PSTAT 126

Final Project

ZhiCheng Zhange and Kalvin Goode

Tuesday 6:00-6:50 pm

Introduction

This project aims to determine the relationship between happiness and people’s gender, hours of work, and quality of love relationships. The given dataset collects responses from 100 volunteers, and is categorized into four parts, namely happy, sex, workhrs, and relationship, in which “happy” and “relationship” are rated on a 10-point scale. In “gender” category, 0 and 1 stand for male and female, respectively. From the summary of dataset, we see that happiness ranges from 3 to 9, with mean 5.42. The mean of gender is 0.52, which indicates that numbers of male and female are evenly distributed with slightly more female. Hours of working ranges from 13 hours to 37 hours, with mean of 23.76 hours, and finally, quality of relationship ranges from 1 to 10, with mean of 5.69. Our goal is to explore the possibility of the relationship between happiness, sex, hours of work, and quality of love. Before implement different research, we predict that there is relationship between happiness and hours of work, quality of love. More specifically, happiness is positively related to quality of love and is negatively related to hours of work. In addition, we predict that gender does not have a significant relationship with happiness in general.

Method

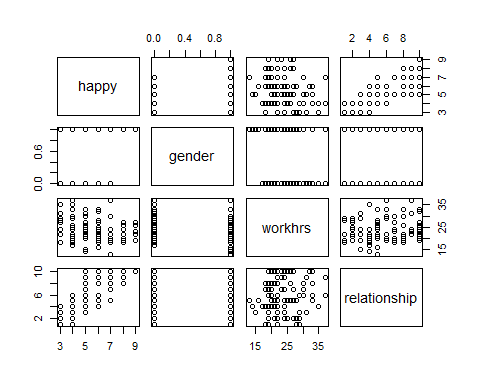
First of all, we want to know some statistics of each variable in the data set by using summary function in R, then check the general relationship between variables using scatterplot matrix (Figure 1), at the same time fit a simple first order model that uses all the variables we have in the data set. The overall p-value in summary reflects overall fitness of the first order model. And the individual p-value of each predictor tells us if three predictors have significant partial SS in this model. Then we use anova function to see if all three predictors have significant sequential SS.

After the analysis of first order model, we want to find other significant interaction by creating another model that involve second order interaction. We use anova function to compare these two models, namely the one with all 2-way interactions and the one without interactions. If the p-value in anova is significant, we do have at least one significant interaction term so that we can evaluate the summary of the larger model to identify which interactions are significant.

To decided the final model that includes interaction terms, we perform a stepwise regression using forward addition, backward elimination, and “both” direction to confirm the model has the smallest AIC value amongst other models. Finally, we check the overall fit of final model by summary function and potential violations of assumption of final model by using plotting residuals, qqnorm, qqline and histogram.

Result

Below is a scatterplot matrix for all the variables in the data set, we see that there is a positive strong linear relationship between happy and relationship, a weak linear relationship between happy and workhrs and relationship between happy and gender.



The hypothesis testing of overall regression model starts with H0: b1=b2=b3=0, Ha: not all slopes are 0. From the summary function of first order model "model1", we find that the overall p-value is <2.2e-16, which is smaller than 0.05. Thus, we reject H0 and conclude that not all slopes are 0. A significant p-value in this model indicates that all predictors in the model has some relationship to outcome 'happy'. We also see that multiple R2=0.907, so 90.7% of variance in the predictor 'happy' can be explained by knowing gender, work hours, and relationship. The initial model we get is **happy=3.54123+1.55447gender-0.07118workhrs+0.48538relationship**

The summary function illustrates that each individual p-value is <2e-16, 2.52e-9, <2e-16 for predictors gender, work hours, relationships respectively. All p-values are smaller than 0.05, which implies each predictor is significant given the other two predictors are in the model. All three predictors have significant partial SS in this model. From anova function, we find that three predictors gender, work hours, and relationship have p-value of <2e-16, 0.005536, and <2e-16 respectively. Thus we conclude that all three predictors have significant sequential SS, that is, gender is significant given nothing in the model, work hours is significant given gender is in the model, and relationship is also significant given gender and work hours are in the model.

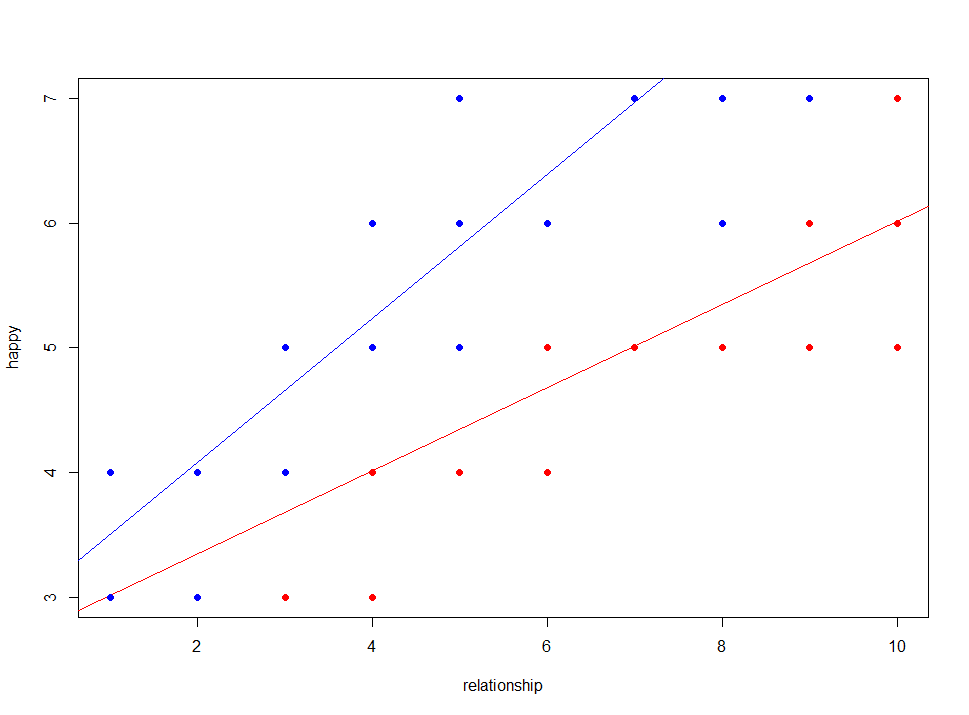
To see if there are interaction terms in the model, we create a new second-order model called "model2" that includes all possible interaction terms in the model. With anova function that compares model2 and model1, we have the following hypothesis test where b4, b5, b6 are coefficients of interaction terms of model2:

H0: b4=b5=b6=0 Ha: Not all b4, b5, b6 are zero

The p-value in anova is 1.047e-12, which is smaller than 0.05, so we reject H0 and we conclude that we at least one interaction term is significant in the model. By summary function of model2, we can identify the specific interactions that are significant. Since only the interaction gender\*relationship has a p-value=3.26e-14 smaller than 0.05, we conclude that gender\*relationship is the only interaction that is significant.

To determine the final model, we perform stepwise regression with forward addition and backward elimination, then use both direction to confirm the final model with smallest AIC. And these three directions give us the same model: **happy=4.28774+0.35210\*relationship+0.17835\*gender-0.07026\*workhrs+0.24158\*relationship\*gender**

From summary function of the final model, we see that p-value is smaller than 2.2e-16, and R2=0.95, and partial p-values for relationship, gender, work hours, relationship\*gender are <2e-16, <2e-16, 3.01, 5.85e-14, 1.84e-14 respectively. We conclude that all predictors have significant partial p-value except gender. Moreover, the interaction plot we get is as following:



Discussion

We can interpret the final model as follows: the happiness of a person is 4.63984 (4.28774+0.35210) if the person is male, relationship is 1 and hours of work is 0. The happiness of a person is 5.05977 (4.28774+0.35210+0.17835+0.24158) if the person is female, relationship is 1 and hours of work is 0. For each additional unit in relationship, male increase happiness by 0.35210 and female increase happiness by 0.59368 (0.35210+0.24158) while other variables held constant. For each additional unit of hours of work each week, the overall happiness decreases by 0.07026. Moreover, the happiness of females is 0.41993 (0.17835+0.24158) higher than the happiness of males if they have the same rate of relationship and work hours each week.

From summary function of the final model, we see that p-value is smaller than critical value 0.05, R-squared is 0.95, so 95% of variance in happy can be explained by knowing gender, work hours, relationship and the interaction of relationship and gender. Hence, we conclude that the final model fit the data.

From residual plots of the final model, we see that the final model is identically distributed with constant variance and no outlier. From qq plot and histogram of residuals, we find that the residuals are normally distributed, therefore, no truncated range problem.

Compare our prediction in the introduction with final model, we conclude that even though it is true that happiness has positive correlation to the quality of love and has negative correlation to the hours of work, we are surprised that gender also has a significant relationship with happiness and gender has interaction with relationship. We also conclude some limitations to the model, for instance, it would be better for relationship and happy to have concrete measurement rather than rate from 1 to 10; we do not know if the survey is self-reported, and the sample size it limited. For future research, we like to see the survey done by more large sample size and we expect to add some predictors such as income, age, etc.

**Appendix**

Kalvin Goode and Zhicheng Zhang

proj=**read.table**("D:/downloads/temp/Stats 126/projdata.txt", header=T)  
**summary**(proj)

## 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

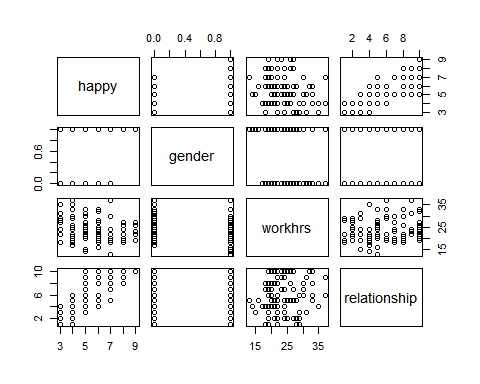
model1=**lm**(happy~gender+workhrs+relationship, data=proj)  
**summary**(model1)

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

**pairs**(proj)  
**anova**(model1)

## Analysis of Variance Table  
##   
## Response: happy  
## Df Sum Sq Mean Sq F value Pr(>F)   
## gender 1 61.439 61.439 218.5215 < 2.2e-16 \*\*\*  
## workhrs 1 2.265 2.265 8.0557 0.005536 \*\*   
## relationship 1 199.666 199.666 710.1608 < 2.2e-16 \*\*\*  
## Residuals 96 26.991 0.281   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**pairs**(proj)



**anova**(model1)

## Analysis of Variance Table  
##   
## Response: happy  
## Df Sum Sq Mean Sq F value Pr(>F)   
## gender 1 61.439 61.439 218.5215 < 2.2e-16 \*\*\*  
## workhrs 1 2.265 2.265 8.0557 0.005536 \*\*   
## relationship 1 199.666 199.666 710.1608 < 2.2e-16 \*\*\*  
## Residuals 96 26.991 0.281   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

model2=**lm**(happy~.^2, data=proj)  
**anova**(model2, model1)

## Analysis of Variance Table  
##   
## Model 1: happy ~ (gender + workhrs + relationship)^2  
## Model 2: happy ~ gender + workhrs + relationship  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 93 14.384   
## 2 96 26.991 -3 -12.606 27.168 1.047e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**summary**(model2)

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

null=**lm**(happy~1,data=proj)  
full=**lm**(happy~.^2, data=proj)  
**step**(null,scope=**list**(lower=null,upper=full), 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 = proj)  
##   
## Coefficients:  
## (Intercept) relationship gender   
## 4.28774 0.35210 0.17835   
## workhrs relationship:gender   
## -0.07026 0.24158

**step**(full,diretion="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 = proj)  
##   
## Coefficients:  
## (Intercept) gender workhrs   
## 4.28774 0.17835 -0.07026   
## relationship gender:relationship   
## 0.35210 0.24158

**step**(null,scope=**list**(upper=full), direction="both")

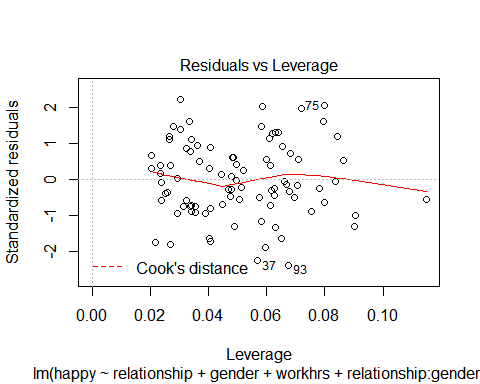
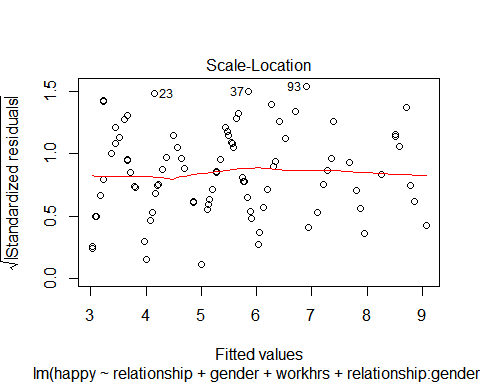
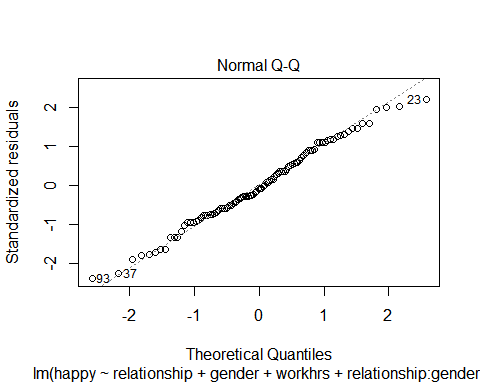
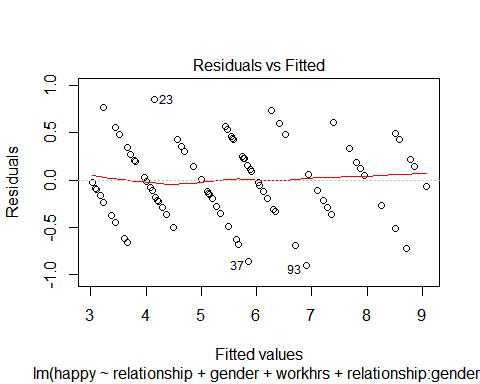
## Start: AIC=108.6  
## happy ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + relationship 1 183.983 106.38 10.182  
## + gender 1 61.439 228.92 86.821  
## + workhrs 1 6.171 284.19 108.447  
## <none> 290.36 108.595  
##   
## Step: AIC=10.18  
## happy ~ relationship  
##   
## Df Sum of Sq RSS AIC  
## + 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 = proj)  
##   
## Coefficients:  
## (Intercept) relationship gender   
## 4.28774 0.35210 0.17835   
## workhrs relationship:gender   
## -0.07026 0.24158

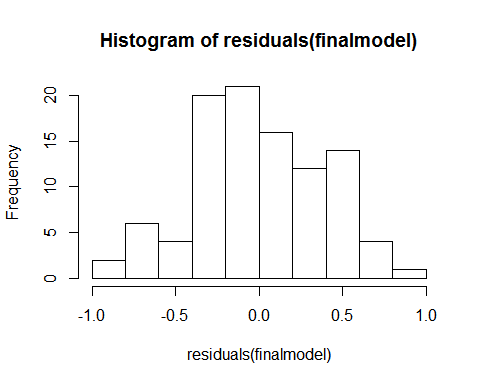
finalmodel=**lm**(formula = happy ~ relationship + gender + workhrs + relationship:gender, data = proj)  
**summary**(finalmodel)

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

**plot**(finalmodel)



**hist**(**residuals**(finalmodel))



**plot**(proj$relationship[proj$gender==0],proj$happy[proj$gender==0],col="red",pch=19, xlab="relationship",ylab="happy")  
**abline**(**lm**(proj$happy[proj$gender==0]~proj$relationship[proj$gender==0]),col="red")  
**points**(proj$relationship[proj$gender==1],proj$happy[proj$gender==1],col="blue",pch=19)  
**abline**(**lm**(proj$happy[proj$gender==1]~proj$relationship[proj$gender==1]),col="blue")

