Midterm Project Report

Chuning Yuan

I.Introduction

The main goal of this project is to find a dataset, and propose an analysis that includes fitting a multilevel model. Based on the goal, the deliverable consisted two parts. First, through Exploratory Data Analysis, I can have an overview of this data set, and visualize the relationships between the different variables. Then we use the outcomes to determine which factors should be included in the analysis model. Second, we used models to conduct an analysis.

II.Data

The data for this project are extract from the dataset in R – Fair's Extramarital Affairs Data, it is a Cross-section data from a survey conducted by Psychology Today in 1969. The data frame containing 601 observations on 9 variables. Below is the data description in detail:

affairs(numeric): How often engaged in extramarital sexual intercourse during the past year? 0 = none, 1 = once, 2 = twice, 3 = 3 times, 7 = 4–10 times, 12 = monthly, 12 = weekly, 12 = daily.

gender(factor): male or female.

age(numeric): numeric variable coding age in years: 17.5 = under 20, 22 = 20–24, 27 = 25–29, 32 = 30–34, 37 = 35–39, 42 = 40–44, 47 = 45–49, 52 = 50–54, 57 = 55 or over.

yearsmarried(numeric): numeric variable coding number of years married: 0.125 = 3 months or less, 0.417 = 4–6 months, 0.75 = 6 months–1 year, 1.5 = 1–2 years, 4 = 3–5 years, 7 = 6–8 years, 10 = 9–11 years, 15 = 12 or more years.

children(factor): are there children in the marriage?

religiousness: numeric variable coding religiousness: 1 = anti, 2 = not at all, 3 = slightly, 4 = somewhat, 5 = very.

education: numeric variable coding level of education: 9 = grade school, 12 = high school graduate, 14 = some college, 16 = college graduate, 17 = some graduate work, 18 = master's degree, 20 = Ph.D., M.D., or other advanced degree.

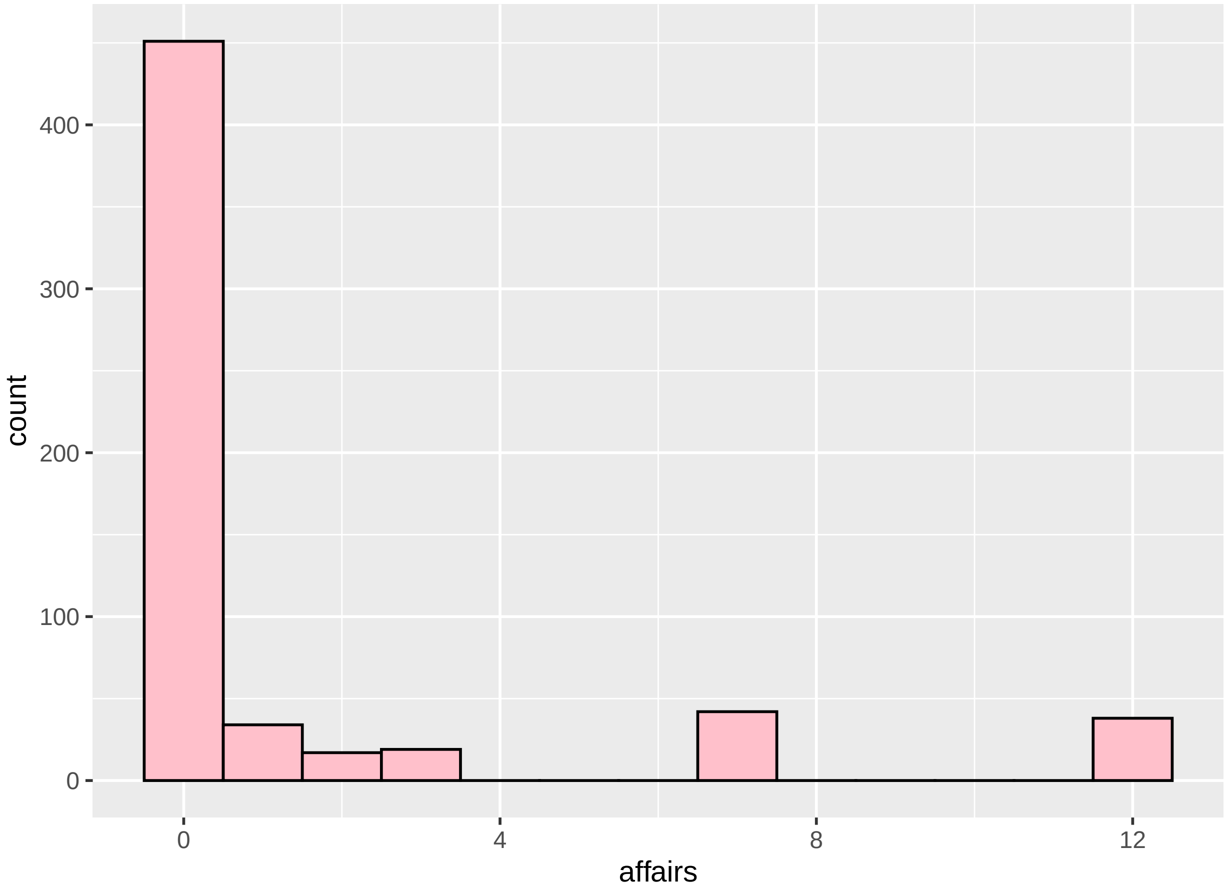
occupation: numeric variable coding occupation according to Hollingshead classification (reverse numbering).

rating: numeric variable coding self-rating of marriage: 1 = very unhappy, 2 = somewhat unhappy, 3 = average, 4 = happier than average, 5 = very happy.

III. EDA

1.What is the distributions of the number of affairs?

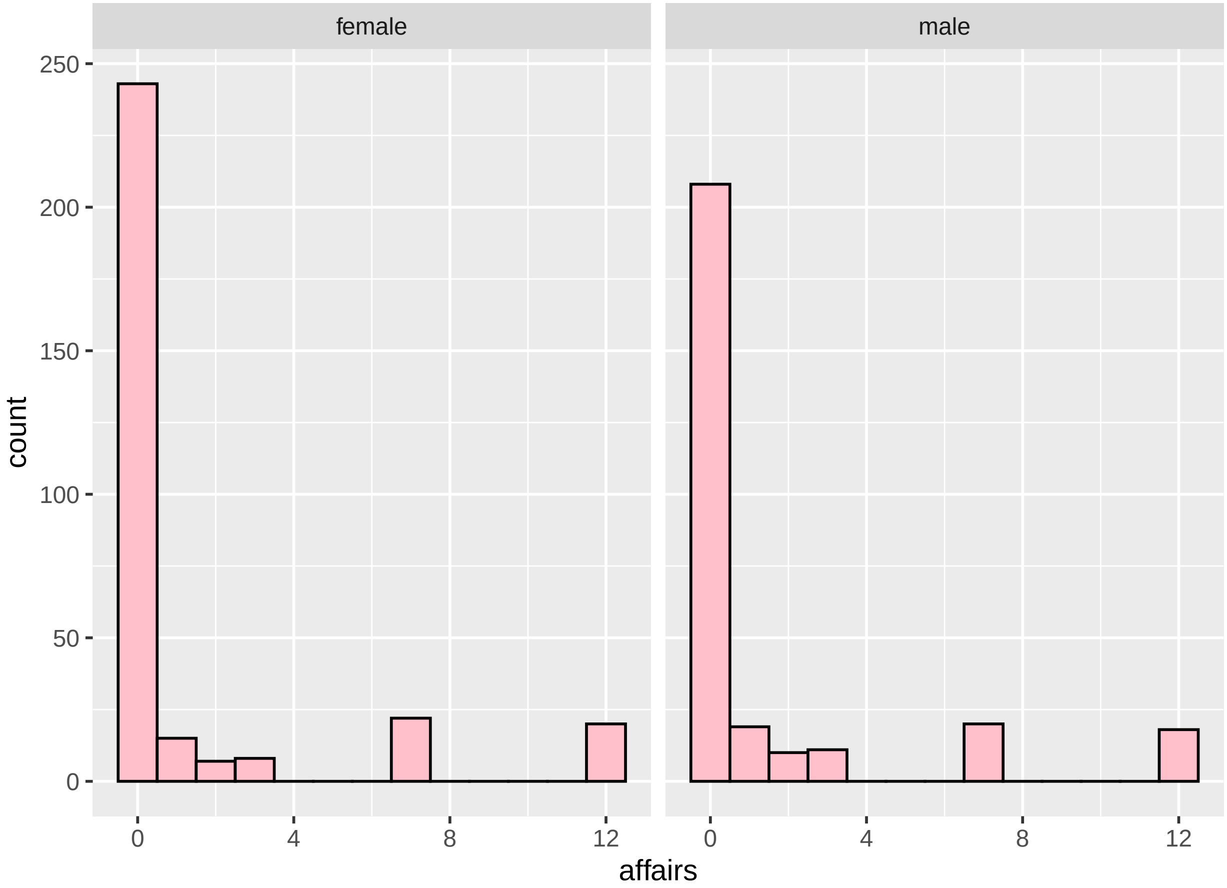
Figure 1. Distribution of affairs



First, we can plot to see the visualization for the variable *affairs*, which is the variable we consider as the outcome for the model. The scale of this variables is discrete integer. From Figure 1, we can see that the number of affairs can take only six non-continuous values, which due to the design of the survey. According to the data description, when the actually number of affairs is between 4 and 12, it will be recorded as 7, when it is greater than 12, it will be recorded as 12.

2. Distributions of number of affairs by gender

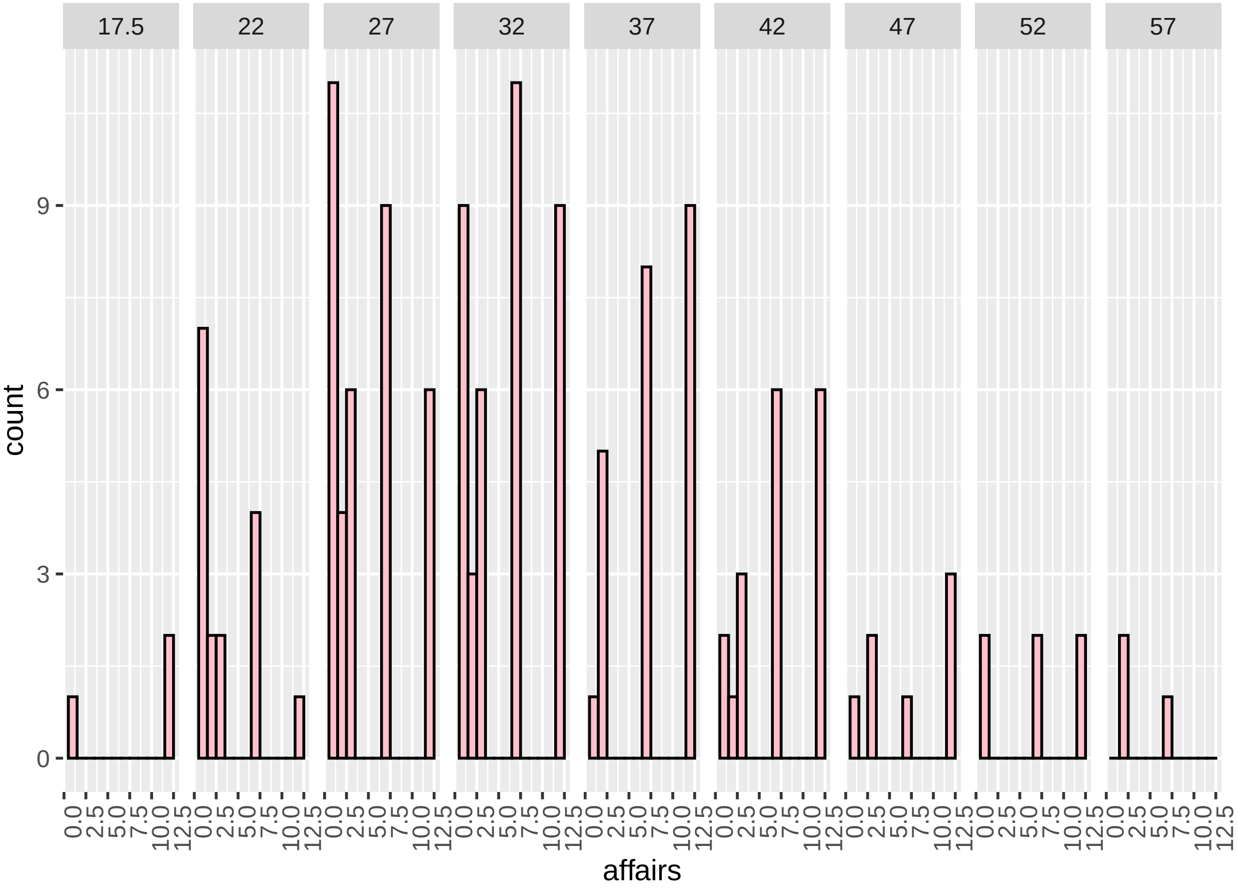
Figure 2. Distribution of affairs by gender



Then we want to check some other predictors will affect the number of affairs. Figure 2 shows the number of affairs distributed by the difference between male and female. From the plot, we can tell that there is not much difference. Therefore, it suggests that we don't need to take into account the difference of gender.

Distribution of number of affairs by age

Figure 4. Distribution of affairs by age



We may suspect that the number of affairs of people in different age group are correlated, below is the histagram of number of affairs between different age group. In each age group, most of the number of affairs are 0, so we filtered those observations and make sure that we can see the difference between each age groups. From Figure 4. we can observe that the younger the age of the person, the higher the number of affairs this person might conduct.

4. How are the variable correlated with each other?

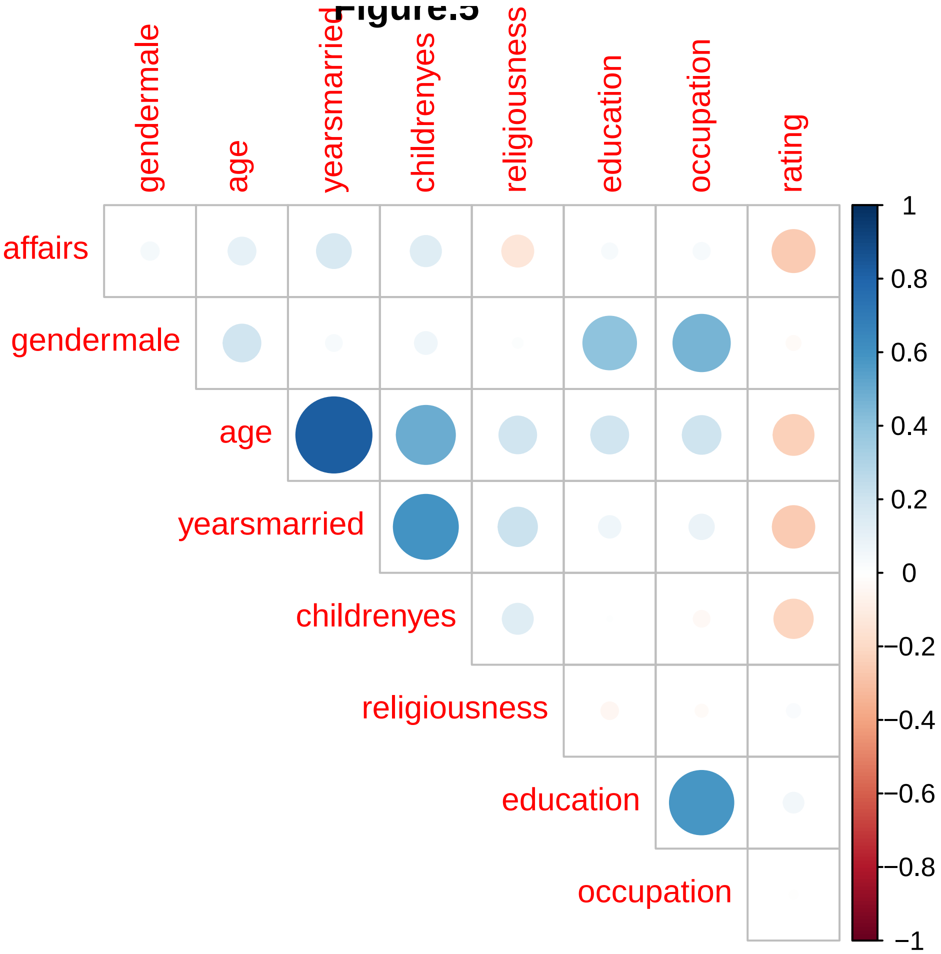


Figure 5. is the correlation plot of all variables, we can observe that number of affairs are not highly correlated with other variables. There are some variables are highly correlated, which can be explained by social context, for example the correlation between gender, education and occupation can be explained by gender discrimination, since the survey was conducted in 1960s. The age and years married of a person are strongly correlated is also very reasonable.

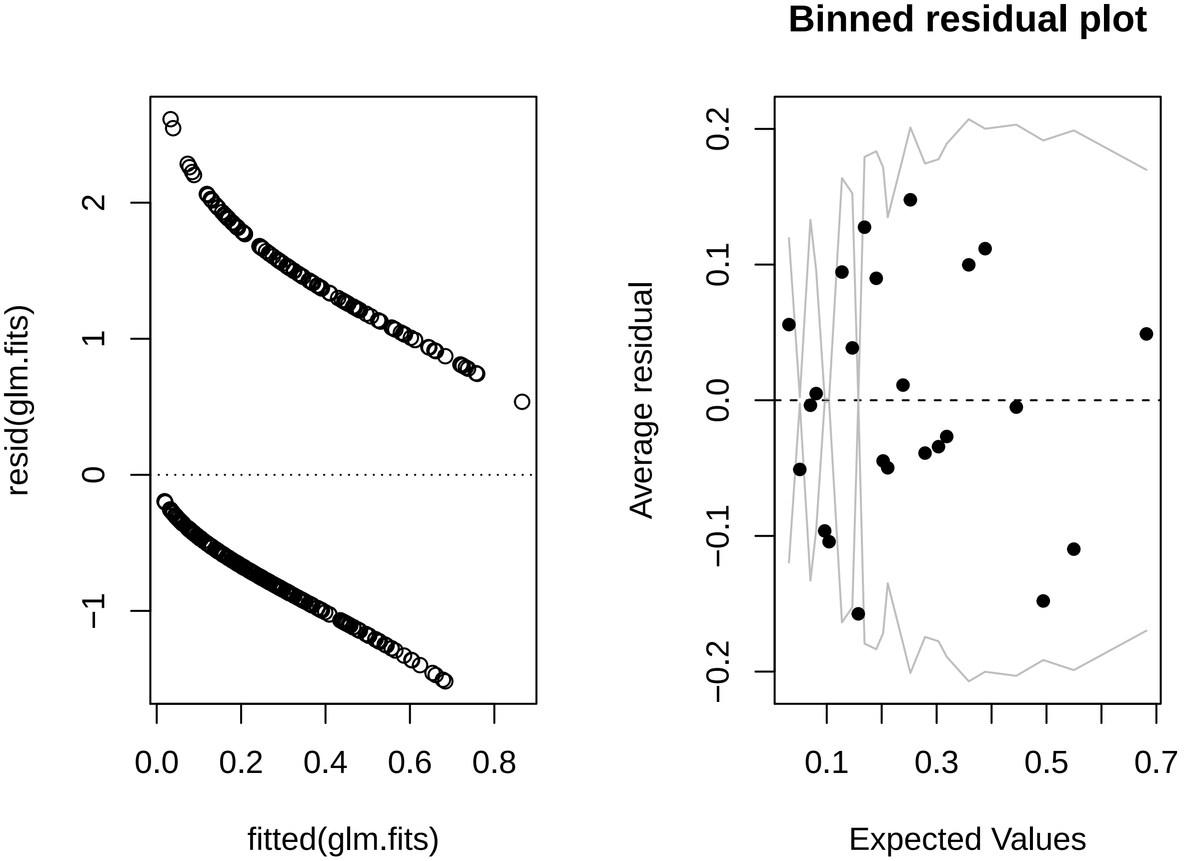
IV. Modeling analysis

1. Logistic Model

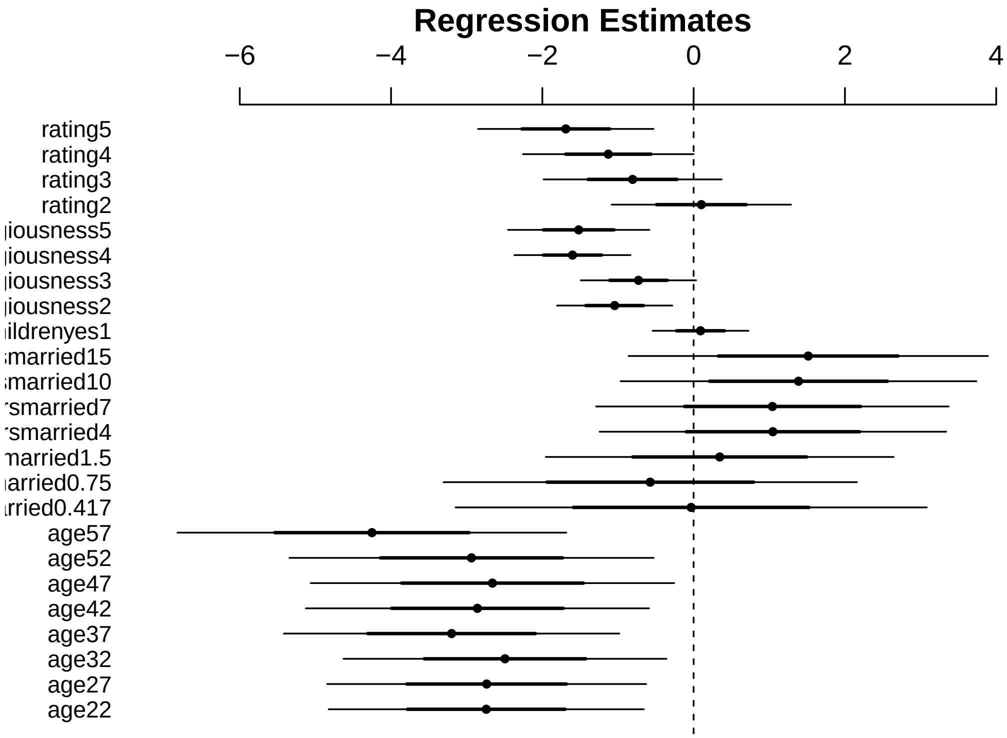
First, we want to know the possibility of a person choosing to have affairs, we consider Logistic Model. According to the correlation plot Figure 4., we choose age, yearmarried, children, religiousness and ratings, below is the result of Logistic Model:

```
##
## Call:
## glm(formula = affairs ~ age + yearsmarried + childrenyes + religiousness +
##     rating, family = binomial(), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5175  -0.7520  -0.5249  -0.1947   2.6137
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        2.65079    1.64385   1.613 0.106842
## age22             -2.74149    1.04160  -2.632 0.008488 **
## age27             -2.73618    1.05460  -2.595 0.009472 **
## age32             -2.49387    1.06748  -2.336 0.019479 *
## age37             -3.19965    1.10909  -2.885 0.003915 **
## age42             -2.85841    1.13440  -2.520 0.011744 *
## age47             -2.66019    1.20134  -2.214 0.026804 *
## age52             -2.93689    1.20381  -2.440 0.014701 *
## age57             -4.25302    1.28530  -3.309 0.000936 ***
## yearsmarried0.417 -0.03263    1.55762  -0.021 0.983286
## yearsmarried0.75  -0.57385    1.36631  -0.420 0.674483
## yearsmarried1.5    0.34506    1.14955   0.300 0.764047
## yearsmarried4      1.04794    1.14566   0.915 0.360344
## yearsmarried7      1.04154    1.16553   0.894 0.371526
## yearsmarried10     1.38642    1.17619   1.179 0.238504
## yearsmarried15     1.51505    1.18788   1.275 0.202159
## childrenyes1       0.09131    0.31685   0.288 0.773212
## religiousness2    -1.04399    0.38141  -2.737 0.006197 **
## religiousness3    -0.72918    0.38144  -1.912 0.055921 .
## religiousness4    -1.60157    0.38420  -4.169 3.06e-05 ***
## religiousness5    -1.51948    0.46767  -3.249 0.001158 **
## rating2            0.10180    0.59276   0.172 0.863641
## rating3           -0.80680    0.58863  -1.371 0.170490
## rating4           -1.12819    0.56435  -1.999 0.045596 *
## rating5           -1.69011    0.58017  -2.913 0.003578 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
## Residual deviance: 586.37  on 576  degrees of freedom
## AIC: 636.37
##
## Number of Fisher Scoring iterations: 5
```

The result shows that the predictors age, religiousness and ratings are significant, they can affect the possibility of one having affairs. Then we do the model checking:



We choose to look at Binned residual plot, we can see all the residuals are distributed around 0, and all of them are inside of the boundry, which suggests our model fits well.



From the Logist model we can know that:

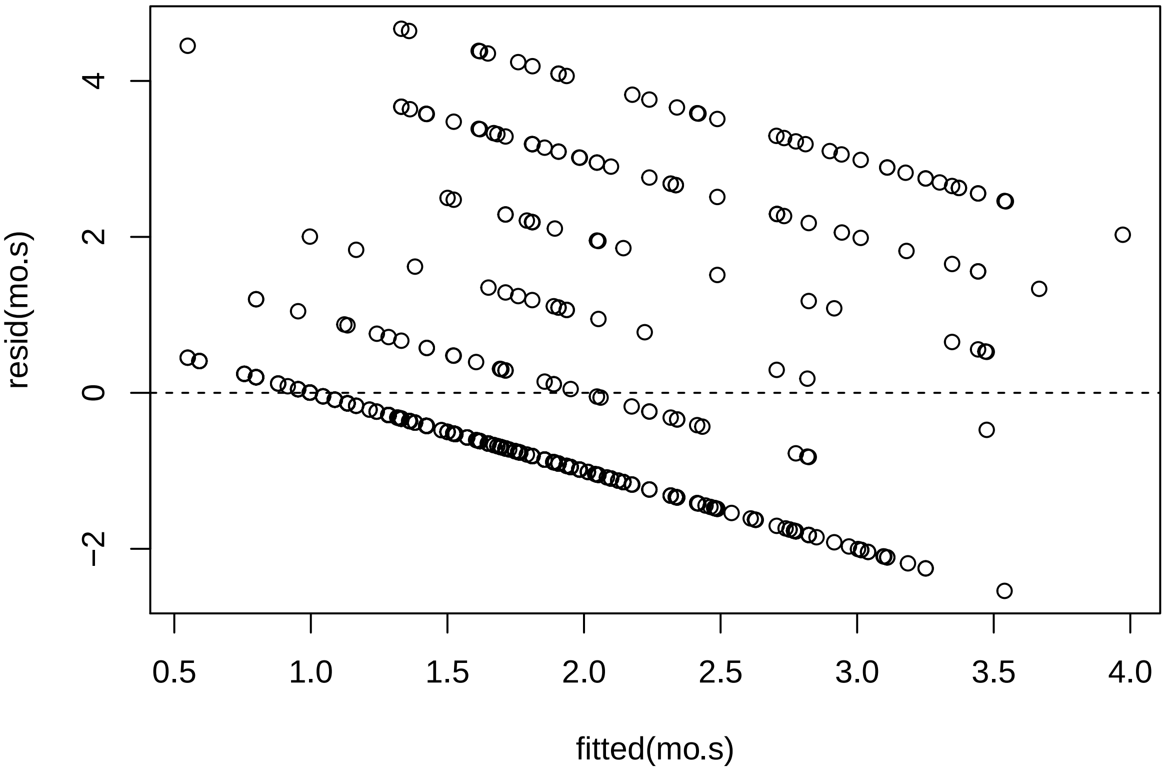
- People who do not have religiousness are more likely to have affairs;
- People who feel more happy about their marriage, are less likely to have affairs;
- Having a children in the marriage will increase the probability of having affairs.

2. Simple Linear Regression

After we know which factors are significant to the possibility of having an affair, we want to know how these factors will affect the exact numbers of affairs. We use simple linear regression in our next step, and also we get rid of the variables that are not statistically significant in previous model, below is the result:

```
##
## Call:
## glm(formula = affairs ~ age + religiousness + rating, family = gaussian(),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5396  -0.8113  -0.3628   0.2005   4.6689
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.95422    0.74366   6.662 6.25e-11 ***
## age22          -1.67686    0.61452  -2.729 0.006549 **
## age27          -1.47991    0.61230  -2.417 0.015956 *
## age32          -1.05190    0.61614  -1.707 0.088308 .
## age37          -1.19247    0.62276  -1.915 0.056004 .
## age42          -0.95501    0.63416  -1.506 0.132625
## age47          -0.82823    0.67679  -1.224 0.221531
## age52          -1.11669    0.68110  -1.640 0.101639
## age57          -1.88426    0.68081  -2.768 0.005825 **
## religiousness2 -0.72173    0.24194  -2.983 0.002972 **
## religiousness3 -0.52975    0.24977  -2.121 0.034348 *
## religiousness4 -1.05636    0.23859  -4.428 1.14e-05 ***
## religiousness5 -1.09947    0.27918  -3.938 9.20e-05 ***
## rating2         0.07013    0.40971   0.171 0.864144
## rating3        -1.05520    0.39968  -2.640 0.008509 **
## rating4        -1.13330    0.38401  -2.951 0.003292 **
## rating5        -1.42152    0.38545  -3.688 0.000247 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.145981)
##
##     Null deviance: 1506.8  on 600  degrees of freedom
## Residual deviance: 1253.3  on 584  degrees of freedom
## AIC: 2183.2
##
## Number of Fisher Scoring iterations: 2
```

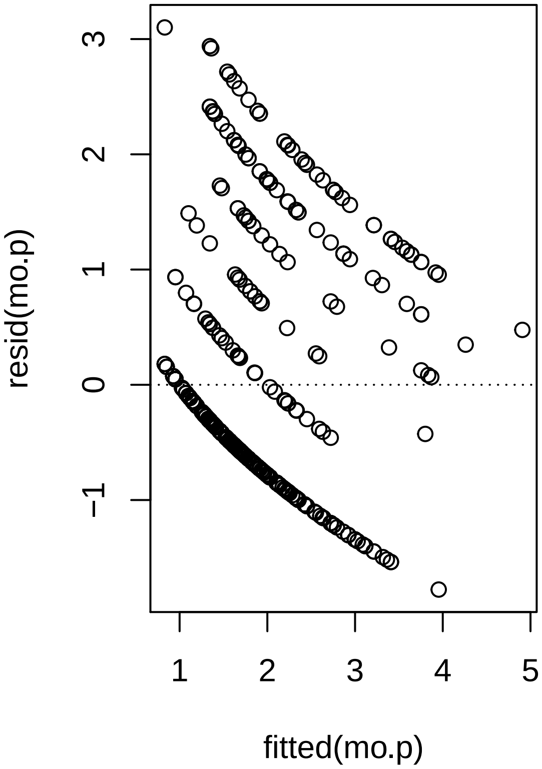
The result shows very similar result that the variable age, religiousness and the rating of the marrige are significant to the numbers of affairs. In general, the older of a person's age, the higher of a person's religious belif and the happier of a person feel about the marriage, will lead to lower number of the affairs he or she wil have. Then we do the model checking:

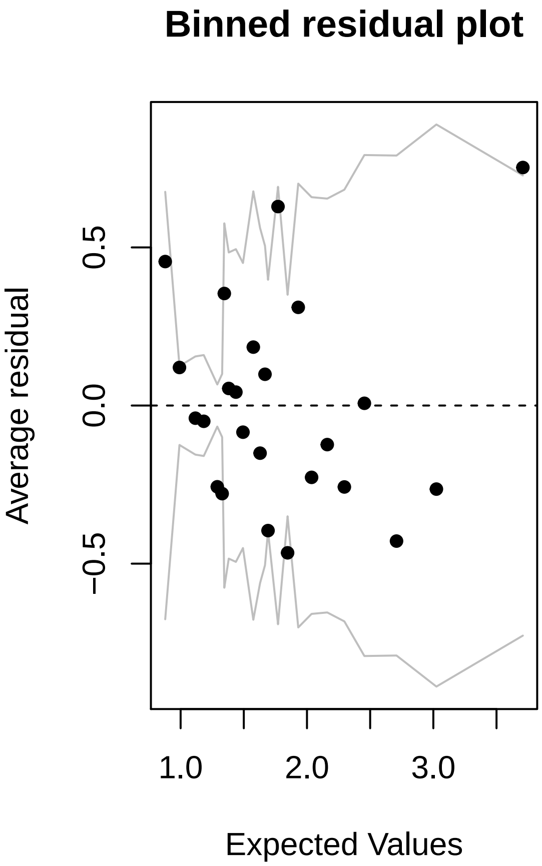


From residuals plot, we can see a clear pattern which is introduced by the categorical response variable. This suggests that we may choose another model, so we choose Poisson for the next step of the analysis, the reason is that the variables *affairs* is a counting of number of affairs.

3. Poisson Regression

```
##
## Call:
## glm(formula = affairs ~ age + religiousness + rating, family = poisson(),
##     data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.77704  -0.58805  -0.31147   0.04856   3.10108
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.05090    0.30564   6.710 1.94e-11 ***
## age22          -0.86168    0.25670  -3.357 0.000789 ***
## age27          -0.71578    0.25274  -2.832 0.004625 **
## age32          -0.49082    0.25348  -1.936 0.052826 .
## age37          -0.54985    0.25739  -2.136 0.032661 *
## age42          -0.43917    0.26247  -1.673 0.094291 .
## age47          -0.36429    0.28525  -1.277 0.201561
## age52          -0.50443    0.29059  -1.736 0.082578 .
## age57          -0.97271    0.30802  -3.158 0.001589 **
## religiousness2 -0.36432    0.11046  -3.298 0.000973 ***
## religiousness3 -0.26783    0.11299  -2.370 0.017768 *
## religiousness4 -0.56348    0.11168  -5.045 4.53e-07 ***
## religiousness5 -0.58970    0.13772  -4.282 1.85e-05 ***
## rating2         0.03071    0.16323   0.188 0.850786
## rating3        -0.44734    0.16609  -2.693 0.007074 **
## rating4        -0.48750    0.15681  -3.109 0.001878 **
## rating5        -0.67470    0.15989  -4.220 2.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 631.32  on 600  degrees of freedom
## Residual deviance: 498.87  on 584  degrees of freedom
## AIC: 1921.8
##
## Number of Fisher Scoring iterations: 5
```





For the model check we use binned residuals plot, and the residuals are located between intervals, fluctuating around 0. The binned residuals show the interval for the estimated coefficients, which is better compared to previous model.This time we have smaller variance for each coefficients, which means the estimation is more precise. We can still draw similar conclusion about the number of affairs, for detailed interpretation:

- The average number of affairs for people who don't have religious belief will be 36.43% lower than people who are anti-religious;
- The average number of affairs for people who slightly have religious belief will be 26.78% lower than people who are anti-religious;
- The average number of affairs for people who have very strong religious belief will be 58.97% lower than people who are anti-religious;

- People who are somewhat unhappy about their marriage will have average number of affairs that is 3.07% lower than those who are unhappy about their marriage;
- People who are very happy about their marriage will have average number of affairs that is 67.47% lower than those who are unhappy about their marriage.

4. Mixed-effect Poisson Regression

Now we consider to build a mixed-effect model, we build the model with random intercepts, ramdom slope and with respect to both religiousness and rating.

```
## Data: data
## Models:
## ml.fit1: affairs ~ religiousness + rating + (1 | age)
## ml.fit2: affairs ~ religiousness + rating + (rating - 1 | age)
## ml.fit3: affairs ~ religiousness + rating + (1 + rating | age)
## ml.fit4: affairs ~ religiousness + rating + (religiousness - 1 | age)
## ml.fit5: affairs ~ religiousness + rating + (1 + religiousness | age)
## ml.fit6: affairs ~ religiousness + rating + (1 + religiousness + rating |
## ml.fit6:     age)
##          Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## ml.fit1 10 1930.3 1974.3 -955.14   1910.3
## ml.fit2 24 1937.1 2042.7 -944.56   1889.1 21.147     14    0.09792 .
## ml.fit3 24 1937.1 2042.7 -944.57   1889.1  0.000      0    1.00000
## ml.fit4 24 1951.4 2057.0 -951.70   1903.4  0.000      0    1.00000
## ml.fit5 24 1951.4 2057.0 -951.70   1903.4  0.000      0    1.00000
## ml.fit6 54 1985.6 2223.1 -938.78   1877.6 25.838     30    0.68335
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

fit1: grouped age as radom effect intercept.

fit2: random slope of rating grouping by age.

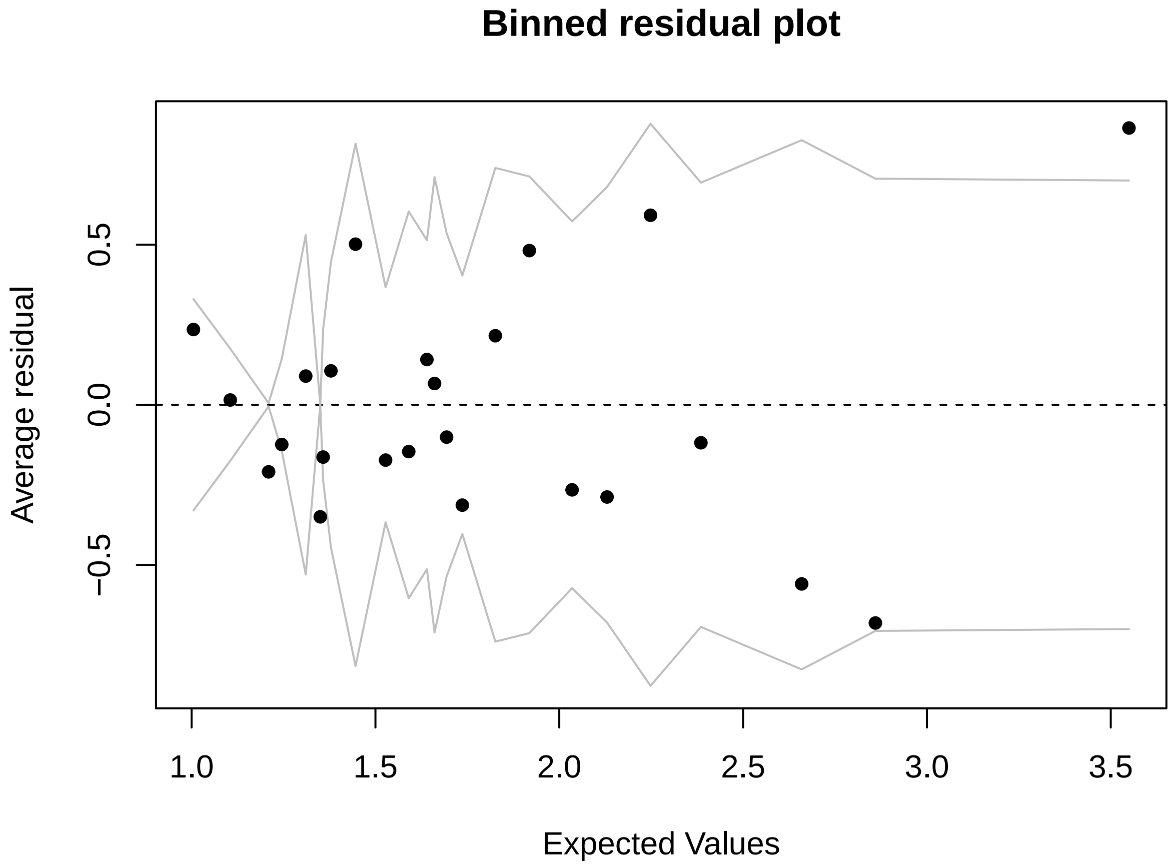
fit3: random slope and intercept for rating grouped by age.

fit4&5: same as fit2&3 with respect to religiouness.

fit6: ramdom slope and intercept for rating and religiousness grouped by age.

We can select model from the result, we see that as the model getting complex, both the test result of AIC and BIC increase steady, which means complex model do not perform better than simple model that only contains random intercept, so we choose the model with only random intercepts.

Model check fit1 :



Still we can see that most residuals are located within interval and around 0.

V.Discussion

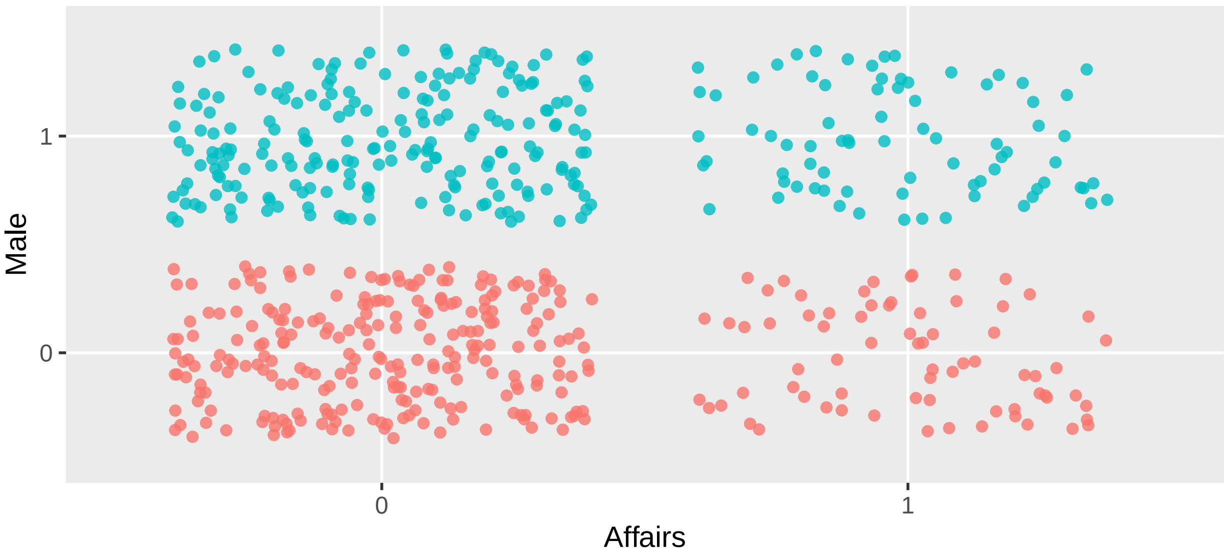
We went through *Logistic Model*, *Simple Linear Model*, *Poisson Model* and *Mixed-effect Poisson Regression*, as we kept modify our model to reflect more features of the data set, we got better result. We find that *attitude towards religions* and *ratings about marriage* are two key factors that will affect the behavior of having affairs in the marriage. The limitation of the model is that the number of affairs is a categorical varibles due to the design of the survey, but we put it into a Possion regression model, this may somehow affact the result of our model.Therefore, the future work of this project would be finding more data that contains non-categorical variables for the number of affairs record and maybe more predictors affacted the extramarital affairs such as the type of marriage(heterosexual or homosexual) of the times of the marriage of a person(1st,2nd,etc).

VI. Reference

Greene, W.H. (2003).*Econometric Analysis*, 5th edition. Upper Saddle River, NJ: Prentice Hall.

Fair, R.C. (1978). A Theory of Extramarital Affairs.*Journal of Political Economy*,86, 45–61.

VI. Appendix



We want to plot the possibility of one choosing to have affairs. Above plot is a scatter plot that we can see for both male and female, the possibility are similar, it is hard to tell the difference between gender in the plot.