Sriharsha Addepalli
Micheal Alexander
Section: Wednesday 3 PM, Wenyi Fu

Project Report

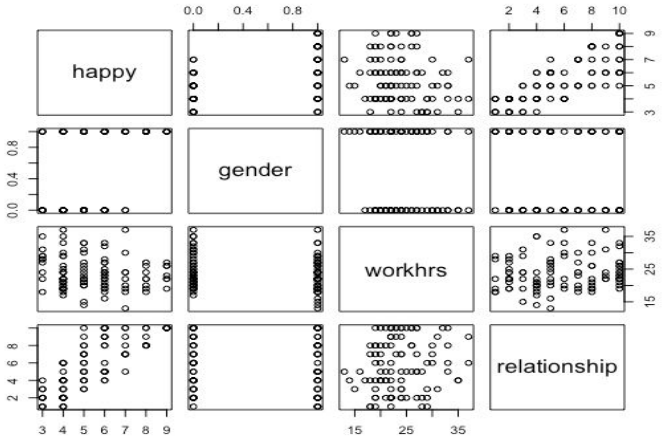
Introduction

The dataset used for this project is comprised of 100 volunteers. We are trying to deduce whether happiness can be explained by gender, hours of work, and satisfaction of one's relationship. We predict that men and women are equally happy. We also predict that there is a negative, linear relationship between happiness and the number of hours worked. Lastly, we predict that there is a positive, linear relationship between happiness and the quality of one's romantic relationship. Overall, we predict that the first order linear model would be significant.

Method

First, we created a scatterplot matrix to gain a holistic view of the data and observe any relationships between the predictors and the outcome variable. Then we fit the first order linear model which contained all the predictors. Next, we evaluated the assumptions and violations of the model using a residual plot, QQ plot for residuals, and a histogram of the residuals. We proceeded to use the extra sum of squares test to see if the model with interactions was more significant than the first order linear model. We then used stepwise regression (with both forward addition and backwards elimination) to arrive at the final model. We also created interaction plots between the predictors to confirm the final model was accurate. We then tested the overall significance of the final model. Lastly, we checked for violations of the final model as well by using the same tests as before.

Results



From the scatterplot matrix of this data, it appears that females have a greater average happiness score than males, whereas we predicted the scores would be equal. Also, it appears there is no relationship between work hours and happiness, contradicting our initial prediction that work hours would decrease one's happiness score. Finally, there is a positive, moderate, linear relationship between relationship and happiness, which we did expect.

The null hypothesis we were testing was that all the slopes were 0 against the alternative hypothesis that at least one slope was non zero.

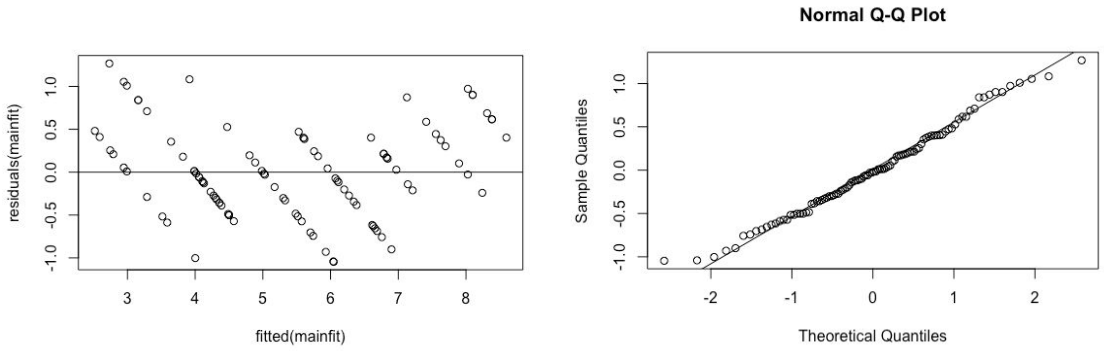
$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$
$$H_1 : \textit{At least one } \beta_1, \beta_2, \beta_3 \neq 0$$

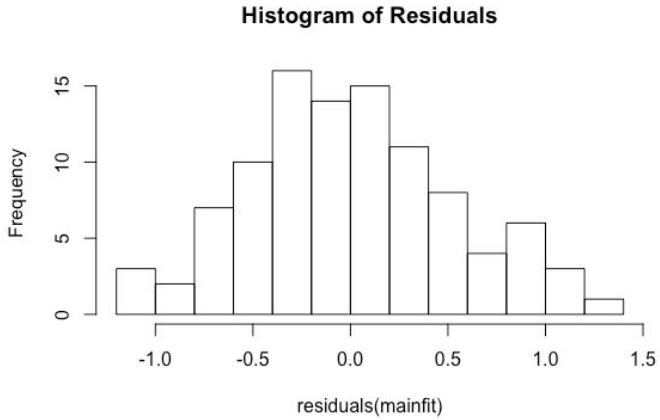
```
Call:
lm(formula = happy ~ relationship + gender + workhrs, data = projdata)

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 ***
relationship  0.48538    0.01821  26.649  < 2e-16 ***
gender        1.55447    0.10700  14.528  < 2e-16 ***
workhrs      -0.07118    0.01082  -6.576 2.52e-09 ***
---
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
```

This is the summary output of the first order model. The regression equation for this model was

$$Y = 3.54123 + 0.48538X_1 + 1.55447X_2 - 0.07118X_3$$

The F-statistic for this test is 312.2 and the overall p-value is 2.2e-16 which is less than 0.05, so we reject the null hypothesis. Thus we conclude at least one of the slopes is significant. There seems to be a relationship between the predictors and happiness.



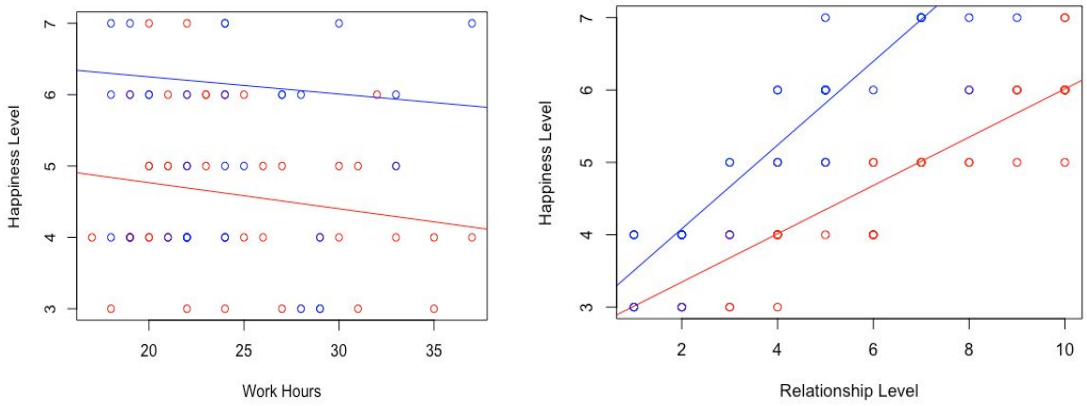


The plot in the top left corner is a plot of the fitted values of our first order model versus the residuals. It appears there is a parabolic pattern to the residuals, which suggests the fitted data is nonlinear. It appears that there is a constant variance. Next is the normal QQ plot. All of the values are close to the QQ line or on it which means the data is normal. The histograms of the residuals supports this conclusion as well.

We then wanted to test if adding interaction terms to the model added any significance. We compared this to the first order model using ANOVA/extra sum of squares test.

```
Analysis of Variance Table

Model 1: happy ~ (gender + workhrs + relationship)^2
Model 2: happy ~ relationship + workhrs + gender
  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
```

Because the p-value is 1.047e-12 < 0.05, this suggests that adding the interaction terms created a more significant model than the original first order model. To see which interactions were significant, we created interaction plots between all three predictors.




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

Residuals:
    Min      1Q  Median      3Q     Max
-1.8123 -0.8462  0.1572  0.7625  1.8270

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.426937   1.107963   3.996 0.000127 ***
relationship        0.554647   0.182315   3.042 0.003028 **
workhrs            -0.073661   0.046836  -1.573 0.119068
relationship:workhrs -0.003006   0.007542  -0.399 0.691108
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9475 on 96 degrees of freedom
Multiple R-squared:  0.7032,    Adjusted R-squared:  0.6939
F-statistic:  75.8 on 3 and 96 DF,  p-value: < 2.2e-16
```

The plot in the top left corner is an interaction plot between gender and work hours. Since the two slopes do not intersect, there is no interaction between these variables. The second plot is an interaction plot between relationship and gender. The two lines intersect, which means there is interaction present between these two predictors. The final test was testing for interaction between relationship and work hours. By looking at the estimate for relationship and work hours, we see that for everyone one unit increase in work hours, slope for relationship decreases by

0.003006. This illustrates that work hours has a negligible effect on relationship. This is also confirmed by the non-significant p-value of 0.691108.

Using stepwise regression, we arrived at the following model:

```
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
```

```
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
```

To ensure that this model was significant, we ran a hypothesis test with the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_1 : \textit{At least one } \beta_1, \beta_2, \beta_3, \beta_4 \neq 0$$

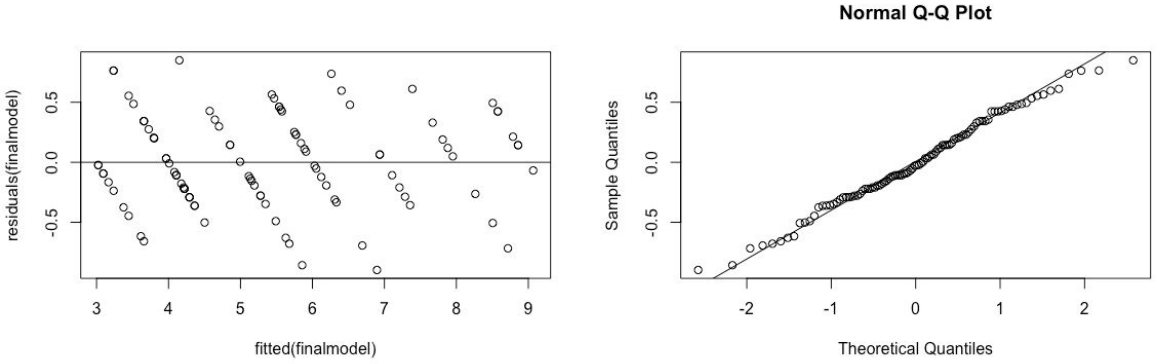
Because the p-value is 2.2e-16 which is less than 0.05, we can reject the null hypothesis, this implies at least one of the slopes is not zero. There is a relationship between happiness and relationship, gender, work hours, and the interaction between relationship and gender, the overall model is significant.

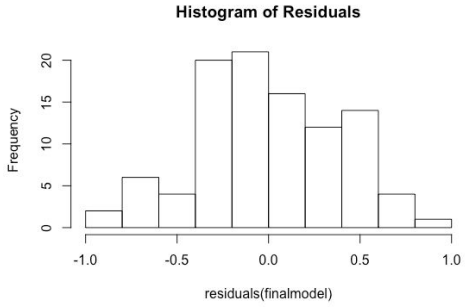
The partial p-values of relationship, work hours, and interaction between relationship and gender being less than 0.05 shows us that their effect individually given all the other predictors in the model are significant. The partial p-value of 0.301 in gender shows that the mean happiness between males and females is not very different given all the other predictors. However, gender is still an important predictor because the interaction term between relationship and gender is significant.

$$Y = 4.28774 + 0.35210X_1 - 0.17835X_2 - 0.07026X_3 + 0.24158X_1X_2$$

This model is the best fitting model. Happiness can best be described by relationship, gender, work hours, and the interaction between relationship and gender.

We predict that a male who has a zero relationship level and has zero work hours has a happiness level of 4.287745. We predict that for males, each additional relationship level will increase their happiness level by 0.352098. We predict that a female who has a zero relationship level and has zero work hours has a happiness level that is 0.178353 greater than males. We predict that for males, each additional work hour decreases their happiness by 0.070259. We predict that an additional level of relationship will increase the happiness level for a female by 0.241580 more than the increase of a male. The coefficient of determination of the model is 0.95, so 95% of the variance in happiness can be explained by knowing the relationship, work hours, gender, and the interaction between relationship and gender.





The residuals plot shows that the fitted data is linear and has constant variance. The points on the QQ plot are all along the QQline, so the fitted data appears to be normally distributed. The histogram confirms the normality. There does not seem to be any outliers so we can safely assume that the data is from the same population.

Discussion

As stated before, happiness can best be described by relationship, gender, work hours, and the interaction between gender and relationship:

$$Y = 4.28774 + 0.35210X_1 - 0.17835X_2 - 0.07026X_3 + 0.24158X_1X_2$$

Overall, the results did fit our predictions well. We did believe that the predictors would have an overall relationship with happiness. We also did predict that relationship would have a significant effect on happiness and that gender would not have a significant effect, which in general was the case in this final model.

One limitation of the model is that we cannot imply that the predictors cause happiness. Also, the model does not account for all of the factors that could contribute to happiness. The test was conducted with only 100 volunteers and the results would be more accurate given a larger sample size. Lastly, since the individuals were volunteers, the test is susceptible to bias.

In the future, this test could be conducted to observe the effects of the predictors on a different emotion other than happiness. We might also want to run another independent study to confirm our final model.

Appendix

```
projdata <- read.table(file.choose(),header=TRUE)
head(projdata)
attach(projdata)

#Scatterplot matrix##
pairs(happy~gender+workhrs+relationship,data=projdata)

##1st order linear model##
mainfit<-lm(happy~relationship+gender+workhrs,data=projdata)
summary(mainfit)
anova(mainfit)

##violations of first order##
plot(fitted(mainfit),residuals(mainfit))
abline(h=0)
qqnorm(residuals(mainfit))
qqline(residuals(mainfit))
hist(residuals(mainfit),breaks=9,main='Histogram of Residuals')

#interactions
interactionfit<-lm(happy~.^2,data=projdata)
summary(interactionfit)
```

```
anova(interactionfit,mainfit)

##interaction plots##
plot(workhrs[gender==0],happy[gender==0], xlab='Work Hours', ylab='Happiness Level',col="red")
abline(lm(happy[gender==0]~workhrs[gender==0], data=projdata), col="red")
points(workhrs[gender==1], happy[gender==1], col="blue")
abline(lm(happy[gender==1]~workhrs[gender==1], data=projdata), col="blue")

plot(relationship[gender==0], happy[gender==0], xlab= 'Relationship Level', ylab= 'Happiness Level',col="red")
abline(lm(happy[gender==0]~relationship[gender==0], data=projdata), col="red")
points(relationship[gender==1], happy[gender==1], col="blue")
abline(lm(happy[gender==1]~relationship[gender==1], data=projdata), col="blue")

interactiontest<-lm(formula=happy~relationship*workhrs)
summary(interactiontest)

##stepwise regression##

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

step(null,scope=list(lower=null,upper=full),direction='forward')
step(full,direction='backward')
step(null,scope=list(upper=full),direction='both')

### final model = lm(formula = happy ~ relationship + gender + workhrs + relationship:gender, data = projdata) ###


##final model##
finalmodel<-lm(happy~relationship+gender+workhrs+relationship*gender,data=projdata)
summary(finalmodel)
anova(finalmodel)

##model violations##
plot(fitted(finalmodel),residuals(finalmodel))
abline(h=0)
qqnorm(residuals(finalmodel))
qqline(residuals(finalmodel))
hist(residuals(finalmodel),main='Histogram of Residuals')
```