

# Data Analytics with R

Yi Zhou

5/10/2021

## 1 Preface

This document is a brief demonstration of my data analytics techniques using R. For non-technical audiences, I've added some explanations for some of the statistical and programming terms. In each section, its questions and findings are placed at the beginning.

Since this is not an oral presentation, I've explained my thoughts and findings in detail to avoid confusion. Please let me know if you have any questions about this demonstration. I'd be happy to explain more.

## 2 Introduction

**Question:** What role does age play in the relationship between one's income and one's satisfaction level with his/her financial condition?

**Main finding:** More income would indicate more satisfaction with one's financial situation. This effect gets larger as age increases. The number of children also matters in this relationship for middle-aged and very old people. My result is significant only when we are studying fully-employed people.

As we all know, if you are doing a business or selling a product, it is crucial to know your customers' income levels. Information about customers' income levels can help companies in various ways, including targeting their markets, improving their services, establishing marketing plans, and pricing. The research paper <https://www.sciencedirect.com/science/article/pii/S0140673613624174> is also an example that demonstrates the importance of people's income information.

In addition to income information, knowing if the customers are satisfied with their financial situation is also essential. For example, customers with different levels of confidence in their financial conditions would most likely be interested in different types of financial products.

Usually, income is a good predictor of people's confidence/satisfaction level of their financial conditions. Many studies have shown that more income would indicate a higher satisfaction level with one's financial condition. However, not every researcher has considered the effect of age on this relationship. With this extra information, companies can better target their customer groups and further improve their marketing plans. In this project, I am going to study the role of age in the relationship between one's income and one's satisfaction level with his/her financial condition.

## 3 Methods and materials

```
library(foreign)
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.0.4
```

```
library(ggplot2)
```

```
gss = read.dta("GSS2016.dta")
gss = data.table(gss)
#gss
gss_subset = gss[, .(sex = as.numeric(sex == "male"),
degree = as.numeric(degree),
work = as.numeric(wrkstat),
marital = as.numeric(marital),
age,
age_decade = as.factor(floor(age / 10)),
race = as.numeric(race == "white"),
hispanic = as.numeric(hispanic != "hispanic"),
attend = as.numeric(attend), ## code as ordinal
educ,
income = as.numeric(income16), ## code as ordinal
srcbelt = as.numeric(srcbelt == "other rural"),
happy = as.numeric(happy),
satfin = as.numeric(satfin),
childs
)]
## remove rows that are all NA
gss_subset = na.omit(gss_subset)
#head(gss_subset)
```

To explore this topic, we need a dataset that contains information about subjects' age, income level, and satisfaction level with his/her financial condition. Our dataset also needs to include control variables, such as gender, mental health, education level, marital status, and so on. In addition, to ensure the internal validity of our study, we need to make sure that our dataset accounts for confounding variables. (In statistics, a confounder is a variable that influences both the dependent variable and independent variable, causing a spurious association.)

After careful consideration, I decided to use a subset of the **GSS 2016 Dataset**. The General Social Survey (GSS) dataset monitors societal change and studies the growing complexity of American society. It contains more than 5000 variables with data on a wide range of subjects. The GSS Data Explorer, from NORC at the University of Chicago, can be found here: <https://gssdataexplorer.norc.org/>

Let's take a look at some important variables:

*satfin(response)*: Satisfaction with financial situation

Questions asked to the subjects in this dataset: A. We are interested in how people are getting along financially these days. So far as you and your family are concerned, would you say that you are pretty well satisfied with your present financial situation, more or less satisfied, or not satisfied at all?

A value of 1 means "satisfied." A value of 2 means "more or less satisfied." A value of 3 means "not at all satisfied."

*work(named wrkstat in the original dataset)*: Labor force status

Questions asked to the subjects in this dataset: Last week were you working full time, part time, going to school, keeping house, or what?

A value of 1 means working full time. A value of 2 means working part time. A value of 3 means not working at the time(temp). A value of 4 means unemployed. A value of 5 means retired. A value of 6 means in school. A value of 7 means keeping house.

*age*: age

*age\_decade*: decade of age. For example, a value of 5 means 50-60 years old.

*income*:(named *income16* in the original dataset) Total family income

Please refer to this link for detailed information <https://gssdataexplorer.norc.umd.edu/variables/6167/vshow> A value of 1 indicates an annual family income of under 1000 dollars. A value of 26 indicates an annual family income of over 170000 dollars. Basically, a higher value indicates a higher annual family income.

*childs*: Number of children

Other variables are control variables.

I have fitted 4 linear models in this project. The first one will use `gss_subset` as the dataset. Others will use subsets of `gss_subset` based on different values of the variable `work`. In my models, I used `satfin` as the response variable; `age`, `childs`, `work`, `income` as response variables (some of these four will also serve as control variables in some of my models); others as control variables. Interactions between variables will be included. I will use `ggplots` to visualize my results. I've printed 6 rows of my dataset `gss_subset`. I've also printed some distributions of the important variables.

```
head(gss_subset)
```

```
##      sex degree work marital age age_decade race hispanic attend educ income
## 1:    1     4     1      1  47           4     1         1      1    16     26
## 2:    1     2     1      5  61           6     1         1      1    12     19
## 3:    1     4     5      1  72           7     1         1      8    16     21
## 4:    0     2     2      1  43           4     1         1      7    12     26
## 5:    0     5     2      1  55           5     1         1      1    18     26
## 6:    0     3     7      1  53           5     1         1      1    14     20
##      srcbelt happy satfin childs
## 1:         0     2      1      3
## 2:         0     2      3      0
## 3:         0     1     2      2
## 4:         0     2     1      4
## 5:         0     1     1      2
## 6:         0     1     2      2
```

```
table(gss_subset$work)
```

```
##
##      1     2     3     4     5     6     7     8
## 1233  314   56  105  493   65  239   66
```

```
table(gss_subset$satfin)
```

```
##
##      1     2     3
## 726 1143  702
```

```
table(gss_subset$childs)
```

```
##
##    0    1    2    3    4    5    6    7    8
## 713 423 663 417 193  79  45  19  19
```

```
table(gss_subset$income)
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20
##  34   35   23   12   19   15   22   47   88   69   63   65   94  101  115  128  126  218  206  256
##  21   22   23   24   25   26
## 212 179 125  96  61 162
```

```
table(gss_subset$age_decade)
```

```
##
##    1    2    3    4    5    6    7    8
##   30  399  472  415  516  416  208  115
```

## 4 Results

### 4.1 Analysis 1

**Question:** What role does age play in the relationship between one's income and one's satisfaction level of his/her financial condition?

**Finding:** More income would indicate more satisfaction with one's financial situation. This effect gets larger as age increases.

This analysis wants to answer the question: What role does age play in the relationship between one's income and one's satisfaction level with his/her financial condition?

**Model:**

```
fit1 = lm(satfin ~ work + sex + marital + income + degree + educ + age + childs +
          happy + srcbelt + attend + race + hispanic + age*income, data = gss_subset)
summary(fit1)
```

```
##
## Call:
## lm(formula = satfin ~ work + sex + marital + income + degree +
##     educ + age + childs + happy + srcbelt + attend + race + hispanic +
##     age * income, data = gss_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93728 -0.49932  0.02426  0.49223  1.63124
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  1.9622776  0.6897625   2.845 0.004478 **
## work        -0.0136039  0.0065856  -2.066 0.038957 *
## sex         -0.0909466  0.0269917  -3.369 0.000764 ***
## marital     -0.0024835  0.0098251  -0.253 0.800461
## income      -0.0070448  0.0068924  -1.022 0.306820
## degree      -0.0785446  0.0219892  -3.572 0.000361 ***
## educ        0.0165388  0.0091160   1.814 0.069754 .
## age         0.0022217  0.0023761   0.935 0.349867
## child       0.0259241  0.0094248   2.751 0.005990 **
## happy       0.2456541  0.0215883  11.379 < 2e-16 ***
## srcbelt     -0.0113013  0.0428840  -0.264 0.792162
## attend      -0.0063553  0.0049657  -1.280 0.200716
## race        -0.0644576  0.0321557  -2.005 0.045117 *
## hispanic    0.1155170  0.6721767   0.172 0.863565
## income:age  -0.0005082  0.0001336  -3.803 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6702 on 2556 degrees of freedom
## Multiple R-squared:  0.1959, Adjusted R-squared:  0.1915
## F-statistic: 44.47 on 14 and 2556 DF,  p-value: < 2.2e-16
```

A quick note to explain how we interpret the summary above (for non-technical audiences): The summary contains information about the model we fitted above. In the coefficients section, we can determine if a variable is significant by looking at its p-value, specifically the  $\text{Pr}( > |t| )$  column. In null-hypothesis significance testing, the p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct. The null hypothesis is that the observed difference is due to chance alone. Typically, a p-value of 5% or less is a good cut-off point, meaning that the variables with p-values less than 0.05 are significant. For example, the variable work has a p-value of 0.038957, which means that if the variable work does not have any influence on our response variable, satfin, the possibility that we obtain test results at least as extreme as the result actually observed is less than 3.9%. Therefore, we conclude a relationship between the variables work and satfin. The work variable is statistically significant in our model.

More information about the interpretation of the output can be found here: <https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R>

Now back to our analysis. I've fitted a model with satfin as the response and other variables as the predictors and control variables. I've also included the interaction term income\*age in my model. Our interaction term is significant. This interaction term will help us directly answer the research question by showing how income interact with age when predicting one's satisfaction level of his/her financial condition. Also, note that work and child are significant, we will explore them in future analysis. Now let's visualize our model and see what we can find.

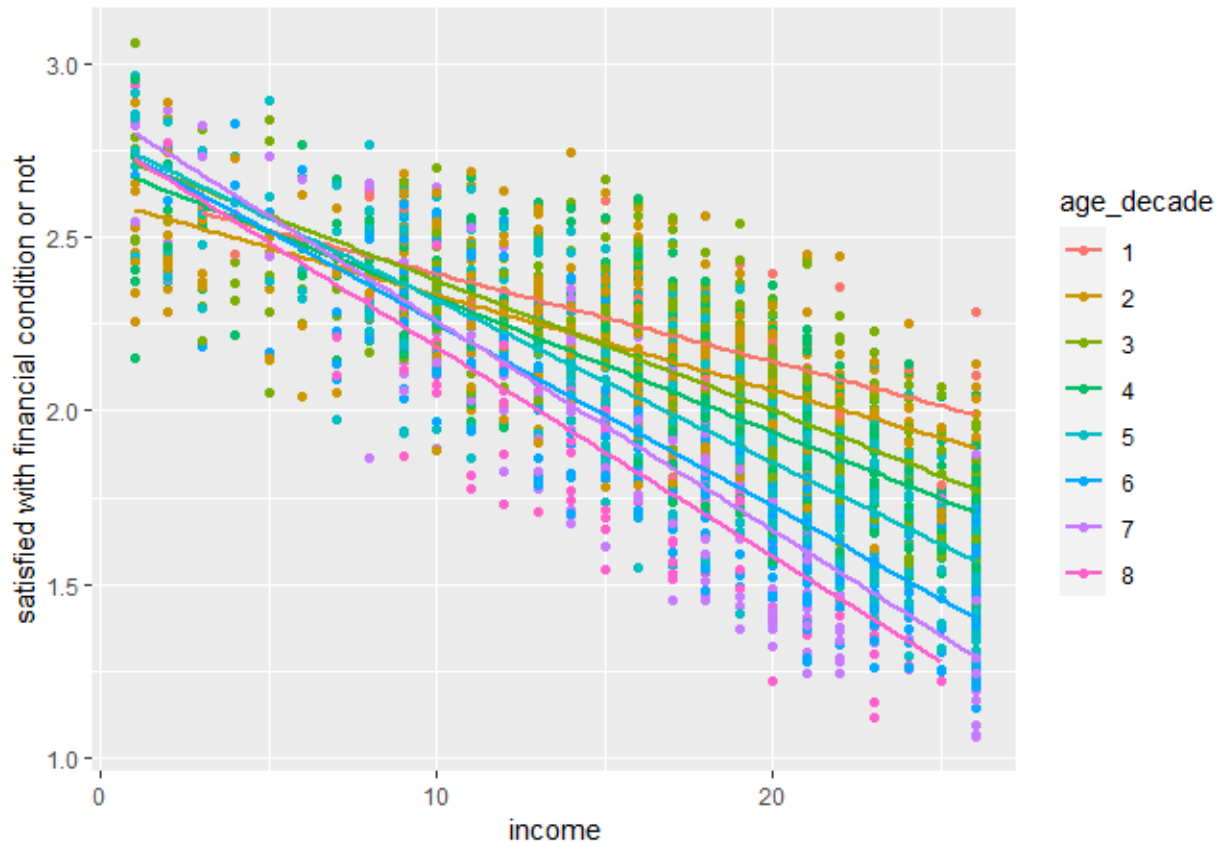
### Visualization:

```
viz_1 = gss_subset[, .(income,
satfin = predict(fit1,
newdata = gss_subset,
type = "response"),
age_decade)]

ggplot(data = viz_1) +
aes(x = income, y = satfin, color = age_decade) +
geom_point() +
```

```
geom_smooth(method = "lm", se = FALSE) +
ylab("satisfied with financial condition or not") +
xlab("income")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



Note that a higher value in satfin indicates less satisfaction. As you can see from the plot, More income would indicate more satisfaction with one's financial situation. Also, we notice that there is a difference in slopes for different age decades. The absolute value in slope increases as age increases, meaning that the effect of income on satfin gets larger as age increases. In analysis 2 and analysis 3, I will explore why this difference in slope(effect) exists.

## 4.2 Analysis 2

**Question:** Could the difference in effects be explained by different employment situations?

**Finding:** The difference in effect for different age groups is not explained by different employment statuses. Our analysis is only significant when the subject is working full-time.

In analysis one, we discovered that more income would indicate more satisfaction with one's financial situation. This effect gets larger as age increases. One possible explanation is that this relation is influenced by different employment statuses. For example, it is possible that some old people are more easily satisfied with their financial condition because they are retired. However, given the same family income, an old person that is still working full time may not be satisfied with their financial condition. Therefore, it is possible that the differences in the effects for different age groups could be caused by changes in employment status. The

effect gets larger because as age increases, more people have become retired, and thus have become satisfied with their financial condition. In this analysis, I want to test this explanation.

I want to analyze the following question: Could the difference in effects be explained by different employment situations?

This analysis will help me better understand my analysis 1 results by testing a possible explanation. It could also potentially help me modify my dataset so that my analysis 1 stays significant.

## Models:

```
subset_fulltime <- gss_subset[ which(work==1),]
#subset_fulltime
fit2 = lm(satfin ~ work + sex + marital + income + degree + educ + age + childs+
          happy + srcbelt + attend + race + hispanic + age*income,data = subset_fulltime)
summary(fit2)
```

```
##
## Call:
## lm(formula = satfin ~ work + sex + marital + income + degree +
##      educ + age + childs + happy + srcbelt + attend + race + hispanic +
##      age * income, data = subset_fulltime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75124 -0.44969  0.00588  0.42649  1.54425
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.2343075  0.3058260   4.036 5.78e-05 ***
## work                NA           NA      NA      NA
## sex              -0.1070338  0.0377063  -2.839 0.004606 **
## marital           0.0141018  0.0130046   1.084 0.278415
## income           0.0227975  0.0138602   1.645 0.100266
## degree          -0.0776354  0.0303800  -2.555 0.010725 *
## educ             0.0187897  0.0130087   1.444 0.148886
## age              0.0240576  0.0064562   3.726 0.000203 ***
## childs           0.0428169  0.0144130   2.971 0.003029 **
## happy            0.2234677  0.0324099   6.895 8.62e-12 ***
## srcbelt         -0.0632732  0.0617451  -1.025 0.305685
## attend           0.0075137  0.0071894   1.045 0.296176
## race            -0.0233556  0.0442477  -0.528 0.597708
## hispanic          NA           NA      NA      NA
## income:age       -0.0014770  0.0003271  -4.515 6.95e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6478 on 1220 degrees of freedom
## Multiple R-squared:  0.1844, Adjusted R-squared:  0.1764
## F-statistic: 22.99 on 12 and 1220 DF,  p-value: < 2.2e-16
```

I've subsetting my dataset so that only fully-employed people are involved. Then, I fitted the same model using the new dataset to see if the result is different. As you can see from the summary above, the interaction term is still significant.

```
subset_retired <- gss_subset[ which(work==5),]
#subset_retired
fitretire = lm(satfin ~ work + sex + marital + income + degree + educ + age + childs+
               happy + srcbelt + attend + race + hispanic + age*income,data = subset_retired)
summary(fitretire)
```

```
##
## Call:
## lm(formula = satfin ~ work + sex + marital + income + degree +
##      educ + age + childs + happy + srcbelt + attend + race + hispanic +
##      age * income, data = subset_retired)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.79394 -0.48771 -0.01607  0.46134  1.70890
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.489e+00  6.852e-01   5.092 5.09e-07 ***
## work          NA          NA      NA      NA
## sex          -8.531e-02  6.098e-02  -1.399  0.16246
## marital      -1.863e-02  2.830e-02  -0.658  0.51056
## income       -3.770e-02  3.547e-02  -1.063  0.28842
## degree       -1.021e-01  4.922e-02  -2.074  0.03858 *
## educ          6.793e-03  2.082e-02   0.326  0.74431
## age          -1.144e-02  8.635e-03  -1.325  0.18583
## childs        1.629e-03  1.830e-02   0.089  0.92911
## happy         1.472e-01  4.651e-02   3.165  0.00165 **
## srcbelt       3.588e-02  8.751e-02   0.410  0.68202
## attend       -1.727e-02  1.025e-02  -1.685  0.09258 .
## race         -2.142e-01  8.441e-02  -2.538  0.01147 *
## hispanic      NA          NA      NA      NA
## income:age   -3.549e-05  5.016e-04  -0.071  0.94362
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6507 on 480 degrees of freedom
## Multiple R-squared:  0.2513, Adjusted R-squared:  0.2326
## F-statistic: 13.43 on 12 and 480 DF,  p-value: < 2.2e-16
```

This is another subset of my original dataset(gss\_subset). This dataset only includes people that have retired. Now we fit the same model using this dataset. As you can see from the summary, the interaction term is no longer significant in this model. For other values in the work variable, I don't think there are enough observations for the models/analysis to be significant. Therefore, I conclude that our result in analysis 1 is only significant when we are considering fully-employed people.

### Visualizations:

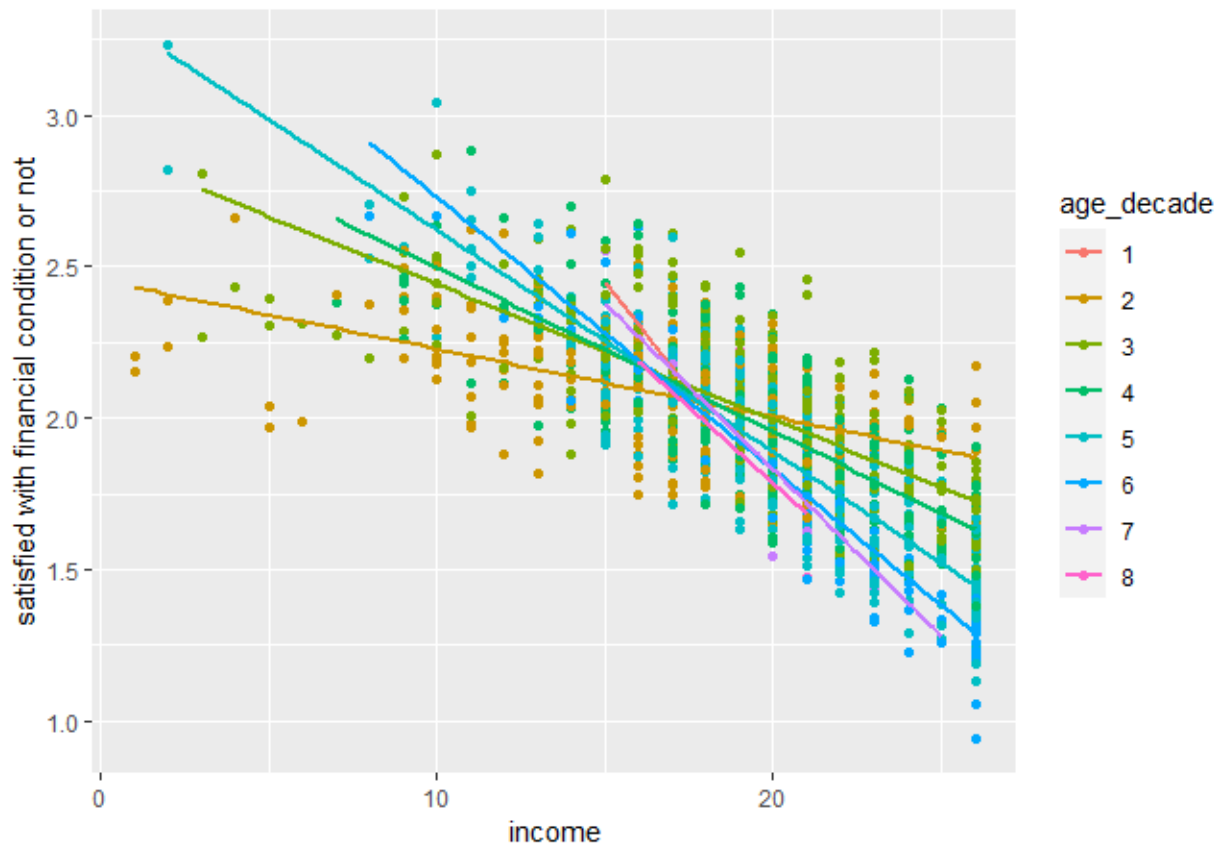
```
viz_2 = subset_fulltime[, .(income,
satfin = predict(fit2,
newdata = subset_fulltime,
type = "response"),
age_decade)]
```



```
## Warning in predict.lm(fit2, newdata = subset_fulltime, type = "response"):  
## prediction from a rank-deficient fit may be misleading
```

```
ggplot(data = viz_2) +  
  aes(x = income, y = satfin, color = age_decade) +  
  geom_point() +  
  geom_smooth(method = "lm", se = FALSE) +  
  ylab("satisfied with financial condition or not") +  
  xlab("income")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



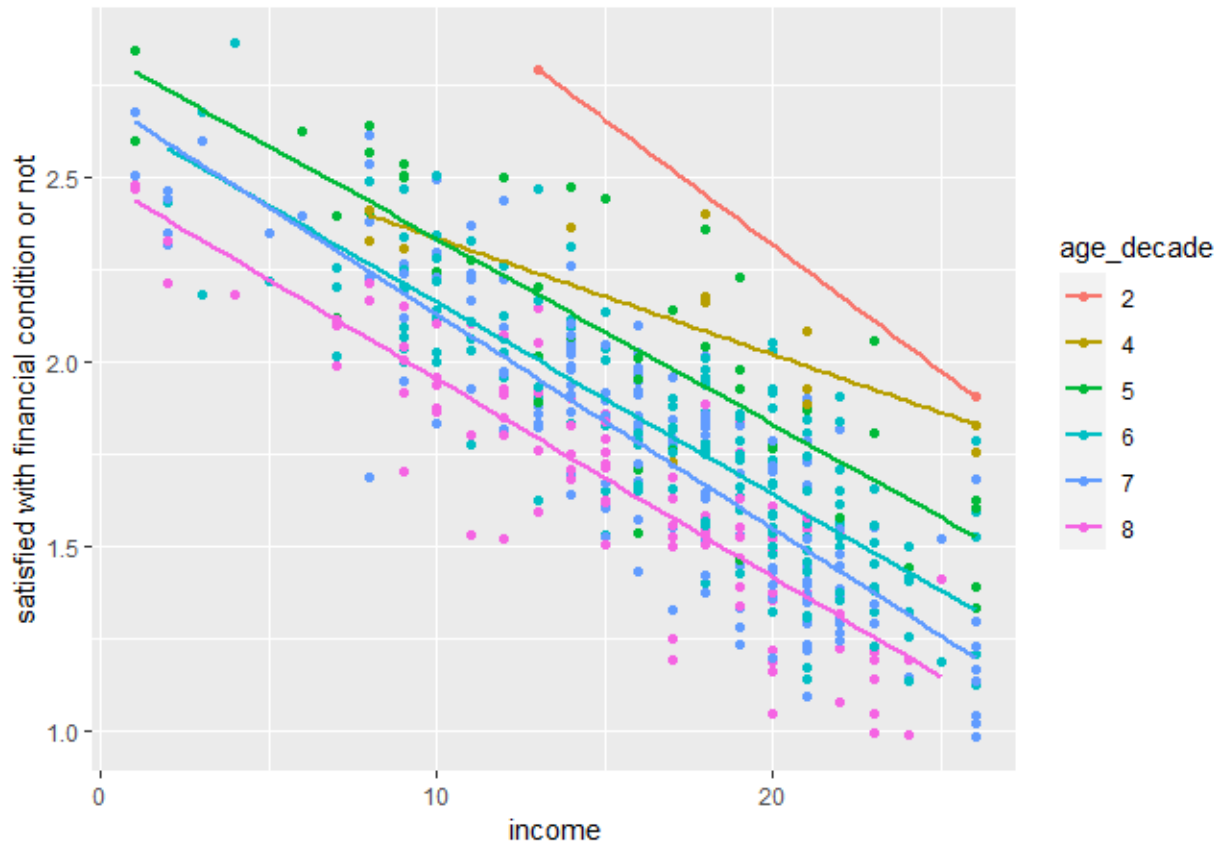
As you can see, the plot has changed a little bit compared to the one in our first analysis. However, the general trend stayed the same. More income would indicate more satisfaction with one's financial situation. This effect gets larger as age increases. Therefore, we reject the explanation I proposed in this analysis. We conclude that the difference in effect for different age groups is not explained by different employment statuses. Our analysis is only significant when the subject is working full-time.

```
viz_retire = subset_retired[, .(income,  
  satfin = predict(fitretire,  
    newdata = subset_retired,  
    type = "response"),  
  age_decade)]
```

```
## Warning in predict.lm(fitretire, newdata = subset_retired, type = "response"):  
## prediction from a rank-deficient fit may be misleading
```

```
ggplot(data = viz_retire) +
  aes(x = income, y = satfin, color = age_decade) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ylab("satisfied with financial condition or not") +
  xlab("income")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



The interaction term for the retired model is not significant. I just want to use this graph to show that the interaction term is indeed not significant in this model. We are not drawing conclusions from this graph.

### 4.3 Analysis 3

**Question:** Does interaction between the number of children and age/income play an important role in the relationship between income and the satisfaction level with one's financial condition?

**Answer:** The interaction between the number of children and age does play an important role in predicting the satisfaction level with one's financial condition. As the number of children increases, it becomes harder for middle-aged and very old people to be satisfied with their financial condition.

We have shown that the difference in effect is not explained by different employment statuses. Now we want to propose an alternative explanation for the result of analysis 1. So technically, analysis 3 is more related to analysis 1 and it is not closely related to analysis 2. I hope this is fine. I know it's not exactly what the

rubric asked for, but since we rejected the explanation in analysis 2, it is natural that we propose another explanation. Analysis 2 did contribute a lot by finding a proper subset in which our analysis is significant.

The variable `childs` have been significant in both models in the previous two analyses. It is possible that the number of children also plays an important role in our analysis. Its interaction with age or income could also be significant. For example, it is possible that having children will result in a higher standard for one's financial situation. When people grow old, that standard may have dropped to normal because their children are now able to take care of themselves. In this analysis, we will test this explanation.

We want to answer the following question in this analysis: Does interaction between the number of children and age/income play an important role in our analysis?

**Model:**

```
fit3 = lm(satfin ~ work + sex + marital + income + degree + educ + age + childs +
          happy + srcbelt + attend + race + hispanic + childs*age + age*income + childs*income, data =
summary(fit3)
```

```
##
## Call:
## lm(formula = satfin ~ work + sex + marital + income + degree +
##      educ + age + childs + happy + srcbelt + attend + race + hispanic +
##      childs * age + age * income + childs * income, data = subset_fulltime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76298 -0.46963  0.01744  0.42756  1.52212
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.1189553   0.3083260   3.629 0.000296 ***
## work              NA              NA      NA      NA
## sex            -0.1004995   0.0377434  -2.663 0.007854 **
## marital         0.0168914   0.0130372   1.296 0.195348
## income          0.0203401   0.0138666   1.467 0.142676
## degree         -0.0786781   0.0303415  -2.593 0.009626 **
## educ           0.0186174   0.0130029   1.432 0.152461
## age            0.0277782   0.0069491   3.997 6.79e-05 ***
## childs         0.1487822   0.0706512   2.106 0.035420 *
## happy          0.2185090   0.0324018   6.744 2.38e-11 ***
## srcbelt        -0.0631439   0.0616403  -1.024 0.305853
## attend         0.0078026   0.0071775   1.087 0.277211
## race           -0.0181580   0.0442006  -0.411 0.681284
## hispanic        NA              NA      NA      NA
## age:childs     -0.0029641   0.0011216  -2.643 0.008327 **
## income:age     -0.0014645   0.0003438  -4.259 2.21e-05 ***
## income:childs  0.0012977   0.0032458   0.400 0.689373
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6465 on 1218 degrees of freedom
## Multiple R-squared:  0.1891, Adjusted R-squared:  0.1798
## F-statistic: 20.29 on 14 and 1218 DF, p-value: < 2.2e-16
```

I've included three interaction terms in this model, `age:childs`, `age:income`, and `income:childs`. The first two

interaction terms are significant, meaning that the result in analysis 1 holds and the proposed explanation in analysis 3 could be true. Let's take a look at the plots.

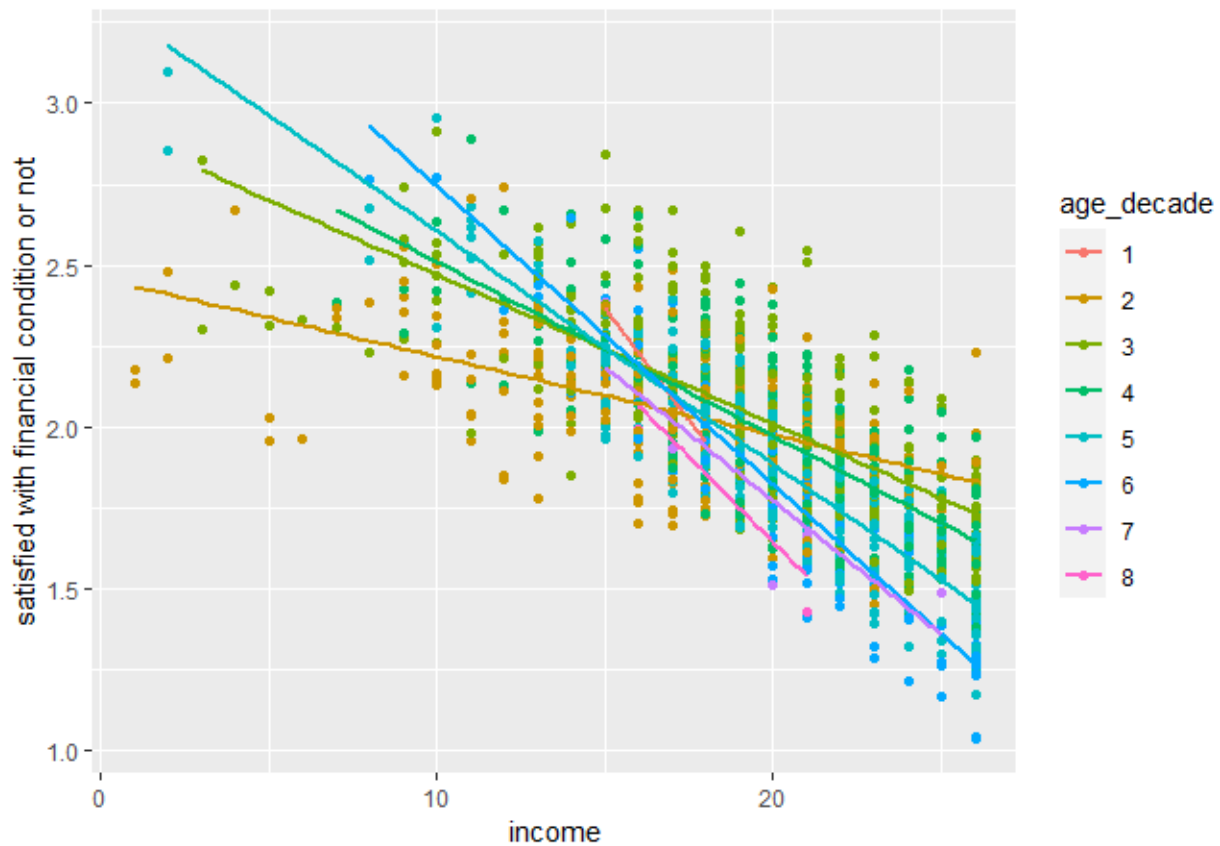
### Visualizations:

```
viz_full1 = subset_fulltime[, .(income,
satfin = predict(fit3,
newdata = subset_fulltime,
type = "response"),
age_decade)]
```

```
## Warning in predict.lm(fit3, newdata = subset_fulltime, type = "response"):  
## prediction from a rank-deficient fit may be misleading
```

```
ggplot(data = viz_full1) +  
  aes(x = income, y = satfin, color = age_decade) +  
  geom_point() +  
  geom_smooth(method = "lm", se = FALSE) +  
  ylab("satisfied with financial condition or not") +  
  xlab("income")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



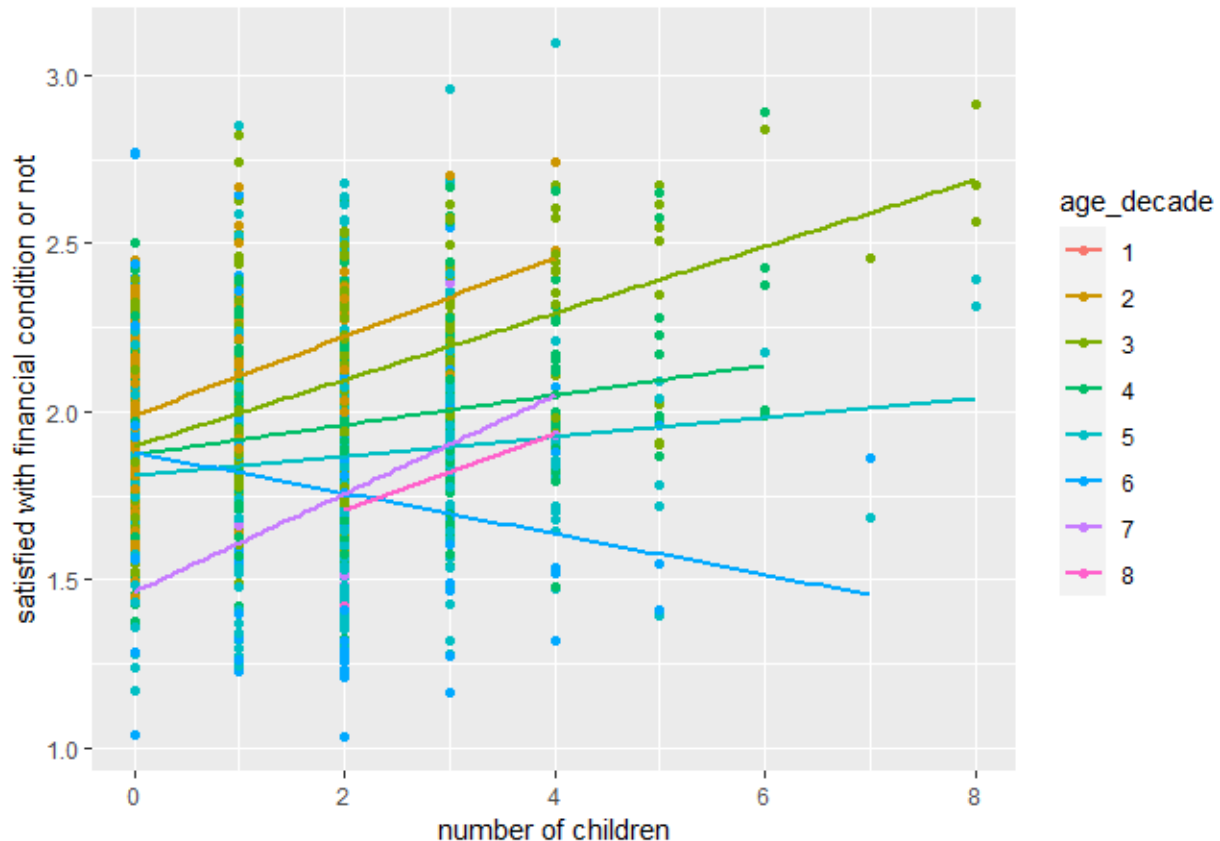
The plot about the interaction term  $\text{income} * \text{age}$  yields similar results as before, which is good.

```
viz_full2 = subset_fulltime[, .(childs,
satfin = predict(fit3,
newdata = subset_fulltime,
type = "response"),
age_decade)]
```

```
## Warning in predict.lm(fit3, newdata = subset_fulltime, type = "response"):
## prediction from a rank-deficient fit may be misleading
```

```
ggplot(data = viz_full2) +
aes(x = childs, y = satfin, color = age_decade) +
geom_point() +
geom_smooth(method = "lm", se = FALSE) +
ylab("satisfied with financial condition or not") +
xlab("number of children")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



This is the plot about the interaction term  $childs \times age$ . As you can see from the plot, as the number of children increases, it becomes harder for middle-aged and very old people to be satisfied with their financial condition while the number of children does not matter that much for people between 40-69 years old. (Absolute value of the slopes are smaller. The slope even becomes negative for people in their 60s.) Therefore, we conclude that the interaction between the number of children and age does play an important role in predicting the satisfaction level of one's financial condition. As the number of children increases, it becomes harder (need more income) for middle-aged (or younger) and very old people to be satisfied with their financial condition.

## 5 Discussion

**Conclusion: More income would indicate more satisfaction with one's financial situation. This effect gets larger as age increases. The number of children also matters in this relationship for middle-aged and very old people. My result is significant only when we are studying fully-employed people.**

Combining the three analyses, my answer to the initial research question would be: More income would indicate more satisfaction with one's financial situation. This effect gets larger as age increases. The number of children also matters in this relationship for middle-aged and very old people. My result is significant only when we are studying fully-employed people.

Let me summarize how I've arrived at this conclusion. To explore the role of age in the relationship between one's income and one's satisfaction level with his/her financial situation, I fitted a model that uses income to predict satisfaction and included an interaction term  $\text{age} \times \text{income}$ . From the interaction term, I found out that the positive effect of income on satisfaction gets larger as age increases. Then, I proposed a possible explanation: Is it possible that this effect that age has results from changes in employment statuses? After testing, I found out that this explanation does not stand. I also found out that our results are only significant when we are considering fully-employed people. Thus, I subsetted the dataset. After the explanation was rejected, I proposed an alternative explanation for analysis 3: Is it possible that this effect that age has results from interactions between age and number of children? After fitting a model with more interaction terms, I found out that the term  $\text{age}:\text{childs}$  is indeed significant, meaning that our alternative explanation could stand.

I would say that my answer could be useful for companies and salesmen. I found out that the positive effect of income on satisfaction gets larger as age increases. This indicates that among people with high family income, a higher age would probably indicate a higher probability of being satisfied with their financial condition. This should be a piece of useful information in business. For example, if I was a salesman that sold financial products, I might want to offer a product that helps secure my customer's money (low risk, relatively-low return) if I notice that my customer is rich and old. Also, it might also be helpful to know if my customer has any children. Middle-aged people with kids might prefer products with lower risks because they have responsibilities to their children. (I'm sure real salesmen know much better than I do. I'm just providing a situation where this study could be useful.)

## 6 References:

<https://www.sciencedirect.com/science/article/pii/S0140673613624174>

<https://gssdataexplorer.norc.umd.edu/>