Introduction

In this project, we are going to find the best fit model for the relationship between people's happiness level and factors such as genders, number of hours they work per week, and quality of their love relationship(s). In this dataset with 100 observations, we'll use happiness as the outcome variable(Y), and gender(X_1), number of hours of work per week(X_2), and quality of love relationship(s)(X_3) as predictor variables. Gender is a dummy variable with 0 being male and 1 being female, quality of love relationship is rated on a 10-point scale with 1 being very lonely and 10 being deeply in love.

We expected that female be happier than male, and quality of relationship having positive effect on happiness level, and number of work hours having negative effect on happiness level. Also, the effect of quality of relationship and work hours on happiness might vary between female and male.

Method

First, we reviewed the dataset by looking at its scatterplots. We got our initial exploration of how gender, work hours and love relationship was related to happiness level in terms of to what extent and in what directions they are related.

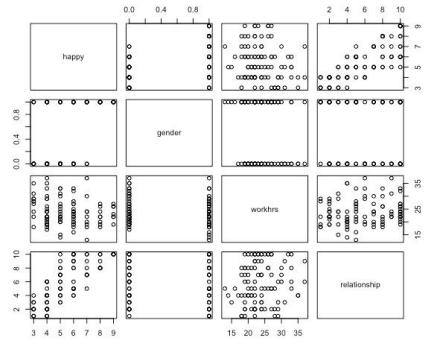
Then, we fit a 1st order linear model to the data by running a regular multiple linear regression $Y' = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ with happiness level as outcome Y', gender x_1, work hours x_2, and relationship x_3 as predictors and name it as model 1. We did a hypothesis test on whether or not the slopes are all zero to determine whether or not all three predictors were significant enough to contribute to the model.

We then used extra sum of squares test to evaluate its two-way interaction model $Y' = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3$, where we added three multiplicative variables to see if adding interactions could contribute even more to the model and named it as model

2. We firstly used ANOVA function to test whether model 2 was significant enough to be kept in general, otherwise we had to use model 1. Then we used SUMMARY function to test on whether or not all interaction variables were significant enough to be kept. We looked at the partial p-values when variables entered last to determine which interaction to keep. And we used forward, backward, and both-direction techniques in stepwise regression to arrive at our final model. After confirming the final model, we interpreted and diagnosed potential violations of the model assumptions and then evaluated it again from its SUMMARY table to see if everything was improved. We also used interaction plots to see if slopes were acting corresponding to our choice of interaction.

Results

By looking at the scatterplots, we found a strong positive relationship between happiness level and quality of love relationship and a weak negative relationship between happiness level and work hours. The scatterplot between happiness level and gender showed a slightly higher level of happiness within females.



Then, we did the hypothesis test to test whether model 1 was significant. The null hypothesis was $\beta_1=\beta_2=\beta_3= 0$ and the alternative test was that at least one of the coefficient was not 0. From the summary table, we got overall p-value less than 2.2e-16, which was less than significance level of 0.05. Thus, we rejected the null hypothesis and concluded that the model 1 $Y' = 3.54123 + 1.55447x_1 - 0.07118x_2 + 0.48538x_3$ was significant.

We added three interaction variables in model 2 and used ANOVA function to compare model1 and model 2. The null hypothesis was $\beta_{12}=\beta_{13}=\beta_{23}=0$ and the alternative hypothesis was at least one of the interaction term not 0. We found the overall p-value

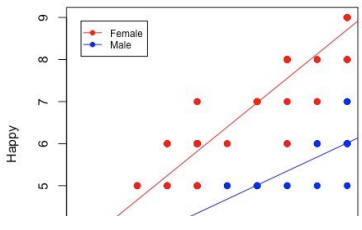
1.047e-12, which was far less than 0.05. Therefore, we concluded that model 2 that is with interaction variables was better than model 1.

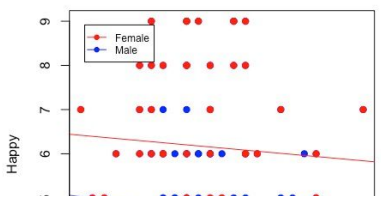
By looking at the partial p-values in the SUMMARY function of model 2, we found that among the three interaction variables only the p-value of β_{13}, the coefficient for gender*relationship, was less than 0.05, which means only this interaction was significant enough to contribute. We also noticed that among the three initial variables, only p-values for work hours and relationship were significant, but we chose to keep gender in our final model to make sure that each variable in gender*relationship was included. Therefore, our final model was $Y' = 4.28774 + 0.17835X_1 - 0.07026X_2 + 0.35210X_3 + 0.24158X_1X_3$.

All of the backward, forward and both ways of stepwise regression yielded the same model $Y' = 4.28774 + 0.17835X_1 - 0.07026X_2 + 0.35210X_3 + 0.24158X_1X_3$, which was just our final model.

The overall p-value for this final model was less than 2.2e-16, which was far less than 0.05, which implied that the model is overall significant. The partial p-values for work hours, relationship, and gender*relationship were all far less than 0.05, which was significantly improved from model 2's values. While p-value for gender was still larger than 0.05, we kept it for its existence in gender*relationship interaction.

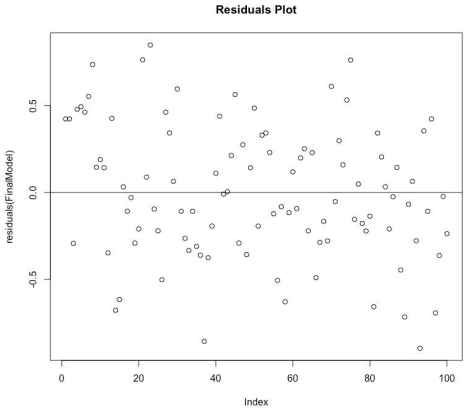
The R-Square was 0.9505, which means 95% of the variance of the data is explained by our final model. β_0=4.28774 means the happiness level for male who doesn't work and has zero relationship level is 4.28774. The slope for gender(X_1) 0.178353 is the difference between intercepts of male and female, so we predict that a woman who doesn't work and has 0 relationship will be 0.178353 happier than a man. The slope for workhr(X_2)=0.07026 means for an extra hour of work, the happiness level will decrease 0.07026. The slope for

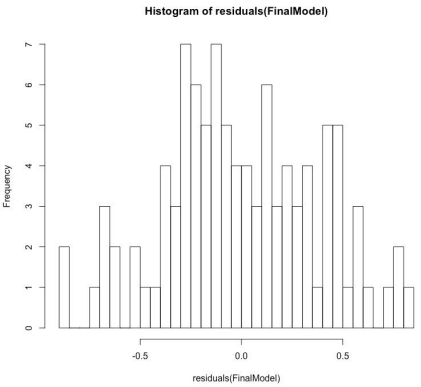




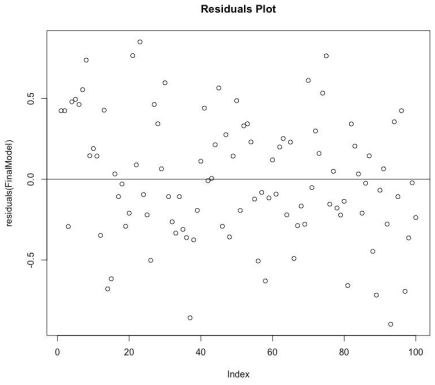
relationship(s)(X_3)=0.35210 means when the relationship level increase one, a male's happiness level will increase 0.35210. The slope for X_1X_3=0.24158 means differences in slopes between male and female, so we predict that an additional level of relationship will increase happiness for a female 0.24158 more than its increase for a male.

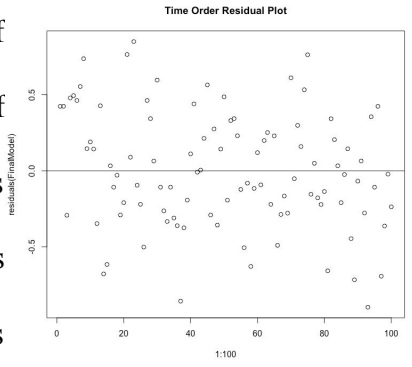
By looking at the interaction plots above, we clearly recognized that the interaction between work hours and gender was insignificant because of its almost parallel lines for female and male, and a significant interaction between relationship and gender since the slopes are notably different. This result again confirmed our choice of interaction variable to be kept in final model.

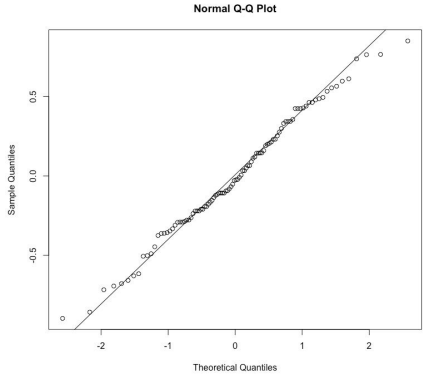




Last but not least, we used analysis of residuals for potential violations of assumptions of our final model. Since there's no clear pattern in the residuals plot, it is generally linear. According to the residuals plot and QQ plot, there is no apparent outlier, which implies it is identically distributed. Since the dots in the residual plot is evenly spreaded, it has constant variance. According to the histogram and QQ plot, it is not skewed much, which means it is normally distributed. The time order residuals plot also shows







an evenly spreaded of dots around zero line so it is also independently distributed.

Discussion

Our final model is $Y' = 4.28774 + 0.17835X_1 - 0.07026X_2 + 0.35210X_3 + 0.24158X_1X_3$. β_0=4.28774 means the happiness level for male who doesn't work and has zero relationship level is 4.28774. The slope for gender(X_1) 0.178353 is the difference between intercepts of male and female, so we predict that a woman who doesn't work and has 0 relationship will be 0.178353 happier than a man. The slope for workhr(X_2)=0.07026 means for an extra hour of work, the happiness level will decrease 0.07026. The slope for relationship(s)(X_3)=0.35210 means when the relationship level increase one, a male's happiness level will increase 0.35210. The slope for X_1X_3=0.24158 means differences in slopes between male and female, so we predict that an additional level of relationship will increase happiness for a female 0.24158 more than its increase for a male. These results are pretty much what we expected.

However, there are still some problems about this model. First, the happiness level is subjective. Everyone has his/her own standard. It is hard to value it only by scaling it from 0 to 10. Furthermore, the sample size 100 is not large enough. Thus, the result may not be convincing enough. Some questions such as whether sample was independently distributed is not clear from information given, so further research may be needed.

Appendix(plots are included in Results section)

```
> projdata=read.table("/Users/xiaokailiu/Desktop/UCSB/2018 Spring PSTAT 126
/Project/projdata.txt",header=T)
> head(projdata)
  happy gender workhrs relationship
1     6      1      18            4
2     9      1      26           10
3     4      0      30            6
4     7      1      13            5
5     9      1      27           10
6     6      1      27            5
```

```
> summary(projdata)
    happy          gender         workhrs       relationship
 Min.   :3.00   Min.   :0.00   Min.   :13.00   Min.   : 1.00
 1st Qu.:4.00   1st Qu.:0.00   1st Qu.:20.00   1st Qu.: 3.00
 Median :5.00   Median :1.00   Median :22.50   Median : 5.00
 Mean   :5.42   Mean   :0.52   Mean   :23.76   Mean   : 5.69
 3rd Qu.:6.00   3rd Qu.:1.00   3rd Qu.:27.00   3rd Qu.: 8.00
 Max.   :9.00   Max.   :1.00   Max.   :37.00   Max.   :10.00
>
> #Scatterplot
> pairs(projdata)
>
> #First order model
> attach(projdata)
The following objects are masked from projdata (pos = 3):
   gender, happy, relationship, workhrs
The following objects are masked from projdata (pos = 4):
   gender, happy, relationship, workhrs
The following objects are masked from projdata (pos = 5):
   gender, happy, relationship, workhrs
The following objects are masked from projdata (pos = 6):
   gender, happy, relationship, workhrs
The following objects are masked from projdata (pos = 7):
   gender, happy, relationship, workhrs
The following objects are masked from projdata (pos = 8):
   gender, happy, relationship, workhrs
The following objects are masked from projdata (pos = 9):
   gender, happy, relationship, workhrs
The following objects are masked from projdata (pos = 12):
   gender, happy, relationship, workhrs
The following objects are masked from projdata (pos = 13):
   gender, happy, relationship, workhrs
> model1 = lm(happy~gender+workhrs+relationship)
> summary(model1)

Call:
lm(formula = happy ~ gender + workhrs + relationship)

Residuals:
    Min      1Q  Median      3Q     Max
-1.04590 -0.35802 -0.02218  0.37697  1.26763

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.54123    0.28090  12.607  < 2e-16 ***
gender        1.55447    0.10700  14.528  < 2e-16 ***
workhrs      -0.07118    0.01082  -6.576 2.52e-09 ***
relationship  0.48538    0.01821  26.649  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5302 on 96 degrees of freedom
Multiple R-squared:  0.907,             Adjusted R-squared:  0.9041
F-statistic: 312.2 on 3 and 96 DF,  p-value: < 2.2e-16

>
> #2-way interactions (Extra SS test):
> model2 = lm(happy~.^2,data = projdata)
```

```
> summary(model2)

Call:
lm(formula = happy ~ .^2, data = projdata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.86671 -0.26448 -0.04598  0.30179  0.86016

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.906103   0.502528   7.773 1.01e-11 ***
gender               0.379229   0.399279   0.950   0.3447
workhrs             -0.053897   0.020836  -2.587   0.0112 *
relationship         0.401203   0.077494   5.177 1.30e-06 ***
gender:workhrs      -0.008898   0.016067  -0.554   0.5810
gender:relationship  0.243410   0.027169   8.959 3.26e-14 ***
workhrs:relationship -0.002106   0.003132  -0.672   0.5030
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3933 on 93 degrees of freedom
Multiple R-squared:  0.9505,     Adjusted R-squared:  0.9473
F-statistic: 297.4 on 6 and 93 DF,  p-value: < 2.2e-16

> anova(model1,model2)
Analysis of Variance Table

Model 1: happy ~ gender + workhrs + relationship
Model 2: happy ~ (gender + workhrs + relationship)^2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     96 26.991
2     93 14.384  3    12.606 27.168 1.047e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> #Interaction plot with separate regression line
> #workhrs vs happy
> plot(workhrs,happy,col="blue",pch=19,xlab="Work Hours",ylab="Happy")
> points(workhrs[gender==1],happy[gender==1],col="red",pch=19)
> abline(lm(happy[gender==0]~workhrs[gender==0]),col="blue")
> abline(lm(happy[gender==1]~workhrs[gender==1]),col="red")
> legend("topleft", inset=.05,cex=.75,pch=19,lty=c(1,1),col=c("red","blue"),legend=c("Female","Male"))
>
> #relationship vs happy
> plot(relationship,happy,col="blue",pch=19,xlab="Relationship",ylab="Happy")
> points(relationship[gender==1],happy[gender==1],col="red",pch=19)
> abline(lm(happy[gender==0]~relationship[gender==0]),col="blue")
> abline(lm(happy[gender==1]~relationship[gender==1]),col="red")
> legend("topleft", inset=.05,cex=.75,pch=19,lty=c(1,1),col=c("red","blue"),legend=c("Female","Male"))
>
>
> #Final Model
> FinalModel =  lm(happy~relationship + gender + workhrs + relationship*gender, data = projdata)
> summary(FinalModel)

Call:
lm(formula = happy ~ relationship + gender + workhrs + relationship *
```

```
   gender, data = projdata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.89700 -0.26709 -0.02701  0.28099  0.84955

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.287745   0.222865  19.239  < 2e-16 ***
relationship      0.352098   0.019935  17.662  < 2e-16 ***
gender            0.178353   0.171396   1.041    0.301
workhrs          -0.070259   0.007978  -8.807 5.85e-14 ***
relationship:gender  0.241580   0.026716   9.043 1.84e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3908 on 95 degrees of freedom
Multiple R-squared:   0.95,             Adjusted R-squared:  0.9479
F-statistic: 451.7 on 4 and 95 DF,  p-value: < 2.2e-16

>
> #residual plot for FinalModel
> plot(residuals(FinalModel), main = 'Residuals Plot')
>
> #fitted residual plot for FinalModel
> plot(fitted(FinalModel),residuals(FinalModel),main = 'Fitted Residuals Plot ')
> abline(h=0)
>
> #QQ plot
> qqnorm(residuals(FinalModel))
> qqline(residuals(FinalModel))
>
> #histogram
> hist(residuals(FinalModel), breaks = 30)
>
> #Time order residual plot
> plot(1:100, residuals(FinalModel),main='Time Order Residual Plot')
> abline(h=0)
>
> #Model selection
> null = lm(happy~1,data = projdata)
> full = lm(happy~.^2,data = projdata)
> step(full,direction  ='backward')
Start:  AIC=-179.9
happy ~ (gender + workhrs + relationship)^2

                    Df Sum of Sq    RSS     AIC
- gender:workhrs        1    0.0474 14.432 -181.57
- workhrs:relationship  1    0.0699 14.454 -181.42
<none>                           14.384 -179.90
- gender:relationship   1   12.4145 26.799 -119.68

Step:  AIC=-181.57
happy ~ gender + workhrs + relationship + gender:relationship +
   workhrs:relationship

                    Df Sum of Sq    RSS     AIC
- workhrs:relationship  1    0.0737 14.506 -183.06
```

```
<none>                          14.432 -181.57
- gender:relationship   1    12.4494 26.881 -121.37

Step:  AIC=-183.06
happy ~ gender + workhrs + relationship + gender:relationship

                 Df Sum of Sq    RSS     AIC
<none>                         14.506 -183.06
- workhrs             1    11.843 26.348 -125.38
- gender:relationship  1    12.485 26.991 -122.97

Call:
lm(formula = happy ~ gender + workhrs + relationship + gender:relationship,
    data = projdata)

Coefficients:
       (Intercept)             gender            workhrs       relationship  gender:relationship
           4.28774            0.17835           -0.07026            0.35210            0.24158

> step(null,scope = list(lower = null, upper = model2), direction = 'forward')
Start:  AIC=108.6
happy ~ 1

             Df Sum of Sq    RSS     AIC
+ relationship  1   183.983 106.38  10.182
+ gender        1    61.439 228.92  86.821
+ workhrs       1     6.171 284.19 108.447
<none>                    290.36 108.595

Step:  AIC=10.18
happy ~ relationship

         Df Sum of Sq     RSS     AIC
+ gender   1    67.227  39.150 -87.777
+ workhrs  1    20.043  86.335  -8.694
<none>                106.377  10.182

Step:  AIC=-87.78
happy ~ relationship + gender

                  Df Sum of Sq    RSS      AIC
+ gender:relationship  1    12.802 26.348 -125.376
+ workhrs             1    12.159 26.991 -122.967
<none>                           39.150  -87.777

Step:  AIC=-125.38
happy ~ relationship + gender + relationship:gender

         Df Sum of Sq    RSS     AIC
+ workhrs  1    11.843 14.506 -183.06
<none>                26.348 -125.38

Step:  AIC=-183.06
happy ~ relationship + gender + workhrs + relationship:gender

                    Df Sum of Sq    RSS     AIC
<none>                           14.506 -183.06
+ workhrs:relationship  1  0.073733 14.432 -181.57
```

```
+ gender:workhrs        1  0.051244 14.454 -181.42

Call:
lm(formula = happy ~ relationship + gender + workhrs + relationship:gender,
    data = projdata)

Coefficients:
       (Intercept)        relationship              gender             workhrs  relationship:gender
           4.28774             0.35210             0.17835            -0.07026             0.24158

> step(null,scope = list(lower = null, upper = model2), direction = 'both')
Start:  AIC=108.6
happy ~ 1

             Df Sum of Sq    RSS     AIC
+ relationship  1   183.983 106.38  10.182
+ gender        1    61.439 228.92  86.821
+ workhrs       1     6.171 284.19 108.447
<none>                    290.36 108.595

Step:  AIC=10.18
happy ~ relationship

             Df Sum of Sq     RSS     AIC
+ gender        1    67.227  39.150 -87.777
+ workhrs       1    20.043  86.335  -8.694
<none>                    106.377  10.182
- relationship  1   183.983 290.360 108.595

Step:  AIC=-87.78
happy ~ relationship + gender

                   Df Sum of Sq     RSS      AIC
+ gender:relationship  1    12.801  26.348 -125.376
+ workhrs              1    12.159  26.991 -122.967
<none>                           39.150  -87.777
- gender               1    67.227 106.377   10.182
- relationship         1   189.772 228.921   86.821

Step:  AIC=-125.38
happy ~ relationship + gender + relationship:gender

                   Df Sum of Sq    RSS      AIC
+ workhrs              1    11.843 14.506 -183.063
<none>                           26.348 -125.376
- relationship:gender  1    12.802 39.150  -87.777

Step:  AIC=-183.06
happy ~ relationship + gender + workhrs + relationship:gender

                    Df Sum of Sq    RSS     AIC
<none>                           14.506 -183.06
+ workhrs:relationship  1    0.0737 14.432 -181.57
+ gender:workhrs        1    0.0512 14.454 -181.42
- workhrs               1   11.8427 26.348 -125.38
- relationship:gender   1   12.4853 26.991 -122.97

Call:
```

```
lm(formula = happy ~ relationship + gender + workhrs + relationship:gender,
    data = projdata)

Coefficients:
      (Intercept)       relationship             gender           workhrs  relationship:gender
          4.28774            0.35210            0.17835           -0.07026            0.24158
```