9. Yes, respondents who completed more years of studies **generally** earn more income.

d$RINCOME\_DOLLAR = incomeInDollar[d$RINCOME]

meanIncomeByEduc <- aggregate(RINCOME\_DOLLAR ~ EDUC, d, mean)

meanIncomeByEduc

# EDUC RINCOME\_DOLLAR

# 1 0 25000.00

# 2 1 14500.00

# 3 2 12500.00

# 4 3 12500.00

# 5 4 14600.00

# 6 5 23333.33

# 7 6 14100.00

# 8 7 15500.00

# 9 8 12552.63

# 10 9 12408.16

# 11 10 15285.71

# 12 11 14611.65

# 13 12 17975.57

# 14 13 18449.58

# 15 14 19401.52

# 16 15 17781.25

# 17 16 21538.32

# 18 17 20900.83

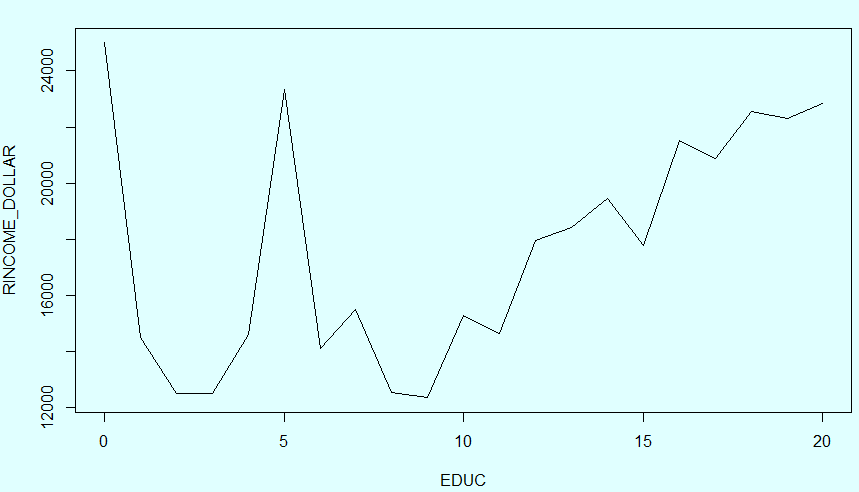
# 19 18 22582.35

# 20 19 22337.84

# 21 20 22856.00

plot(meanIncomeByEduc, type="l")

Plotting average of respondents’ income against his/her number of years in education produces below plot:



It seems that there are very high spikes at 0 and 5 years.

table(d$EDUC)

# 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

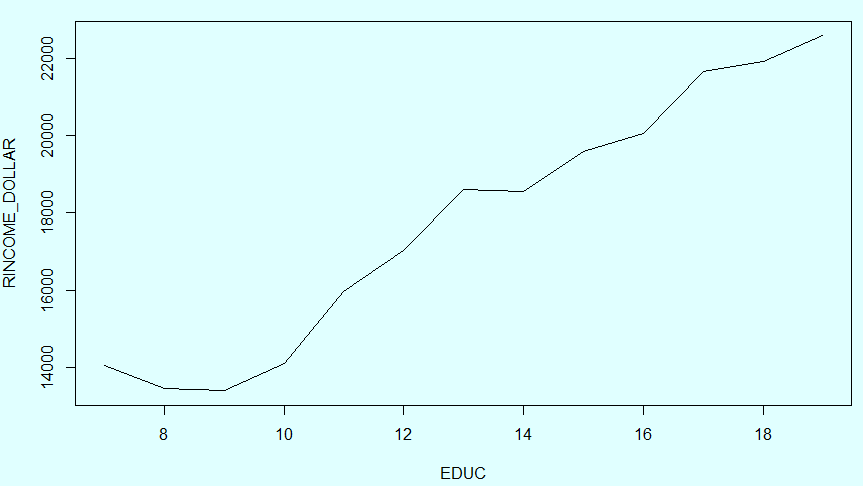
# 1 2 2 2 5 3 30 6 38 49 42 103 696 238 396 128 535 121 170 74 125

There is only 1 sample for person with 0 year education, and 3 sample for person with 5 year education. Furthermore, the number of sample below 6 years are quite small (up to 5 sample for each), and may not be enough to infer the trend from 0 to 5 years. Formal education less than 6 years may have little influence in deducing income. Those without even completing primary education might still be successful, but his/her success could be because of factors other than education.

If we exclude those below 6 years of education, and use moving average of 3 elements (EDUC=7 includes average from 6 to 8 years), it produces below plot:

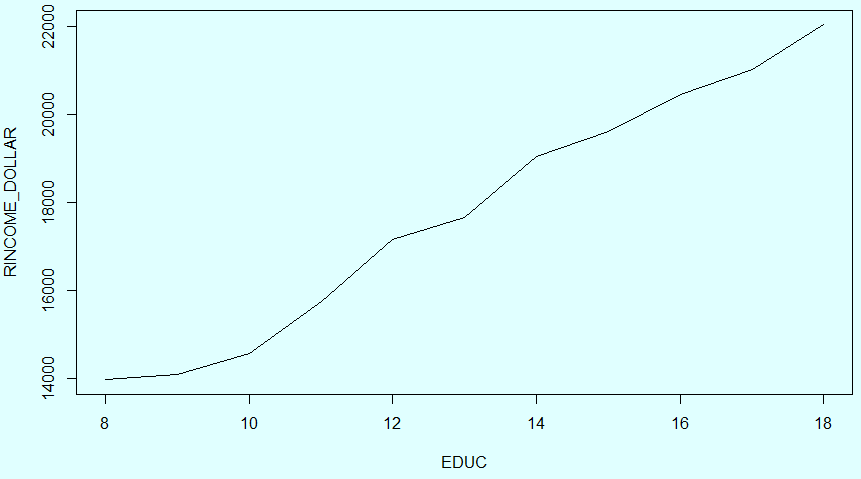
library(zoo)

plot(rollmean(meanIncomeByEduc %>% filter(EDUC > 5), 3), type="l")



With moving average of 5 elements, the trend is clearer that those who completed more years of studies generally earn more income.

plot(rollmean(meanIncomeByEduc %>% filter(EDUC > 5), 5), type="l")

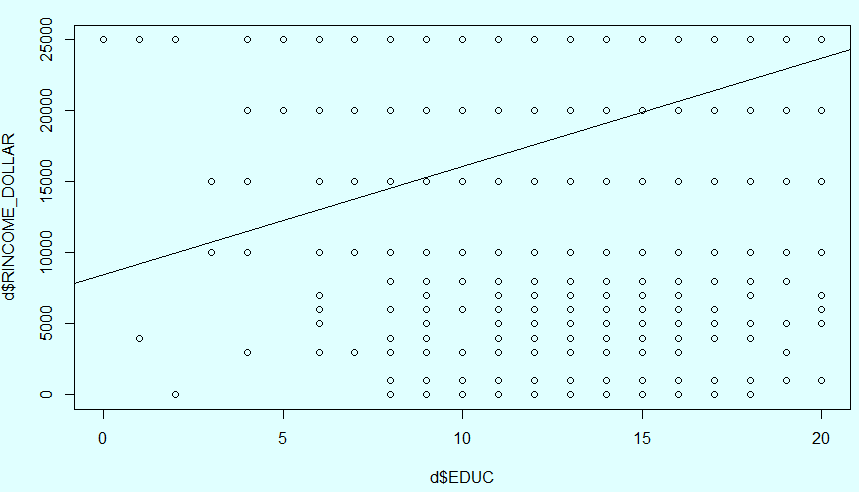


Extra experimental plots:

***Linear regression model***

plot(d$EDUC, d$RINCOME\_DOLLAR)

abline(lm(RINCOME\_DOLLAR ~ EDUC, data=d))



Observation:

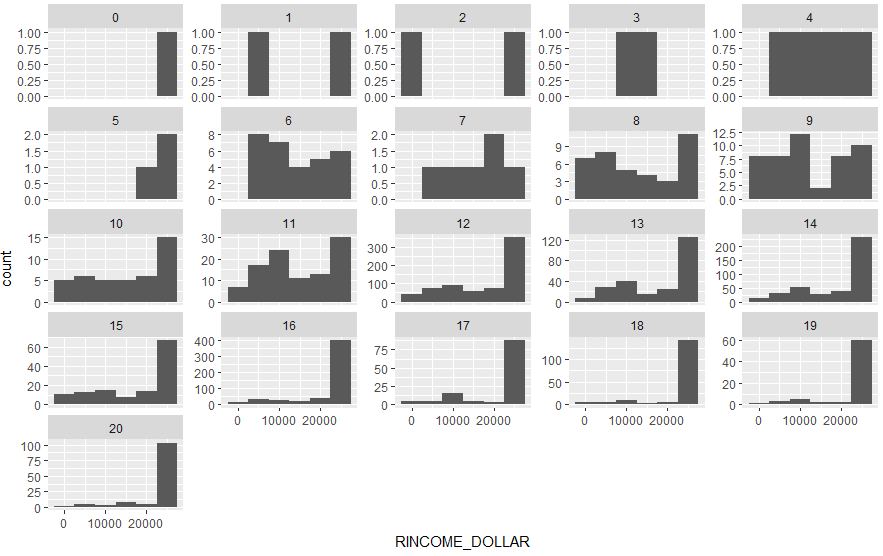
* The trend line is increasing, income is proportional to years of education.

***Histograms with facet***

ggplot(d, aes(x = RINCOME\_DOLLAR)) +

geom\_histogram(binwidth=5000) +

facet\_wrap(~ EDUC, scales="free\_y")



Observations:

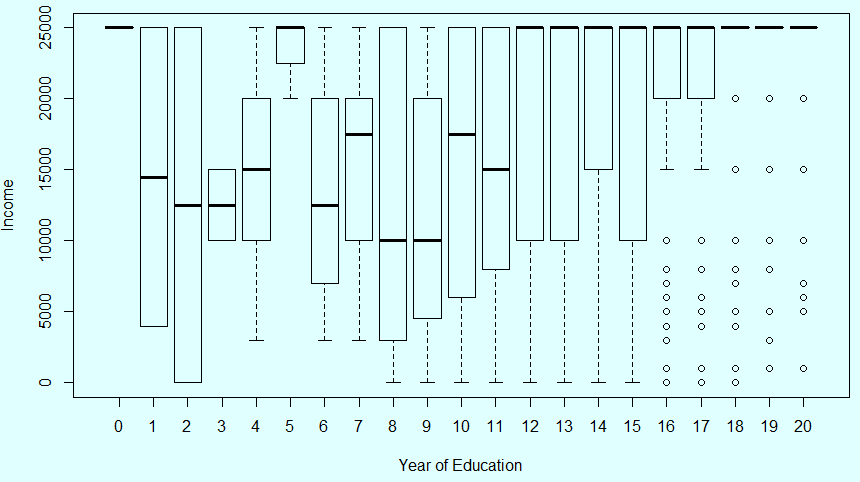
* The higher number of education years, more people will be at the last/highest income range.
* The shape of histograms from 12 years of education started to look similar
* For those with 16 years or more (completed bachelor degree), the proportion of people earning less than $25000 become very small

***Boxplot***

boxplot(d$RINCOME\_DOLLAR ~ d$EDUC,

xlab="Year of Education",

ylab="Income")



Observations:

* After 12 years of education, the median is already at the highest of $25000. More than 50% have earning more than $25000
* After 16 years of education, those earning less than $15000 are outliers
* After 18 years of education, those earning less than $25000 are outliers