The Effect of Educational Level on Income

General Explanation of the Study

The original study that used these data sought to answer what seems a simple question: By how much will another year of schooling most likely raise one's income? Attempts had been made to estimate the value of a year's education in previous studies, but previous estimates may have been imprecise for two reasons. The first, most obvious reason is the difficulty of extracting education's effect on income from the effect that other variables related to education have on income. That is, a worker's natural ability, his family background, and his innate intelligence are all possible confounding factors that must be controlled for to estimate the effect of education on income accurately. Thus this study interviewed twins, collecting information about education, income, and background. Because monozygotic twins (twins from a single egg) are genetically identical and have similar family backgrounds, they provide excellent control for confounding variables.

The second difficulty in measuring the effect of income on education has to do with the false reporting of education levels, and this study is the first to address it. Since people are more likely to report a higher education level than they have actually attained, especially in face-to-face interviews, the data will contain a number of people with lower education levels in the higher education categories. Thus, since education usually increases income, estimates for the precise amount of this increase will be too low. To correct for this bias the researchers interviewed the twins separately and recorded two entries for each individual's education level: his self-reported education level and the education level reported by his twin. This allowed them to estimate the "measurement error" of reported education levels and correct for it. The result was a much higher estimate of the effect a year of education is likely to have on one's income. In fact, this study's estimates were higher than those of all previous studies, which did not correct for measurement error in education level.

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.

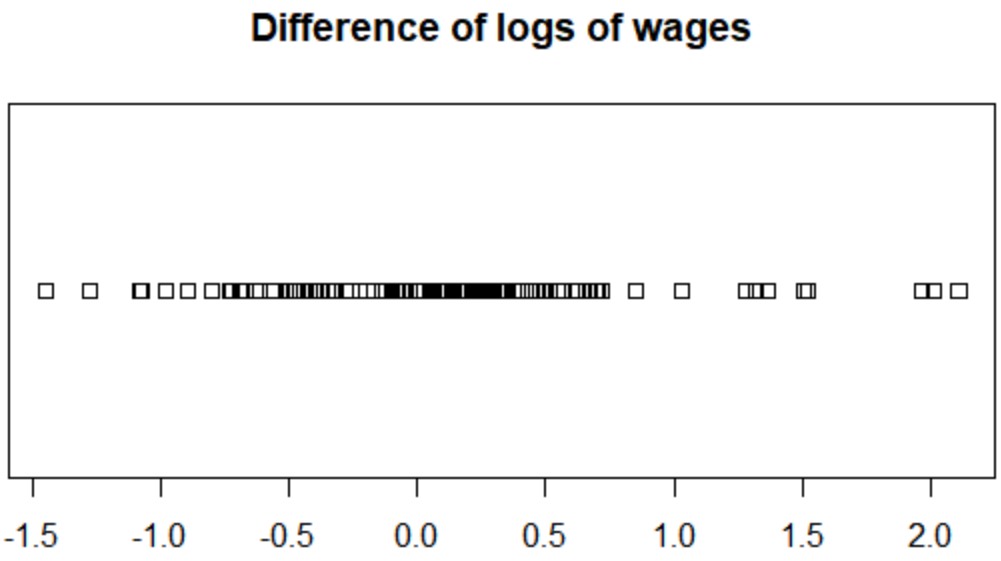
In total, 495 individuals over the age of 18 were interviewed.

Data Processing

The authors chose to analyze the difference of logs of wages. First of all, as the study's goal is to test the effect of educational level on income, wages become a dependable variable in the study. Secondly, the use of logarithm eliminates the heteroscedasticity (unequal variability)

that produces the fanning out effect in the scatterplot of the residuals. Therefore, by using the difference of logs of wages, the author can test his hypothesis with homoscedastic Ys values.

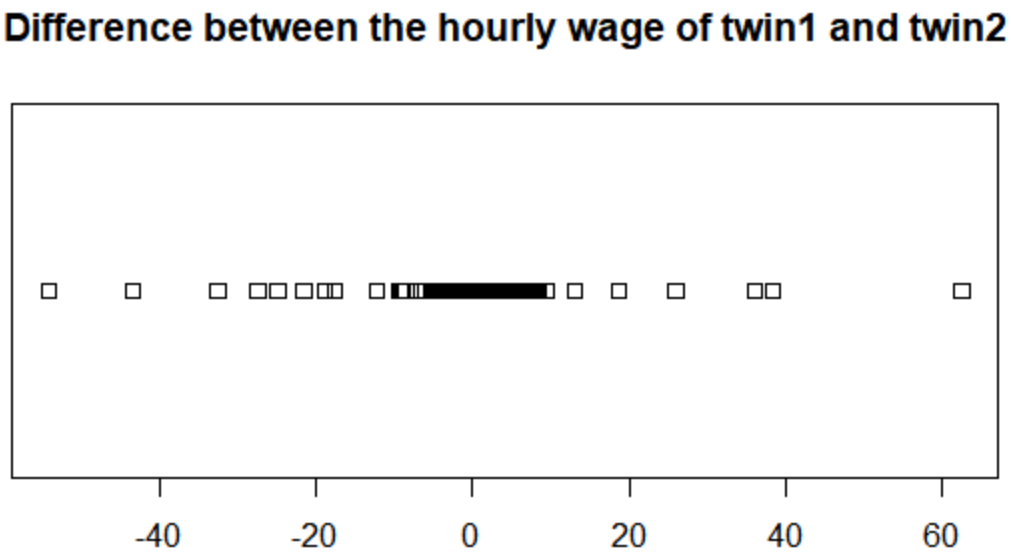
Produce a stripchart of these logs:
> stripchart(Stat_Project_1_cut$DLHRWAGE,main="Difference of logs of wages")



Create a new variable that is the difference between the hourly wage of twin1 and twin 2:
> Stat_Project_1_cut$DHRWAGE<- Stat_Project_1_cut$HRWAGEL - Stat_Project_1_cut$HRWAGEH

Produce a stripchart of the difference:

```
> stripchart(Stat_Project_1_cut$DHRWAGE, main="Difference between the hourly wage of twin1 and twin2")
```



The outliers in the new variable can clearly explain the author's choice. With the application of the logarithm, the influence of the outliers can be reduced. The distribution is more normal as the residuals are approximately symmetrically distributed.

Generate a regression of difference in log hourly wage as a function of the difference in self-reported education:

```
> regression<-lm(formula = Stat_Project_1_cut$DLHRWAGE ~ Stat_Project_1_cut$DEDUC1)
> summary(regression)

Call:
lm(formula = Stat_Project_1_cut$DLHRWAGE ~ Stat_Project_1_cut$DEDUC1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.15581 -0.34628 -0.02782  0.21069  2.03275

Coefficients:
                              Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)                    0.07700    0.04610   1.670   0.097022 .
Stat_Project_1_cut$DEDUC1  0.09135    0.02396   3.812   0.000203 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5578 on 145 degrees of freedom
Multiple R-squared:  0.0911,       Adjusted R-squared:  0.08484
F-statistic: 14.53 on 1 and 145 DF,  p-value: 0.0002029
```

Since the P-value is really small (0.0002029), it can be indicated that increased education is a significant predictor of increased wages.

The expected increase in log hourly wage per additional year of education is 0.09135.

=> Δ (ln(hourlywage1) - ln(hourlywage2)) = 0.09135

⇔ Δ ln(hourlywage1/hourlywage2) = 0.09135

⇔ Δ e^(ln(hourlywage1/hourlywage2)= e^0.09135

⇔ Δ hourlywage1/hourlywage2 = 1.0957

⇔ Δ hourlywage1 = 1,0957 * hourlywage2

⇔ Δ hourlywage1 * 52 * 40 = 1.0957 * hourlywage2 * 52 * 40

⇔ Δ yearlywage1 = 1.0957 * yearlywage2

⇔ Δ (yearlywage1 - yearlywage2) = (e^0.09135 - 1) * yearlywage2

Generate a regression of difference in log hourly wage as a function of the difference in cross-reported education:

```
> regression2<-lm(Stat_Project_1_cut$DLHRWAGE ~ Stat_Project_1_cut$DEDUC2)
> summary (regression2)

Call:
lm(formula = Stat_Project_1_cut$DLHRWAGE ~ Stat_Project_1_cut$DEDUC2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.15528 -0.32304 -0.04672  0.21398  2.03328

Coefficients:
                              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                    0.07647    0.04591   1.666   0.097937 .
Stat_Project_1_cut$DEDUC2  0.09215    0.02321   3.970   0.000113 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.5556 on 145 degrees of freedom
Multiple R-squared:  0.09804,      Adjusted R-squared:  0.09182
F-statistic: 15.76 on 1 and 145 DF,  p-value: 0.0001127
```

The slope of the regression of difference in log hourly wage as a function of the difference in cross-reported education is greater (0.09215 > 0.09135).

This indicates that the data is more precise with the cross-reported method, which leads to a better model. Therefore, as stated in the explanation of the study, “the result was a much higher estimate of the effect a year of education is likely to have on one’s income”. Without cross-reported education, earlier studies depend only on the self-evaluation, which in this case, has been proved to be much less accurate. Thus, the results bear out what the authors say about the underestimation effect of inaccurate reporting.

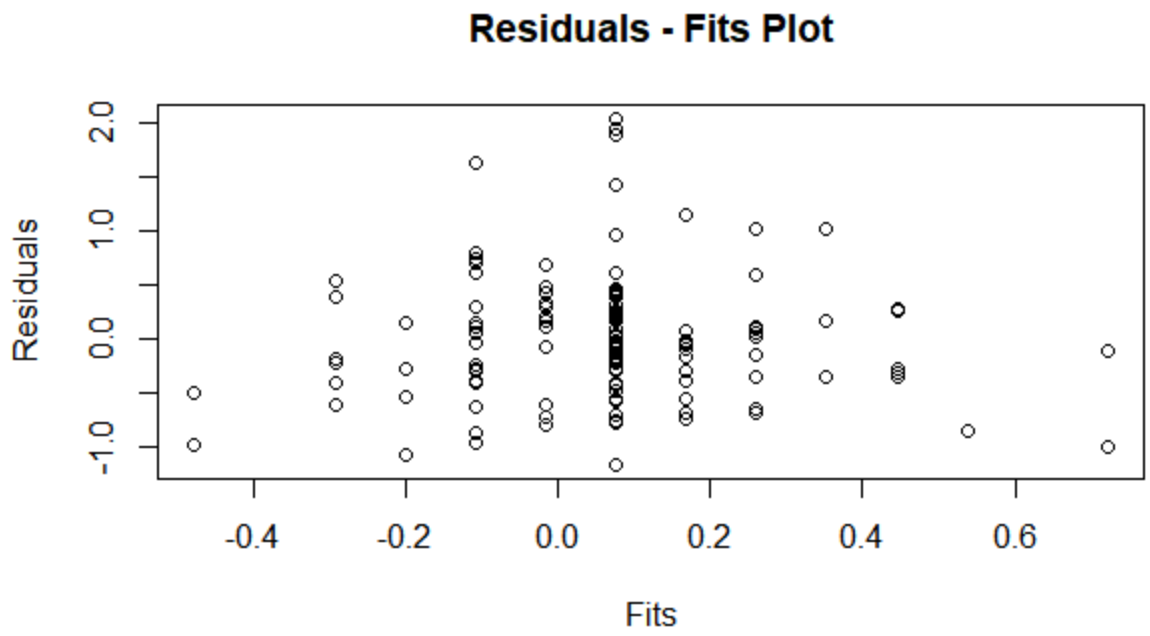
Question 5: Diagnostic checking

1. Produce the residuals and fitted values:

```
> myfits <- fitted(regression2)
> myresids <- residuals(regression2)
```

Graph residuals vs.fits:

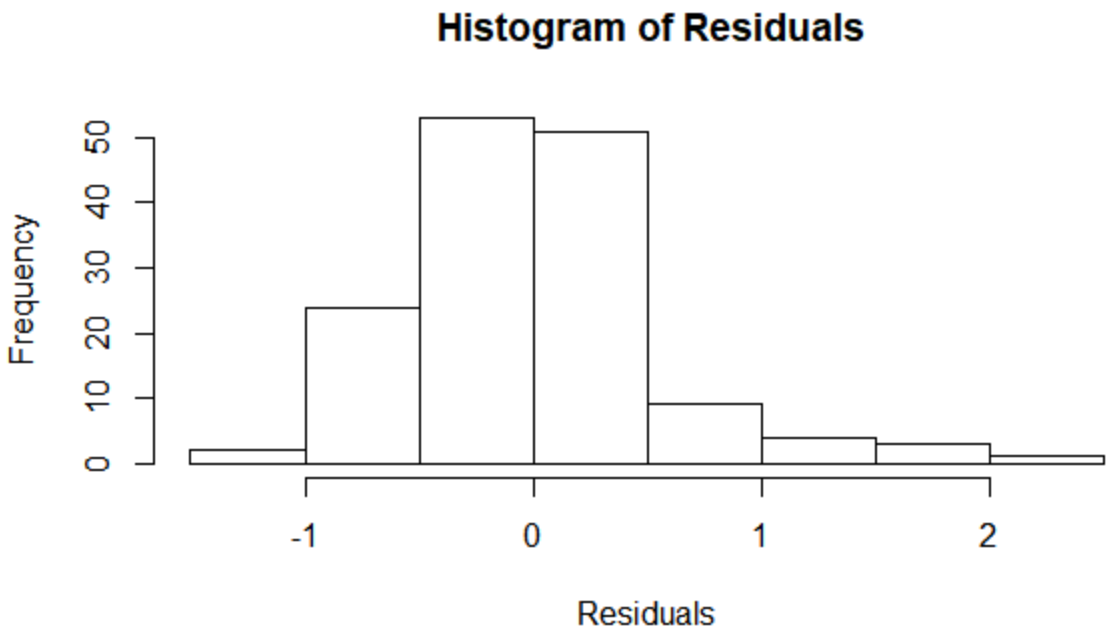
```
> plot(myfits, myresids,main = "Residuals - Fits Plot", xlab = "Fits", ylab = "Residuals")
```



An association can be seen between the fitted values and the residuals: as the Y values get closer to 0, they become more variable. Therefore, the variances are not equal. Assumption 5 does not hold.

Graph a histogram of residuals:

```
> hist(myresids, main = " Histogram of Residuals", xlab="Residuals")
```



Since the histogram shows a normal distribution, assumption 3 of regression holds.