PSTAT 126 Final Project

University of California, Santa Barbara

Spring 2018

By: Megan Godfrey(#8950156) and Jasmine Kellogg(#4266367)

TA: Lizzie Spencer

Section: 5-5:50 M

In this experiment, we seek to measure one's happiness through their gender, number of work hours, and level of satisfaction with their relationships. To explain, a survey was given out to 100 random people, where they were asked to rate their happiness on a 10-point scale, with 1 being extremely unhappy and 10 being extremely happy. After responding to this, participants were then asked to mark their gender(0=male, 1=female), how many hours a week they work, and rate the quality of their love relationship(1=very lonely and 10=deeply in love). Overall, we predict that there will be a negative relationship between happiness and work hours, a positive relationship between happiness and relationship satisfaction, and no relationship between happiness and gender.

To begin the analysis of our happiness data, using the plot function, we put the data into a scatterplot matrix that displayed the relationships between each of the three predictor variables, work hours, gender, and relationship satisfaction, and the outcome variable, happiness. By doing so, we were able to evaluate initial linearity, strength, and direction of the various relationships. After completing this initial review of the scatterplot, we moved to the summary function which displayed the slopes of each predictor variable along with their partial p-values, the overall p-value, the t and F statistics, and the R squared value. By looking at the overall p-value of the linear model, we were able to test the alternative hypothesis(H1) that not all slopes Bi were equal to zero versus the null hypothesis(H0) that all Bi = 0. We were also able to use multiple R squared to determine the percent of the variance in happiness that was explained by knowing gender, work hours, and relationship satisfaction.

We next tested the outcomes of 2-way interactions between the predictor variables by using 2-way interaction plots. Because gender is the variable that has variation, we tested the interaction of male and female work hours and happiness in the first plot and male and female relationship satisfaction and happiness in the second plot. By looking at the interaction of the 2 lines in each of the plots, we were able to assess whether or not there were interactions between the predictor variables analyzed in each of the plots. Following the 2-way interaction plots, we evaluated interaction between predictor variables using the Extra SS test. By using the Extra SS test, we used the anova function with two parameters, the initial first order model and a second model that included all interaction terms. By evaluating the significance of the p-value from the anova output, we can assess whether or not the inclusion of interaction terms creates a model that is better fit than a model without it. We also used the anova function to test the sequential effects of predictor variables and the outcomes that occur when predictor variables are tested, given other variables were assumed to be significant. By using the Extra SS Test, we can assess whether or not a predictor variable or interaction term can be thrown out and whether or not the addition of a variable has a significant effect on the linear model. In order to select the best-fitting model, we used stepwise regression. When evaluating our R-Output, we looked for the model that all three stepwise methods converged on. Using our best-fit model as our final model, we looked at the overall p-value to see if it was significant and also analyzed the multiple R-squared.

Finally we analyzed potential violations of the model by looking at the residuals. First, we assessed the residuals plot for uneven variance, any bulges in the points, or any pattern. We then plotted the histogram of the residuals to assess for distribution of the model and analyzed whether or not the plot was skewed in any direction. Next we plotted the QQ Plot and the QQ Line to look for normality as well as possible outliers in the data.

From the scatterplot matrix, we were able to evaluate linearity, strength, and direction of each relationship. For the relationship between gender and happiness, we can see a weak, positive, linear relationship. For work hours and happiness, we see a weak, negative, linear relationship between the two. Finally, for relationship satisfaction and happiness, we see a very obvious strong, positive, linear relationship between the two. In terms of our initial linear model, we formed the null hypothesis that all Bi=0 and tested it against the alternative hypothesis that not all Bi=0. By looking at the overall p-value of the linear model, we can see that the p-value is $< 2.2e-16$ which is clearly smaller than our alpha value of 0.05. Therefore, we can see that there is sufficient evidence to reject the null hypothesis, that all Bi=0, and conclude that there is in fact a relationship between at least one of the predictor variables and the outcome variable.

For the 2-way interaction portion of our analysis, we evaluated two different 2-way interaction plots. Because gender is the variable that changes, it was necessary for us to look at interaction models that accounted for the different data for males and females. In our first interaction model, we tested the hypothesis that the relationship between work hours and happiness depends on gender. By looking at the plot lines for males and females, we see that they do not interact with one another, they are parallel. Thus, there is no relationship between gender and work hours on a happiness rating. The p-value for this interaction term also indicates no interaction as the p-value is 0.6038, which is not significant. Thus, we conclude that the relationship between work hours and happiness does not depend on whether one is male or female. For our second interaction model, we tested the hypothesis that the relationship between relationship satisfaction and happiness depends on gender. By looking at the plot lines for males and females, we see that the lines are heading in a direction where they will eventually intersect.

So, there is in fact a relationship between gender and relationship satisfaction on a happiness rating. The p-value for this interaction term also displays a significant interaction between these two terms, as the value is 1.84e-14, which is less than our alpha value of 0.05. Thus, we can conclude that there is a relationship between male and female relationship satisfaction and happiness.

Following our 2-way interaction plots, we evaluated interaction of variables using the Extra SS Test with anova. We fit our first order linear model against the linear model that includes all interaction terms and assessed the significance of interaction terms. By looking at the overall p-value of the anova test, we saw that it had a p-value of 1.047e-12, which is less than 0.05. Therefore, we concluded that the addition of interaction terms was significant to the fit of our overall model. We then looked at the partial p-values of the interaction terms to determine whether or not we could eliminate any of the terms. By looking at the p-values, we saw that the interactions between gender and work hours and work hours and relationship satisfaction were insignificant. The two interactions had p-values of 0.5810 and 0.5030 respectively. However, the interaction between gender and relationship was significant with a p-value of 3.26e-14, so we eliminated the former two interaction terms from our final model but included the interaction term between relationship satisfaction and gender.

For our final model, we used all three of the stepwise regression methods to reach a final fit. We used forward, backward, and both to determine a final linear model that included all three predictor variables as well as the interaction term between relationship and gender. All three methods converged on this linear model. Stepwise regression allowed us to confirm the final model that we found in our previous Extra SS Test.

For our analysis of residuals for potential violations of assumptions, we first looked at a residuals scatterplot for our final model. By evaluating the plot, we concluded that there did not appear to be any bulging or trumpet shape of the points and thus, the variance is constant. We then looked at the individual variance plot and found no violations in terms of independence. We then plotted a histogram of our residuals and evaluated the graph for distribution, to which we found that it was symmetrical and bell-shaped and it was not skewed. Thus we concluded that the data was normally distributed. Our final residuals plot was the QQ Plot and the associated QQ Line. By plotting and evaluating these two, we noticed a very normal distribution of the plotted points. The QQ Plot confirmed our conclusion that the distribution is normal. All graphs followed similar and there were no outliers, so we were able to conclude that the data was identically distributed.

In the final model, our overall p-value was <2.2e-16, which is less than alpha, 0.05, so it is significant, meaning a relationship exists. Multiple R-squared determined that 95% of the variance in happiness can be explained by knowing all three predictor variables as well as the interaction between relationship satisfaction and gender. For a male who works zero hours and has a relationship satisfaction of zero, we predict a happiness a level of 4.288. For a female who works zero hours and has a relationship satisfaction of zero, we predict a happiness level of B0+B1, which is 4.4661. B3 indicates there is a 0.24 point difference in the plot lines of males and relationship satisfaction, and females and relationship satisfaction. When all other predictors are held constant, there is a 0.352 point increase in happiness when there is a one unit increase in relationship satisfaction. For every additional hour worked, there is a 0.0703 point decrease in happiness.

Overall, the data shows that there is a significant relationship between happiness and gender, work hours, and relationship.The strongest relationship is between happiness and relationship satisfaction, meaning people are happier when they have someone to share their life with. The other two predictors have weaker relationships with happiness. The regression coefficients in our final model prove our initial predictions to be accurate. However, some limitations of this model include the sample. We do not know the socioeconomic level of each participant, so one participant may have been a part of the upper class, and another a part of the working class, resulting in different lifestyles and happiness. Another limitation can be found in work hours, as some people enjoy work more than others, resulting in a difference in happiness level regardless of how many hours worked. For future research, we would question: How does money contribute to one's happiness. In order to prove this, we would add another predictor, income; happiness would then be measured by knowing one's gender, work hours, relationship, and income.

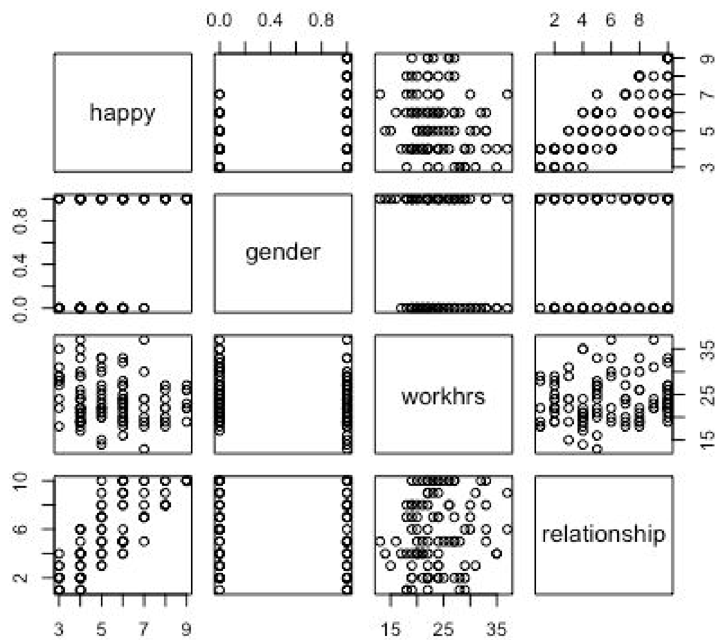
R-Code:

```
getwd()
setwd("/Users/Jasmine/desktop")
projdata<-read.table("/Users/Jasmine/desktop/projdata.txt",header=TRUE)
attach(projdata)
summary(projdata)
#create scatterplots
pairs(happy~gender+workhrs+relationship)
fitprojdata<-lm(happy~gender+workhrs+relationship)
summary(fitprojdata)
anova(fitprojdata)
#two-way interaction
plot(workhrs[gender==0],happy[gender==0],col="red",pch=19,xlab="Work
Hours",ylab="Happiness",main="Work Hours and Happiness")
abline(lm(happy[gender==0]~workhrs[gender==0]),col="red")
points(workhrs[gender==1],happy[gender==1],col="blue",pch=19)
abline(lm(happy[gender==1]~workhrs[gender==1]),col="blue")
text(35,5.5,"Male",col="red")
text(35,5.25,"Female",col="blue")
plot(relationship[gender==0],happy[gender==0],col="red",pch=19,xlab="Relationsh
ip",ylab="Happiness",main="Relationship and Happiness")
abline(lm(happy[gender==0]~relationship[gender==0]),col="red")
points(relationship[gender==1],happy[gender==1],col="blue",pch=19)
abline(lm(happy[gender==1]~relationship[gender==1]),col="blue")
text(2,6,"Male",col="red")
text(2,5.75,"Female",col="blue")
fit0<-lm(happy~gender+workhrs+relationship,data=projdata)
summary(fit0)
#creates linear model with all possible interaction terms
fit1<-lm(happy~.^2,data=projdata)
summary(fit1)
anova(fit0,fit1)
#compares linear model fit0 and fit1
fitfinal<-lm(happy~gender*relationship+workhrs,data=projdata)
summary(fitfinal)
#residuals plot
plot(fitted(fitfinal),residuals(fitfinal),xlab="Fitted",ylab="Residuals",main="
Residuals Plot")
abline(h=0)
plot(residuals(fitfinal),ylab="Residuals",main="Residuals Per Person")
abline(h=0)
#qqnorm plot
qqnorm(residuals(fitfinal))
qqline(residuals(fitfinal))
#histogram plot
hist(residuals(fitfinal))
#stepwise regression
null=lm(happy~1,data=projdata)
full=lm(happy~.^2,data=projdata)
#forward stepwise
step(null,scope=list(lower=null,upper=full),direction='forward')
#backward stepwise
step(full,direction='backward')
#both stepwise
```

```
step(null,scope=list(upper=full),direction='both')
```

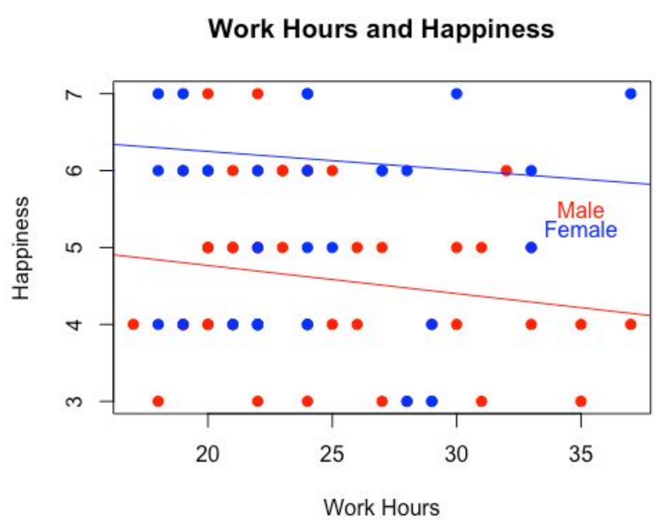
R-Output:

```
>getwd()
[1] "/Users/Jasmine/Desktop"
> setwd("/Users/Jasmine/desktop")
> projdata<-read.table("/Users/Jasmine/desktop/projdata.txt",header=TRUE)
> attach(projdata)
>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
> pairs(happy~gender+workhrs+relationship)
```



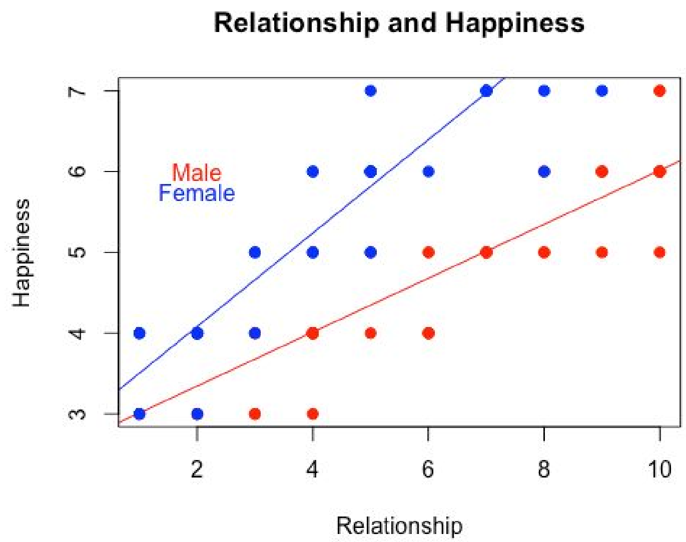

```
> fitprojdata<-lm(happy~gender+workhrs+relationship)
> summary(fitprojdata)

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

> anova(fitprojdata)
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
> plot(workhrs[gender==0],happy[gender==0],col="red",pch=19,xlab="Work 
Hours",ylab="Happiness",main="Work Hours and Happiness")
> abline(lm(happy[gender==0]~workhrs[gender==0]),col="red")
> points(workhrs[gender==1],happy[gender==1],col="blue",pch=19)
> abline(lm(happy[gender==1]~workhrs[gender==1]),col="blue")
> text(35,5.5,"Male",col="red")
> text(35,5.25,"Female",col="blue")
```




```
>plot(relationship[gender==0],happy[gender==0],col="red",pch=19,xlab="Relations
hip",ylab="Happiness",main="Relationship and Happiness")
> abline(lm(happy[gender==0]~relationship[gender==0]),col="red")
> points(relationship[gender==1],happy[gender==1],col="blue",pch=19)
```

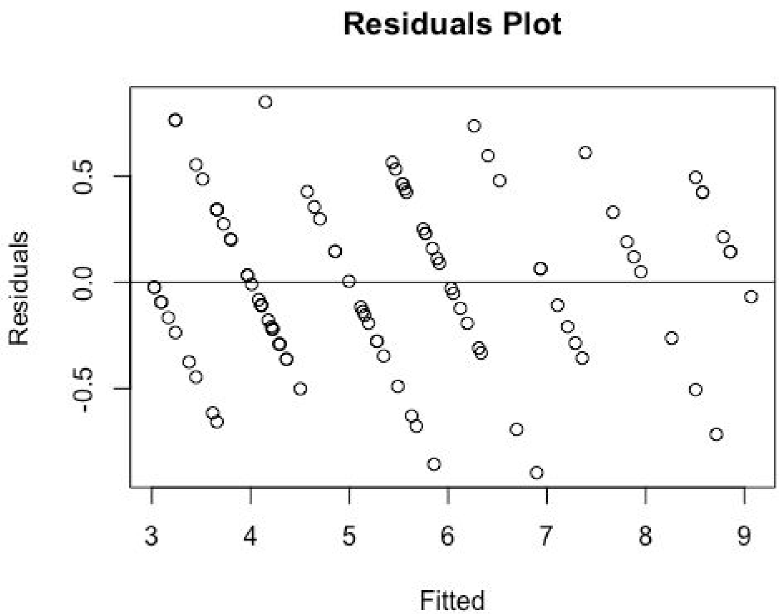
```
> abline(lm(happy[gender==1]~relationship[gender==1]),col="blue")
> text(2,6,"Male",col="red")
> text(2,5.75,"Female",col="blue")
```



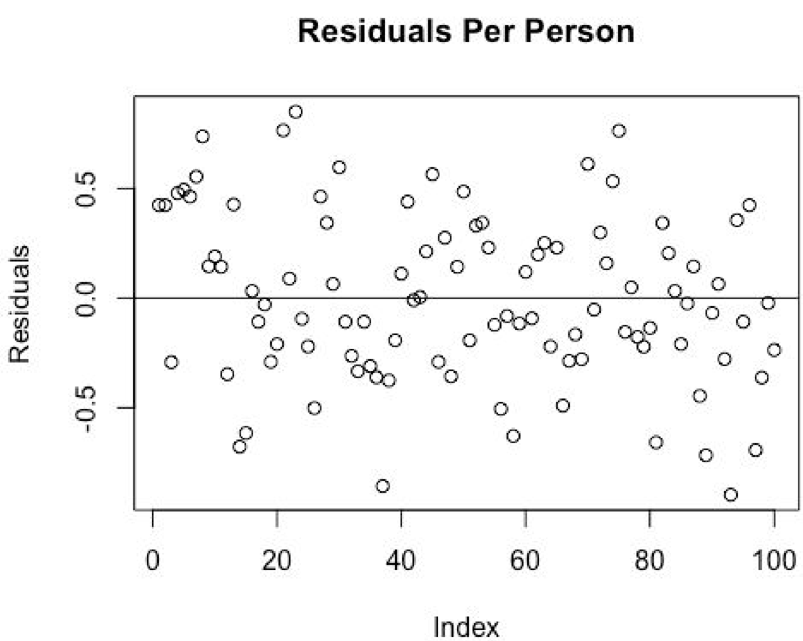

```
>fit0<-lm(happy~gender+workhrs+relationship,data=projdata)
> summary(fit0)

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

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

> fit1<-lm(happy~.^2,data=projdata)
> summary(fit1)

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(fit0,fit1)
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
> fitfinal<-lm(happy~gender*relationship+workhrs,data=projdata)
> summary(fitfinal)

Call:
lm(formula = happy ~ gender * relationship + workhrs, 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 ***
gender               0.178353   0.171396   1.041    0.301
relationship         0.352098   0.019935  17.662  < 2e-16 ***
workhrs             -0.070259   0.007978  -8.807 5.85e-14 ***
gender:relationship  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
```

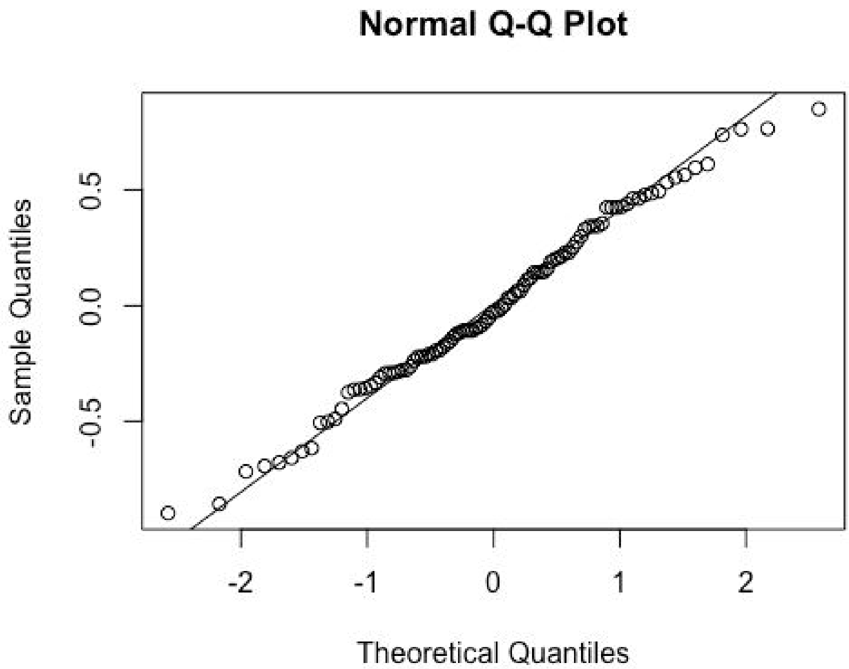
```
>plot(fitted(fitfinal),residuals(fitfinal),xlab="Fitted",ylab="Residuals",main=
"Residuals Plot")
> abline(h=0)
```



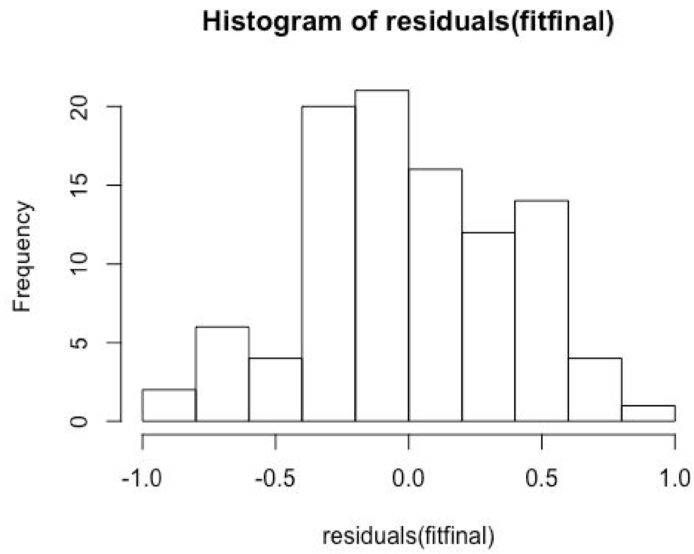

```
>plot(residuals(fitfinal),ylab="Residuals",main="Residuals Per Person")
>abline(h=0)
```



```
> qqnorm(residuals(fitfinal))
> qqline(residuals(fitfinal))
```




```
> hist(residuals(fitfinal))
```




```
> null=lm(happy~1,data=projdata)
> full=lm(happy~.^2,data=projdata)
> 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 = projdata)


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

> 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
            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 = projdata)
```

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