**Exploration of Happiness**

Peter Ding

Project Description

**Introduction:**

The goal of this project is to predict what makes people happy. The factors that we are considering are gender, work hours, and love relationship. We believe that they all contribute to happiness. Our research question are: Is female happier than male? Is people who has less hours of work each week happier than those who has more hours of work? Is people who has a good love relationship happier than those who are lonelier? Base on our hypothesis, we predict that female is generally happier than male, people who has a good love relationship is happier than those who are lonelier, and people who work less time each week is happier than those who has more hours of work.

The data set, projdata.txt, contains 100 sample of volunteers with their information about happiness, gender, work hours, and love relationship. The happiness level is rated on a 10 point scale (1 = very unhappy, 10 = very happy). Gender is either 0(male) or 1(female). The work hours is the number of hours they work each week. The quality of their love relationship(s) is also rated on a 10-point scale (1 = very lonely, 10 = deeply in love).

**Method:**

In this project, we utilize R-language to analyze the relationships in our dataset. Firstly, we use the R command, pair(), to create a Scatter plot for our data. The scatterplot gives us a brief review of the relationships among dependent variable Y and the independent variables X where Y is happy and X is gender, work hours, or relationship. By analyzing the data distribution for X and Y, we can determine whether they are related or not. Then we use the command lm() to fit our first order linear regression model with happy ~ gender+workhrs+relationship and use the command summary() to get information of our first order linear regression model. Because there might be potential violation on the initial model, we want to know if there are interactions between each two variables, so we fit another linear regression model with two-way interaction using command lm() with happy vs (gender + workhrs + relationship) ^2. We then use summary() to analysis the new fitted model. By the partial P-value, we can verify if there is a two-way interaction between variables. Using anova() function, we can compare the full model with the reduced model, and select the best result. After that, we use the “forward”, “backward” and “both” ways stepwise regression to confirm our selected model if they return the same results. In addition, we check if there is any potential violation by analyzing the residual plot, QQ plot and histogram of our final model to see if they follow the characteristics of a proper regression. Lastly, we evaluated the fit of final model by analysis the summary table which provides p-value, r-squared and the coefficients of each variables.

**Result:**

1. Exploring the data with Scatterplots

According to the scatterplot (Appendix 3), we observed that the mean of females’ happy level is higher than males’ happy level. In other words, we can say females have higher happy level than males have. In addition, we found that the happiness and the work hours have weakly negative relation with each other. Furthermore, we also observed that the love relationship and the level of happy have moderate positive relationship.

2. Fitting the 1st order linear model to the data

According to summary table of initial model (Appendix 3), we get the first order model which contains three X variables: Y= 3.54123+1.55447\*X1-0.07118\*X2+0.48538X3+ error. Based Appendix 3, we get the overall P-value is 2.2e-16, which is less than 0.05. Because the P-value is less than 0.05 which is significant, we can reject null hypothesis which is 𝛽0=𝛽1=𝛽2=𝛽3=0. Thus, based on this statistic test, we have enough confidence to conclude that this first order model has at least one slope(β) is not equal to zero.

3. Evaluation of including 2-way interaction and explain the extra ss test.

Based on the summary table of interaction model (Appendix 4), we get the information about the 2-way interactions and the P-values for three interactions. The first interaction is between gender and work hours, and its P-value is 0.5810 which is larger than 0.05. Hence, we have insignificant P-value. There is no sufficient evidence to prove gender and work hours have relationship. In addition, the P-value of interaction between work hours and relationship is 0.5030 which is insignificant. Hence, there is no sufficient evidence to prove relationship and work hours have relationship The interaction between gender and relationship is significant which has P-value(3.26e-14) less than 0.05 which is significant. Hence, we conclude that gender and relationship have relationship. Thus, we should add the interaction between gender and relationship in to our final model.

From the extra ss test, based on Anova table for 2-way interaction (Appendix 5), we get the P-value is 1.047e-12 less than 0.05 which is significant. Thus, there is sufficient evidence for us to conclude that at least one interaction should be included into our final model.

4. Model selection

The results of forward, backward and both stepwise regression methods (Appendix 6) converge on the same final model (happy~gender+relationship+workhrs+gender\*relationship). Therefore, the model selection confirmed our model.

5. Analysis of residuals for potential violations of assumption

Based on our best-fitting model, from the residual plot (Appendix 7), we can conclude that we have constant variance. In addition, based on the qq-normal plot (Appendix 8), all points are around with the line so that we can conclude that the model is normal and linear. Based on the histogram (Appendix 9), and qq-normal plot(Appendix 8), there is no obvious outlier so that we can conclude that we do not have outliers.

6. Describe your final model, including overall significance, variance explained, meaning of each regression coefficient.

Based on above steps, we find out the final model (Appendix 11) which is *Y= 4.287745+ 0.352098X1 + 0.178353X2 - 0.070259X3+0.241580X1\*X2 .* Based on our final model’s summary analysis, we get the overall P-value is 2.2\*e^(-16) which is less than 0.05, so we reject the Null hypothesis that  *β0 = β1 = β2= β3=0.* Thus, we have sufficient evidence to conclude that in this model, at least one slope is not equal to zero. From the table, we know the multiple R^2 is 0.95 which means that 95% variance can be explained by knowing the gender, relationship and work hours.

In our final model, the interception (*β0*) means that we predict that a male whose score of relationship is zero and who work for zero hour each week, has the happy level of 4.287745. The slope of relationship (*β1* ) is 0.352098 for male. Thus, we predict that, for each male who work zero hour each week, each additional relationship will increase their happy level 0.352098 points. The slope of gender (*β2*) is 0.178353. Hence, we predict that a female who has zero relationship and works zero hour each week will have 0.178353 point in happy level higher than a male. The slope of workhrs (*β3*) is -0.070259, so we predict that, for a male who has love relationship in zero score, each additional work hour will decrease 0.070259 points in happy level. The slope of relationship:gender (*β4*) is 0.241580, so we predict that an additional score of relationship will increase the happiness for female 0.241580 points higher than the happiness increase for male. About the interaction part, we create the accompanying interaction plot.

Based on the interaction plot (Appendix 10), we know that people with better love relationship will have higher happy level. In addition, we know females have higher happy level than males. Moreover, the relationship has bigger effect on happy level for females.

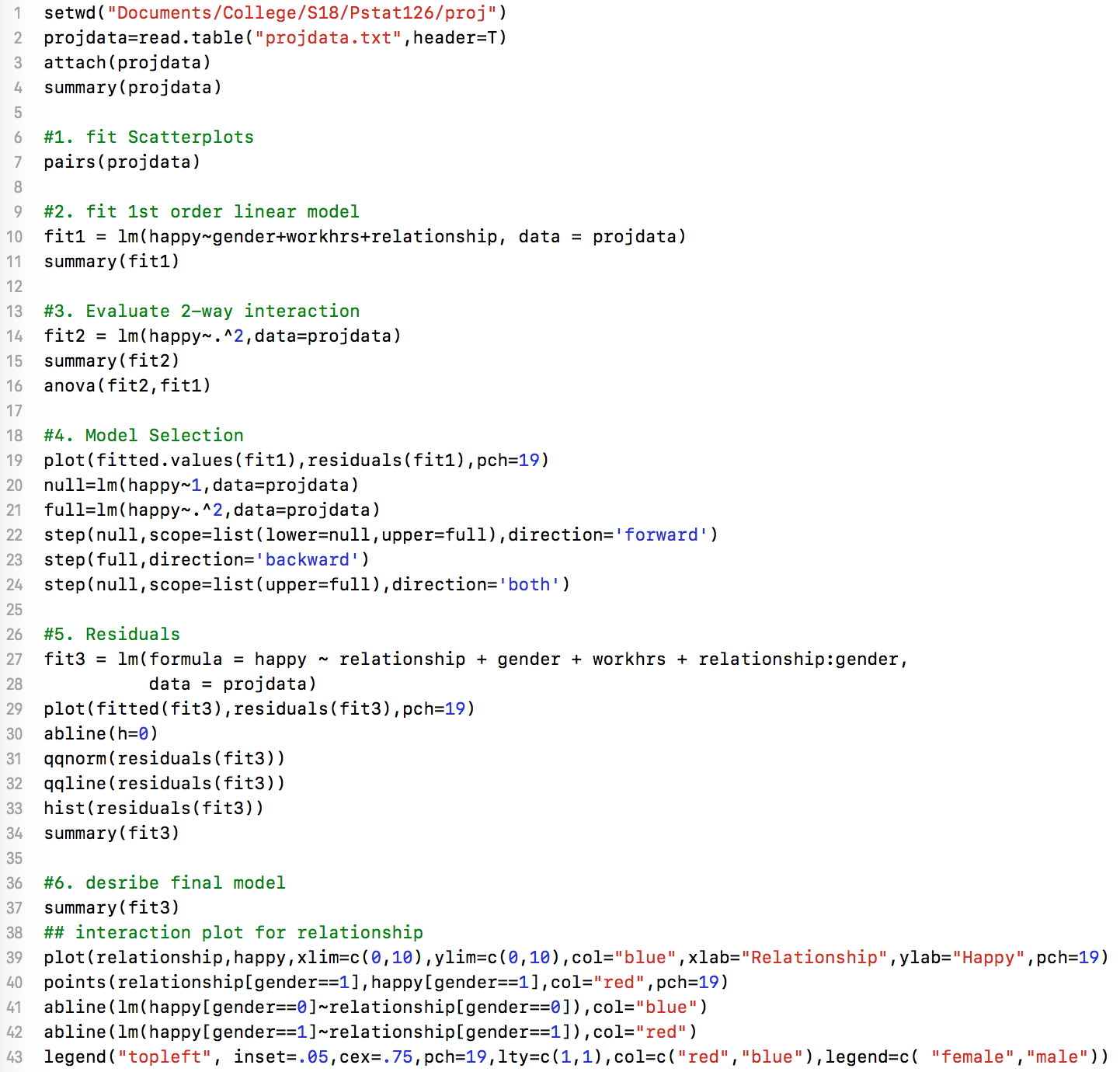
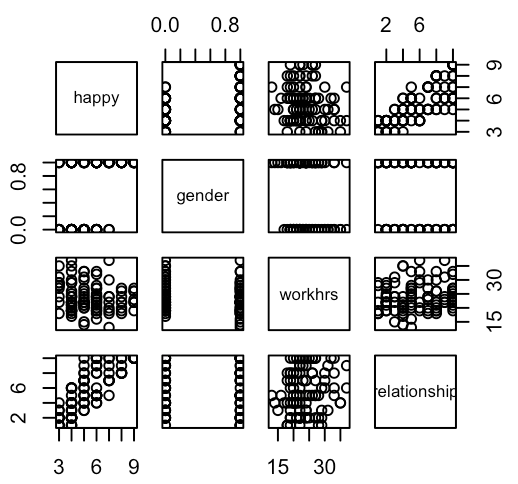
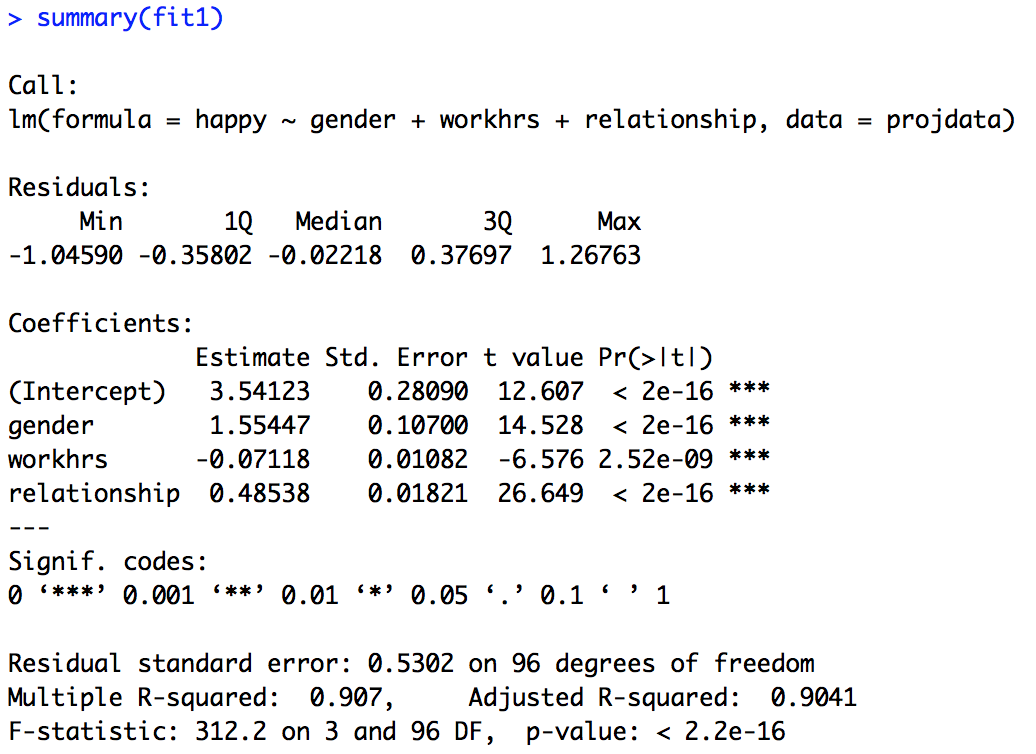
Based on final model’s summary (Appendix 11), we get that the P-value for slope of X1(relationship) is less than 2e-16. Hence, it is significant so that we should conclude X1 in our model. The P-value for X2(gender) is 0.301 larger than 0.05, hence it is insignificant. The P-value for X3 (work hours) is 5.85e-14 less than 0.05 which is significant. Hence, we should include X3 into our model. The P-value for X1:X2 (relationship:gender) is 1.84e-14 less than 0.05, which is significant. Hence, we should include this interaction into our model.

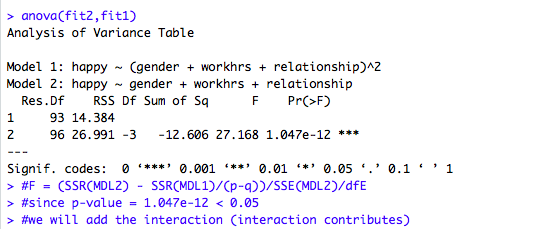
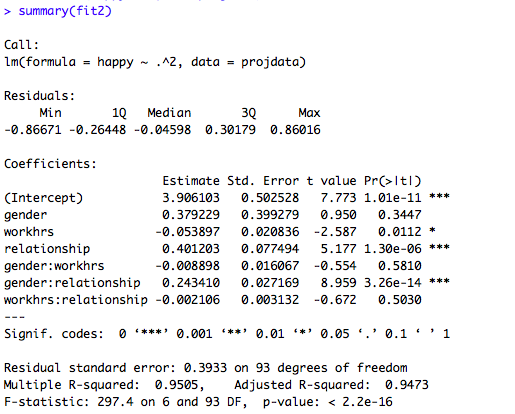
**Discussion:**

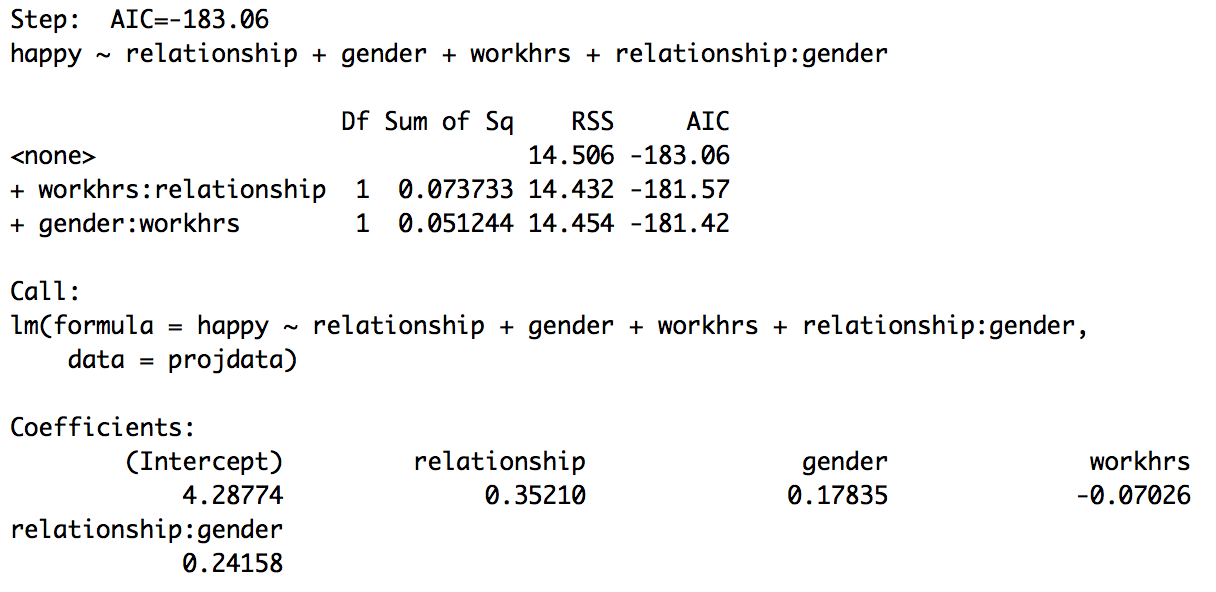
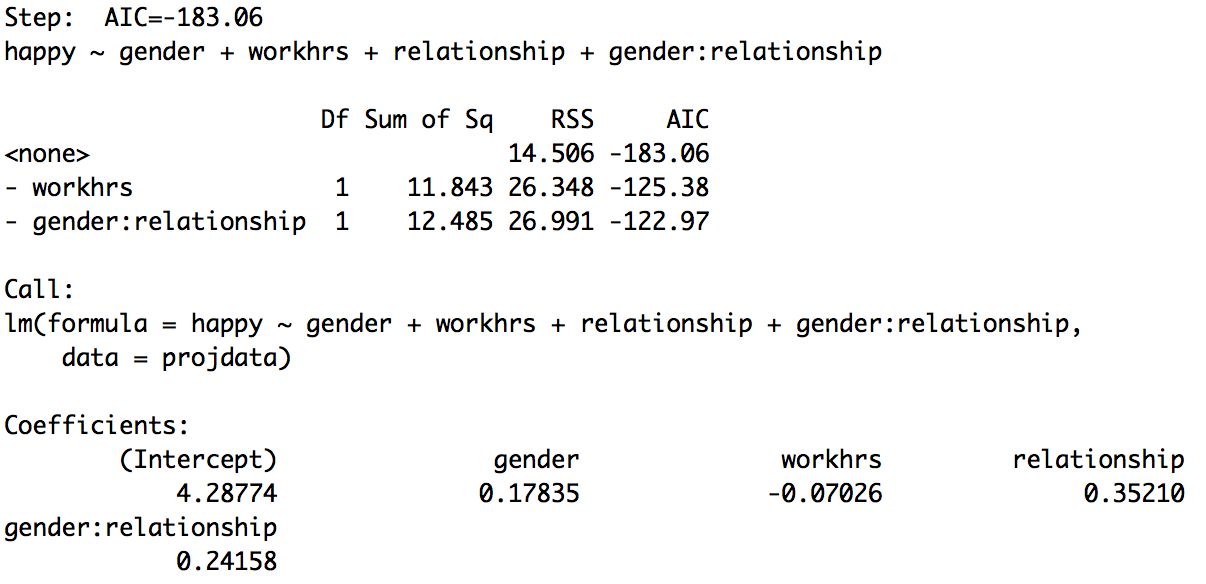
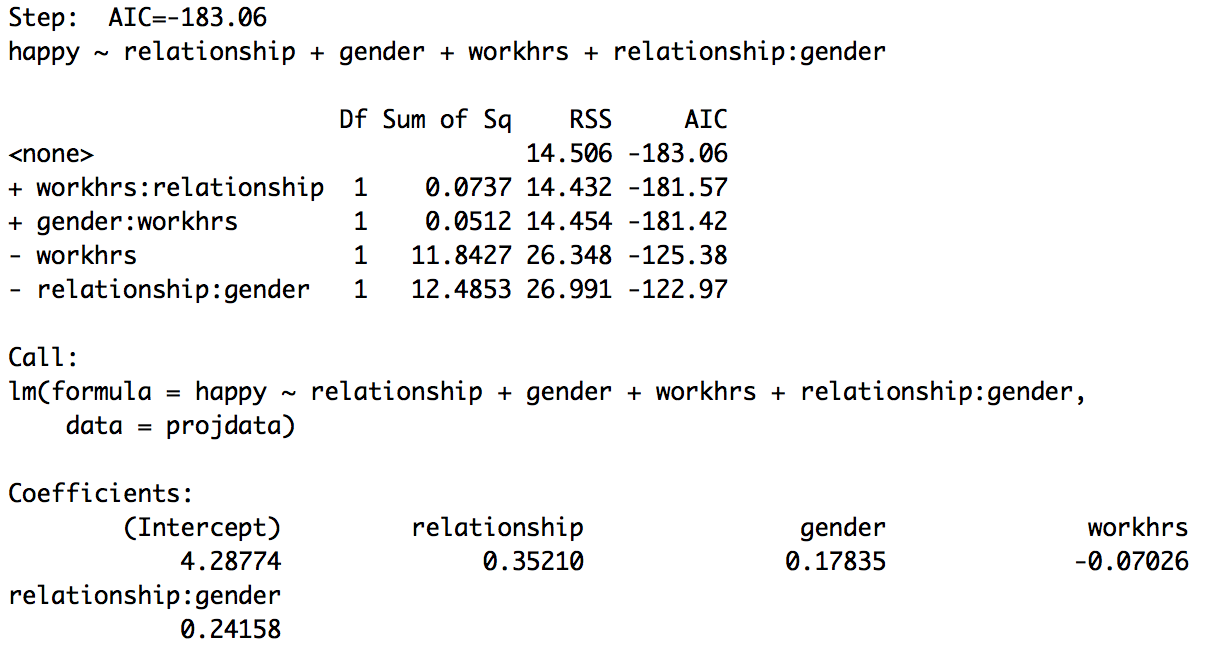
In conclusion, the linear regression results obtained followed our initial belief. In general, female is happier than male. Also, an increase in an individual’s working hour will decrease his or her happiness. A person will be significantly happier if he or she is in a good love relationship. In addition, the relationship between happiness and love relationship is different for male and female while a female is more sensitive towards love relationship.

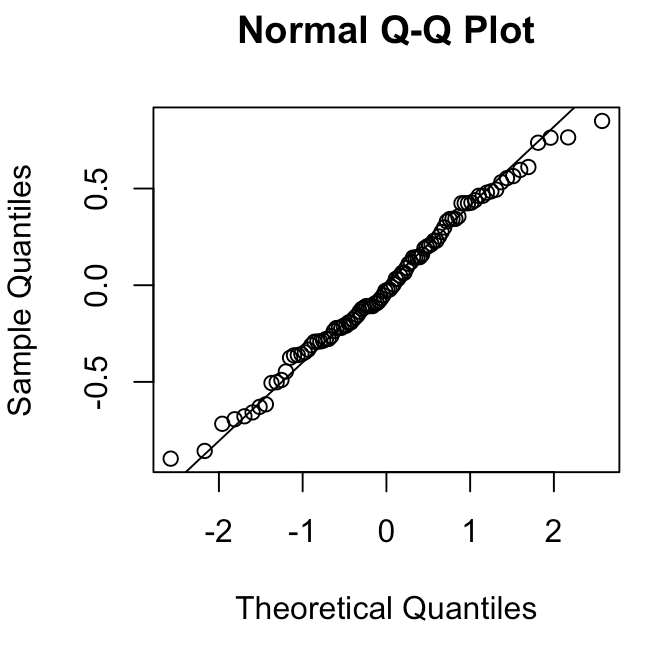
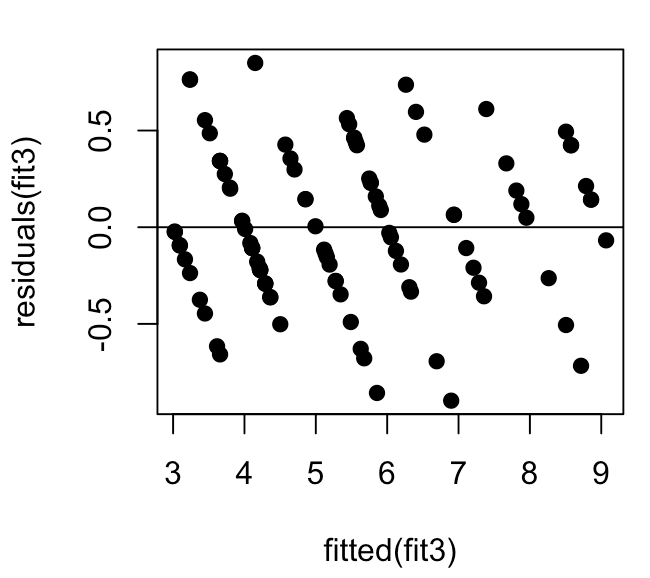
There are also several limitations to the model. The volunteers may have provided inaccurate data since there was no way to verify the information obtained from the survey. Also, the 100 volunteers might from the same background with the same age and dissatisfied with their jobs and love relationship. Given the above loopholes, it would not be accurate to conclude that there is a correlation between happiness and gender, work hours, and relationship. In the future research, it would be appropriate to use a large random sample and take more observations of the individuals such as age in order to obtain a stronger result.

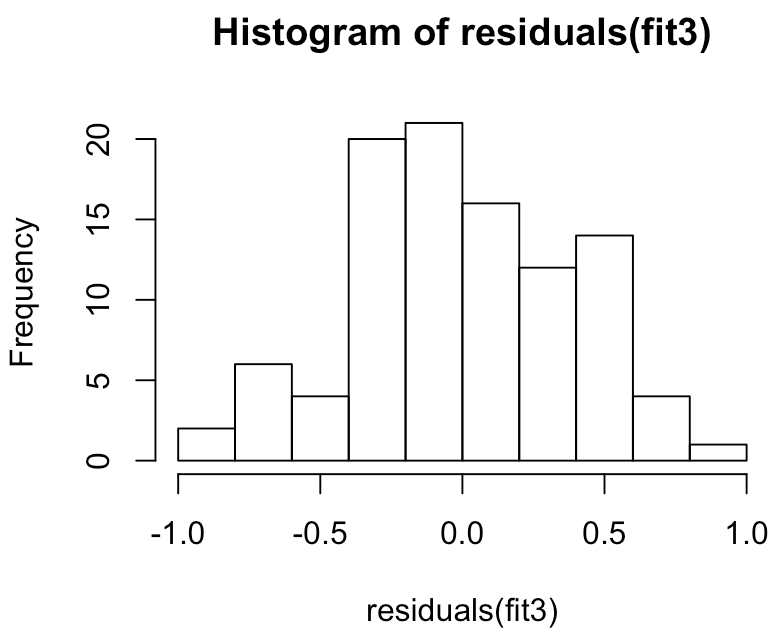
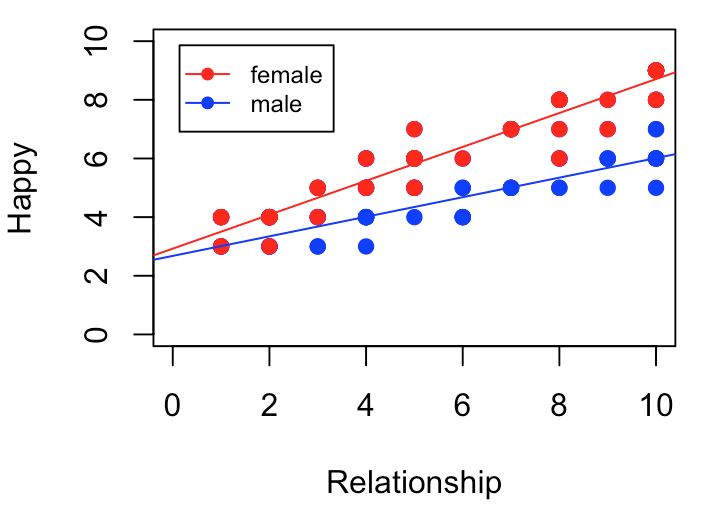
Appendix:

1. R code used in this lab 
2. Scatterplot of data 3. Summary table for initial model 

4. Summary table for 2-way interaction 5. Anova table for 2-way interaction

6. Forward stepwise selection Backward stepwise selection Both stepwise selection

7. Residuals plot of best model 8. QQ plot of best model 

9. Histogram of best model 10. Interaction plot 

11. Summary table of final model