DLHRWAGE)
abline(model1, col="blue")

#-------------------------------------------------------------------------------------------

# Part 4 Second Linear Regression

model2 <- lm(dset$DLHRWAGE ~ dset$DEDUC2); summary(model2)
plot(dset$DEDUC2,dset$DLHRWAGE, main = "Difference in log hourly wage vs cross-reported
education",
    xlab = "cross-reported education",
    ylab = "Difference in log hourly wage",
    pch = 16,
    col = "red")
cor(dset$DEDUC2,dset$DLHRWAGE)
abline(model2, col="blue")
exp(0.09234)

#-------------------------------------------------------------------------------------------

# Part 5 Diagonostic testing

par(mfrow=c(2,2)) # Split screen 2,2

plot(model1)

par(mfrow=c(1,1))

myfits <- fitted(model2)

myresids <- residuals(model2)
```

```
qqnorm(myresids)    # Checks normality of model

# Residual vs fits plot
plot(myfits, myresids, main = "Residual-Fits Plot", xlab = "Fits", ylab = "Residuals")

# Time series plot
plot(myresids, main = "Residual time series plot")

# Residual histogram
hist(myresids)
```