1 Introduction

1.1 Objectives

Is it true that our pay rises in tandem with our years of experience? Do our age, sex, or even drinking habits affect how much money we will earn? While we can answer this question based on our fundamental knowledge of the labour market, we will utilize simple data from a website, “OkCupid”, driven technique to confirm the reality. In this study, we aim to explore what variables in life have an impact on an individual's income. To do so, we want to examine whether or not any combination of our “age” and/or additional dummy variables in the database serves as a more accurate and/or precise predictive model for predicting income level than the most accurate and precise base parameters, while either controlling for predicted influential variables (such as job or education) or assuming excluded variables are independent.

The **research question** of this study include:

- Which variables in the dataset effectively influence OkCupid users’ income level
- Which model including different variables fits or predicts users’ income better

1.2 Background

All variables were collected from the users of OkCupid, which is an online dating, friendship, and social networking website and app headquartered in the United States that uses multiple-choice questions to connect users. Anyone over the age of 18 may join the site, and all members are able to connect with one another through private messages or an instant messaging "chat" function. OkCupid was also the first major dating service to provide free unlimited texting. As of September 2010, OkCupid claimed 3.5 million active users. As a relied data collection site, we intend to use the dataset provided by this website to predict income level.

2 Data Collection and Summaries

2.1 Dataset

The data consists of the public profiles of 59,946 OkCupid users who lived within 25 miles of San Francisco, had active profiles in the 2010s, were online the previous year, and had at least one picture in their profile.

2.2 Motivation

- The variables that affect income level
- The better model fit the income level

2.3 The source of the data

Data for OkCupid Profile Data for Introductory Statistics and Data Science Courses, Journal of Statistics Education July 2015, Volume 23, Number 2. The original manuscript was subsequently revised in 2021.

(https://github.com/rudeboybert/JSE_OkCupid/blob/master/profiles_revised.csv.zip)

2.4 Data Description

The dataset used for the analysis is an extraction of the 2010 data in the article "OkCupid Data for Introductory Statistics and Data Science Courses" by Albert Y. Kim and Adriana Escobedo-Land (Kim and Escobedo-Land 2015). Variables include typical user information (such as sex, sexual orientation, age, and ethnicity) and lifestyle variables (such as diet, drinking habits, smoking habits).

In order to address the aforementioned two questions more precisely, we made some adjustments from the original dataset. For simplifies, we removed the variable including "Body type", "Diet", "Education", "Ethnicity", "Job", "Orientation", "Pets", "Religion", "Sign" and all the blanks in the original dataset. Besides, we combine some factors,

- Drinks: “desperately” & “very often” → “lots”; “often” & “socially” & “rarely” → “sometimes”
- Offspring: “does not have a kid” & “doesn’t want kids”; “and/but might want them”, “wants them”, and “doesn’t want any” → “doesn’t have kids”; “has a kid" & “has kids” & “doesn’t want more” → “has kids”
- Smokes: “yes” & “sometimes” & “when drinking” & “trying to quit” → “yes”
- Status: “single” & “seeing someone” → “available”; “married” & “in an open relationship” → “taken”

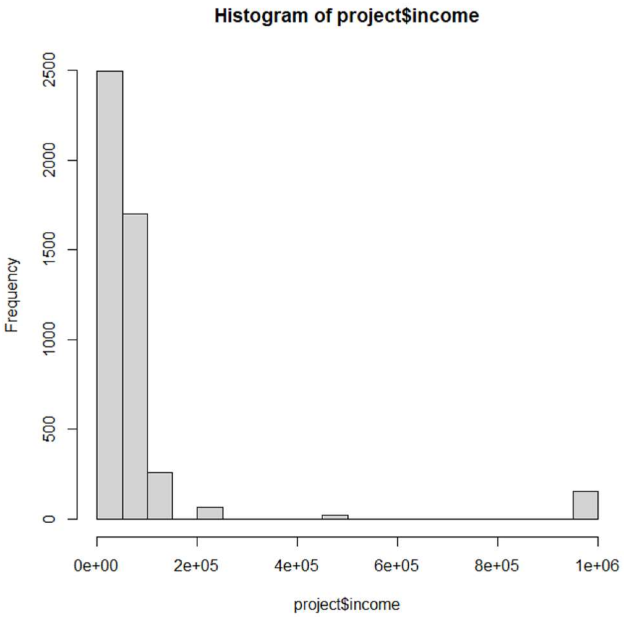
Finalized data variables:

age	User’s age in years (from 17 years old to 70 years old)
drinks	User’s frequency of drinking (Factor levels: not at all, sometimes, lots)
drugs	User’s frequency of taking drugs (Factor levels: never, sometimes, often)
income	User’s income level
offspring	Has offspring (Factor levels: doesn’t have kids, has kids)
sex	User’s gender (Factor levels: male, female)
smokes	Smoker (Factor levels: no, yes)
status	User’s marital status (Factor levels: available, taken)

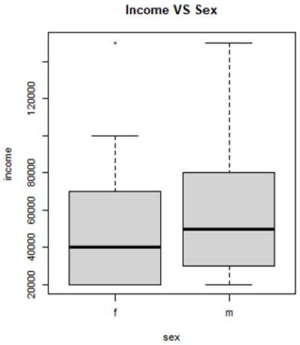
3 Interpretation and Analysis

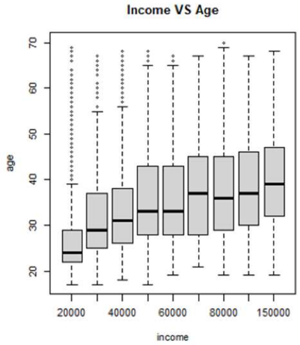
3.1 Discarding the outliers of response variable

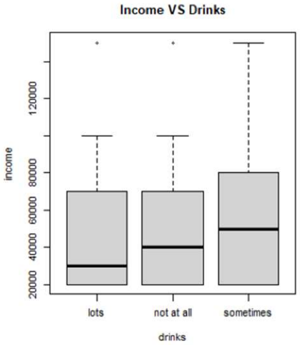
We set the variable “income” as the response variable. By drawing the histogram, we observe that income levels higher than 150000 are outliers and including them may not make our further linear regression model effective. Therefore, we filtered out the income values that are higher than 150000.

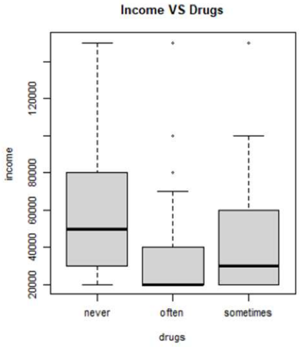


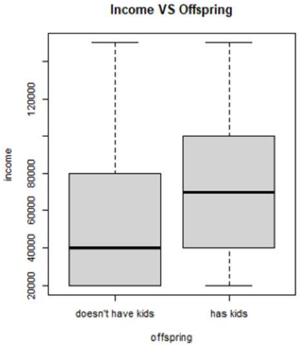
3.2 Visualization and Analysis between Variables

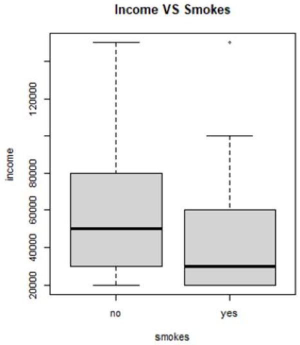


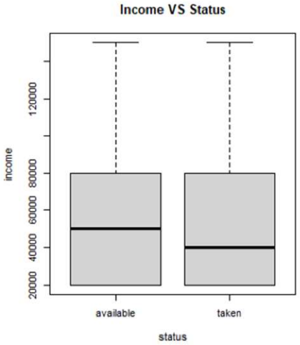












- **Income VS Sex**

First, we want to investigate the relationship between income and sex. After excluding the outliers, we can clearly see from the above figure (plot Income VS Sex) that the mean income level of male workers is significantly higher than that of female workers.

- **Income VS Age**

As can be seen from the second plot above, with the increase of income level, the age of workers corresponding to them is also higher. Older workers have more experience and earn more.

- **Income VS Drinks**

The frequency of drinking and income do not have obvious relationship, but workers who drink lots have a much lower mean income level than people who drink sometimes and never.

- **Income VS Drugs**

The relationship between income and drugs are significant: as the frequency workers take drugs increase, the corresponding level of income decrease.

- **Income VS Offspring**

According to the plot, individuals with kids seems to have higher income. It is reasonable as most people gestate to the next generation when they own adequate financial support.

- **Income VS Smokes**

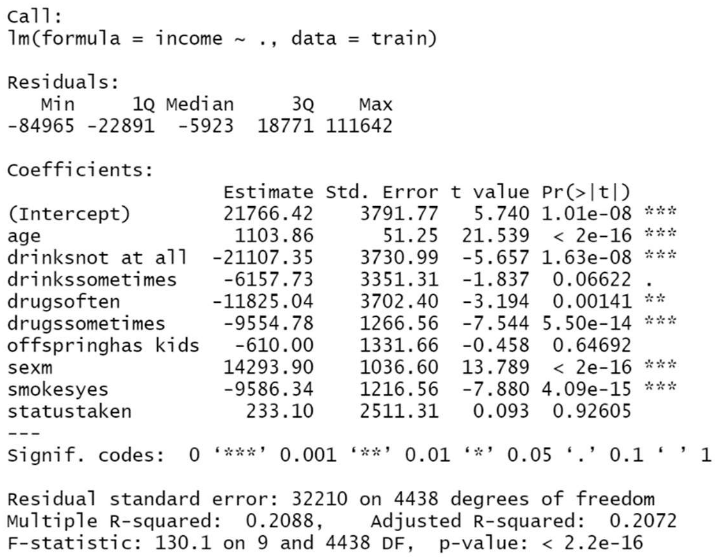
The "Income VS smokes" plot suggests smokers may have a lower average income and range.

- **Income VS Status**

There is no significant difference in income level between the single and group of married, while the married show a lower average earing.

3.3 Fit the model

We fit the “Full model” (all variables) at the beginning to observe which variables are significantly affect the income level in this dataset. As the right picture, the p-values of dummy variable “offspring” and “status” are much larger than 0.05 (assume the significant level is 5%). Hence, we choose the rest explanatory variables and refit the model, called “Model 1”.



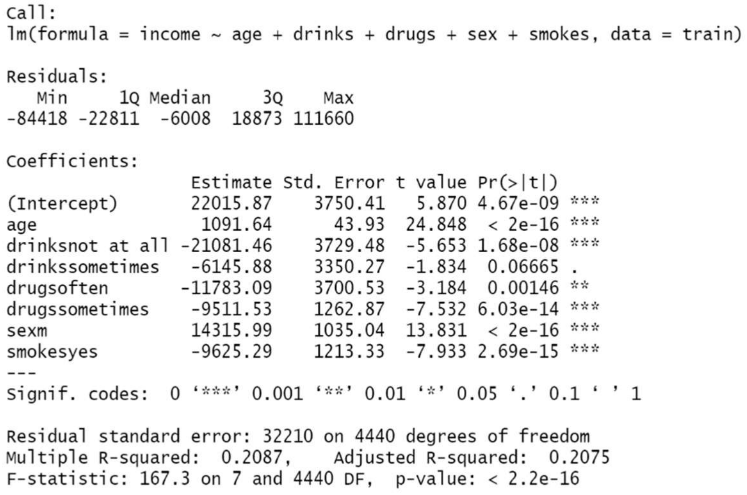
```
Call:
lm(formula = income ~ ., data = train)

Residuals:
   Min     1Q Median     3Q    Max 
-84965 -22891  -5923  18771 111642 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        21766.42    3791.77   5.740 1.01e-08 ***
age                 1103.86      51.25  21.539  < 2e-16 ***
drinksnot at all  -21107.35    3730.99  -5.657 1.63e-08 ***
drinkssometimes    -6157.73    3351.31  -1.837  0.06622 .  
drugsoften        -11825.04    3702.40  -3.194  0.00141 ** 
drugssometimes     -9554.78    1266.56  -7.544 5.50e-14 ***
offspringhas kids   -610.00    1331.66  -0.458  0.64692    
sexm               14293.90    1036.60  13.789  < 2e-16 ***
smokesyes          -9586.34    1216.56  -7.880 4.09e-15 ***
statustaken          233.10    2511.31   0.093  0.92605    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32210 on 4438 degrees of freedom
Multiple R-squared:  0.2088,	Adjusted R-squared:  0.2072 
F-statistic: 130.1 on 9 and 4438 DF,  p-value: < 2.2e-16
```

Full model



```
Call:
lm(formula = income ~ age + drinks + drugs + sex + smokes, data = train)

Residuals:
   Min     1Q Median     3Q    Max 
-84418 -22811  -6008  18873 111660 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       22015.87    3750.41   5.870 4.67e-09 ***
age                1091.64      43.93  24.848  < 2e-16 ***
drinksnot at all -21081.46    3729.48  -5.653 1.68e-08 ***
drinkssometimes   -6145.88    3350.27  -1.834  0.06665 .  
drugsoften       -11783.09    3700.53  -3.184  0.00146 ** 
drugssometimes    -9511.53    1262.87  -7.532 6.03e-14 ***
sexm              14315.99    1035.04  13.831  < 2e-16 ***
smokesyes         -9625.29    1213.33  -7.933 2.69e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32210 on 4440 degrees of freedom
Multiple R-squared:  0.2087,	Adjusted R-squared:  0.2075 
F-statistic: 167.3 on 7 and 4440 DF,  p-value: < 2.2e-16
```

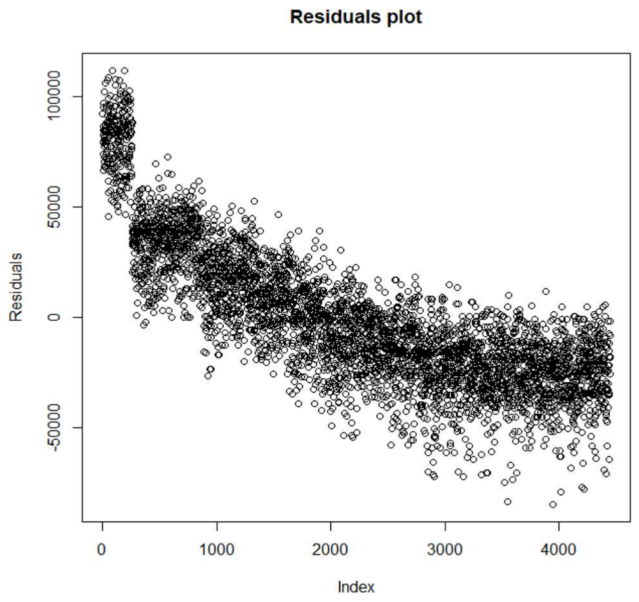
Model 1

We are curious about whether the assumption of linear regression is satisfied, thus we draw a residual plot to check the following assumptions of doing linear regression:

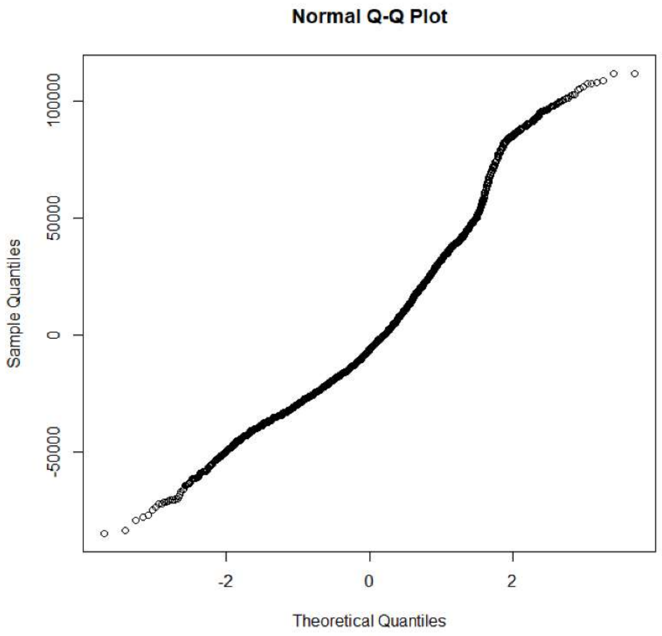
1) Whether there is a linear relationship between the response variable and the explanatory variables

2) Whether the observations are independent

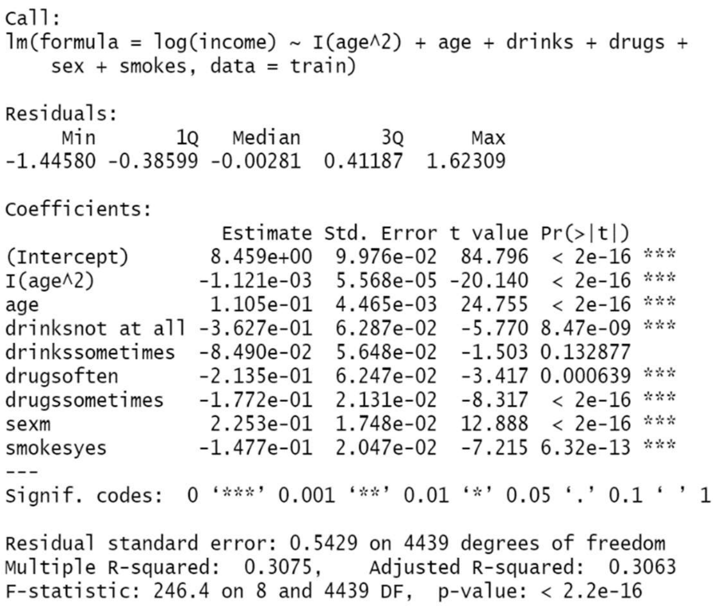
3) Whether variances of random errors are equal



By observing the residual plot, we found that although the dots were scattered, a non-linear pattern was shown. That means that our first assumptions are not satisfied, and as a result, the R-squared value will be small. Therefore, we need further mathematical transformation to the explanatory variables (such as adding a squared explanatory variable into the model) to get a better-fitted model and a higher adjusted R-squared value. There is a noticeable trend in the residual plot, as the residual values take positive with small fitted values and negative for large fitted values. This indicates a serial correlation, so the second assumption is unsatisfied, and we need to do further mathematical computations to the response variable, such as taking the log or the square root of income. In this residual plot, the longitudinal widths were almost the same, indicating equal variances of random errors, which proved that assumption 3 was satisfied.



Second, we draw a Q-Q plot to check the assumption of whether random errors are normally distributed. It shows that the points approximately fall on the line, which indicates this assumption is met.



```
Call:
lm(formula = log(income) ~ I(age^2) + age + drinks + drugs +
    sex + smokes, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.44580 -0.38599 -0.00281  0.41187  1.62309

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.459e+00  9.976e-02  84.796  < 2e-16 ***
I(age^2)         -1.121e-03  5.568e-05 -20.140  < 2e-16 ***
age               1.105e-01  4.465e-03  24.755  < 2e-16 ***
drinksnot at all -3.627e-01  6.287e-02  -5.770 8.47e-09 ***
drinkssometimes  -8.490e-02  5.648e-02  -1.503 0.132877
drugsoften       -2.135e-01  6.247e-02  -3.417 0.000639 ***
drugssometimes   -1.772e-01  2.131e-02  -8.317  < 2e-16 ***
sexm              2.253e-01  1.748e-02  12.888  < 2e-16 ***
smokesyes        -1.477e-01  2.047e-02  -7.215 6.32e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5429 on 4439 degrees of freedom
Multiple R-squared:  0.3075,	Adjusted R-squared:  0.3063
F-statistic: 246.4 on 8 and 4439 DF,  p-value: < 2.2e-16
```

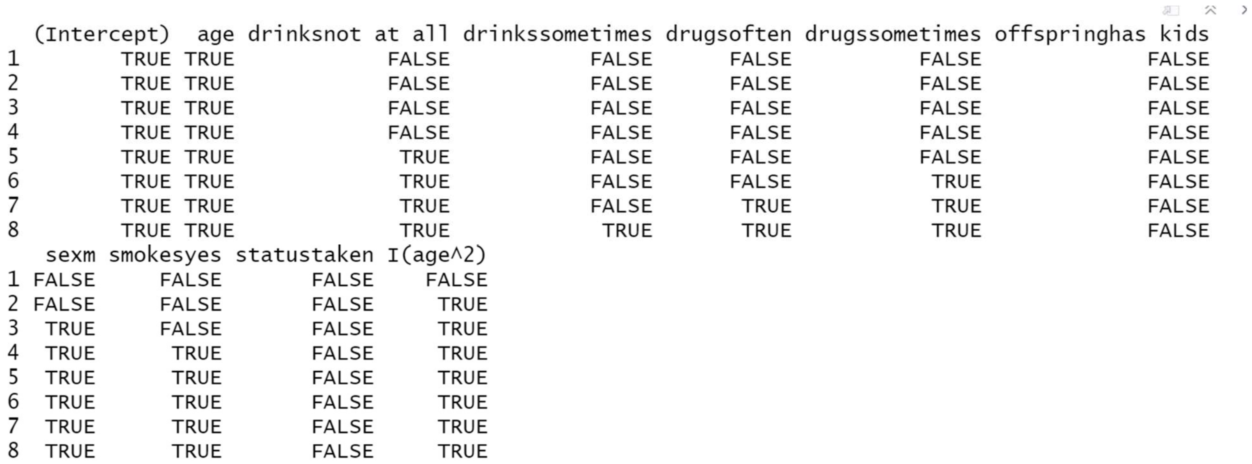
Model 2

Based on the above conditions, we first add [age^2] to the original model and then take a log of income [log(income)] to get model 2. Compared with other models, the adjusted R-squared value of this model was significantly higher, reaching 0.3063. Simultaneously, we also try to add all possible interaction terms, but the adjusted R^2 value of these models is far inferior to model 2. For other models, we obtained different R-squared values, such as 0.05259 for model 1.1, 0.05221 for model 1.2 and 0.1399 for model 1.3. Thus, we determined that model 2 is the best model we can get so far. The formula is:

income = 8.459 - 0.001121*(age^2) + 0.1105*age - 0.3627*drink_not_at_all - 0.0849*drink_sometimes

- 0.2135*drugs_often - 0.1772*drugs_sometimes + 0.2253*sex_male - 0.1477*smoke_yes

Next, we make the model selection through the regsubsets() function, which returns separate best models of all sizes up to 8. We found that the model obtained by our analysis and the eighth model in this output both use the same variables: age, sex, smokes, drinks, drugs. This conclusion suggests these five variables all significantly affect the income level, which also proves the effectiveness of our model 2 from another perspective.



```
  (Intercept)  age drinksnot at all drinkssometimes drugsoften drugssometimes offspringhas kids
1        TRUE TRUE            FALSE           FALSE      FALSE          FALSE             FALSE
2        TRUE TRUE            FALSE           FALSE      FALSE          FALSE             FALSE
3        TRUE TRUE            FALSE           FALSE      FALSE          FALSE             FALSE
4        TRUE TRUE            FALSE           FALSE      FALSE          FALSE             FALSE
5        TRUE TRUE             TRUE           FALSE      FALSE          FALSE             FALSE
6        TRUE TRUE             TRUE           FALSE      FALSE           TRUE             FALSE
7        TRUE TRUE             TRUE           FALSE       TRUE           TRUE             FALSE
8        TRUE TRUE             TRUE            TRUE       TRUE           TRUE             FALSE
   sexm smokesyes statustaken I(age^2)
1 FALSE     FALSE       FALSE    FALSE
2 FALSE     FALSE       FALSE     TRUE
3  TRUE     FALSE       FALSE     TRUE
4  TRUE      TRUE       FALSE     TRUE
5  TRUE      TRUE       FALSE     TRUE
6  TRUE      TRUE       FALSE     TRUE
7  TRUE      TRUE       FALSE     TRUE
8  TRUE      TRUE       FALSE     TRUE
```

4 Discussion and Conclusion

3.1 Discussion and Limitation

1. The response variable "income" is not a precise quantitative variable. Although income is used as a numerical variable in our analysis, it is rounded to several integers from 20K to 15000K with different intervals, which looks more like a categorical variable. This makes the scatter plot that is used to represent the relationship between age and income look strange. Moreover, this problem also makes the fitted model inaccurate.

2. We try to add interaction terms to see if the model fits better, such as age*drinks and age*drugs. It results in adjusted R^2 being smaller. So we decided not to join the interaction.

3. The residual plot of Model 2 still shows a negative linear pattern. Also, QQplot appears a slightly thin trail. Thus, we believe that the relationship between income and other variables in this dataset does not conform to linear regression.

4. The users with the extreme value income are omitted. The dataset doesn't have workers who earn less than 20K, while we filter out those who make higher than 150K in the model. Hence, our result is not suitable for predicting the income of workers whose income is lower than 20K or higher than 150K. Otherwise, we are at the risk of extrapolation.

5. There may be other factors that affect income. For simplicity, we ignore some variables included in the database, such as education and job type. These factors have a high potential to affect changes in income.

6. We assume that each individual entered into the database is independent, but this assumption may not hold.

7. Data may not be true. Since this is a website's message statistics for users, there is no mandatory requirement that the information to be filled in is true. This has a high probability of causing inaccurate information.

3.2 Conclusion

To conclude, the variables affecting OkCupid users’ income level are "age", "sex", "drinks", "smokes", and "drugs". After trying different transformations to the categorical and response variables, model 2 fits users’ income better.

However, the R^2 of model 2 is still too small and the residuals still follow some sort of trend, no matter what non-linear transformation we apply to the model. These evidences suggest we may explore the risk that regression may not be the most appropriate method to fit the income based on this dataset.