Project Report

Sara Bey, Jayanth Sivakumar, Wei Yang

June 8, 2019

Contents

List of Tables

List of Figures

1 Introduction

What factors contribute most to our happiness? Is it money, sex, love, or work? Or, is it a combination of these factors? And, how do each of these factors, and combinations of them, contribute to our happiness? Beyond a certain point, does any additional income generate more happiness? Or, is it possible that having too much money can negatively effect our rating of happiness?

In order to investigate these ideas, we will us the data set *happy*, in the faraway package to carry out our analysis. This data set is a collection of data from 39 students in the MBA class at University of Chicago. Data about the students' happiness, money, sexual activity, love and work are provided. Each student rates their happiness on a scale of 1 to 10. *Money* is measured by looking at family income of the student in thousands of dollars. *Sex* is measured by responding"1" for "satisfactory sexual activity" and "0" for "not" satisfactory sexual activity. *Love* is measured by responding "1" for "lonely", "2" for "secure relationships" or "3" for "deep feeling of belonging and caring". *Work* is measured on a 5 point scale with "1" representing "no job", "3" representing "OK job" and "5" representing "great job". Overall, there are 39 observations with 5 variables. It is important to note that it is not possible for someone to have a rating for their happiness of 0 or 1.

We hypothesize that, on average, people are happier when they have more money, satisfactory sexual activity, deeper feelings of belonging and caring and a great job. However, past a certain earning, money has a diminishing effect on that happiness. Our rationale for investigating these questions, with this dataset in mind, it to see if we can find evidence to support the statement that "money can't buy happiness". We want to discover if there is evidence to suggest that having more money does in fact have an association with a higher rating of happiness. We will investigate this relationship between happiness and each of the other four variables, and combinations of them. We believe that there will be a strong correlation between *money* and *work*, as those who have a great job tend to be earning more. We also believe that past a certain point, the effect on happiness due to money may not be as great. This cold be that there is a certain level of income that someone is no longer worried about money, but more so about the other variables, such as close relationships, amongst other factors. We conjecture that there is a strong relationship between *love* and *sex* since someone in a sexually active relationship tends to have deeper connections and have a stronger relationship.

2 Methodology

In order to find the best fit model for predicting our happiness from the data give in the set *happy*, we use a variety of tests and criterions. We begin by investigating the full model, using all the predictors of *love, sex, work* and *money* to predict *happy*. We then proceed to look into which variables will be significant in this prediction. We do so by using F-tests and hypothesis tests for the different variables. We then carry out some model and subset selection processes to verify the results from our testing. We can back up these results with goodness of fit criterion, such as the Akaike information Criterion (AIC) and Mallow's CP (C_p) (Dang 2018, slide 6-8, 15). These tests will show us how many predictors we should include in our model to get the most accurate predictions. From here we will try to see if different transformations of each variable can produce a model that better fits the observation form the data.

3 Results and Discussion

In order to carry out our testing, we first wanted to get an idea of how many people reported happiness at each level 1 - 10.

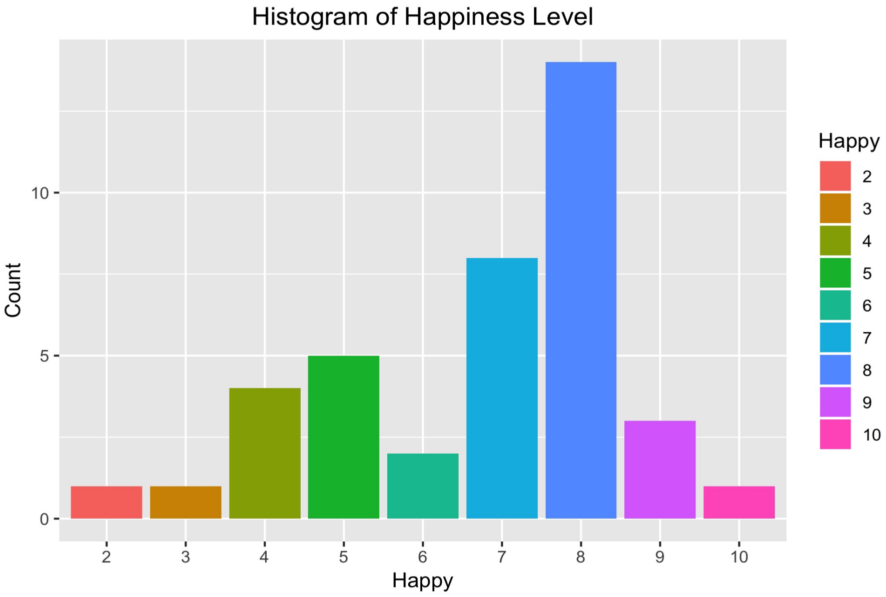


Figure 1: Overall Happiness Rating

From aFigure refoverall, we see that most people have an *happy* level greater than or equal to 7, i.e. people are generally happy.

Then, we wanted to look at the breakdown of how many people gave each rating for the remaining four variables provided in the data set; *money, sex, love*, and *work*.

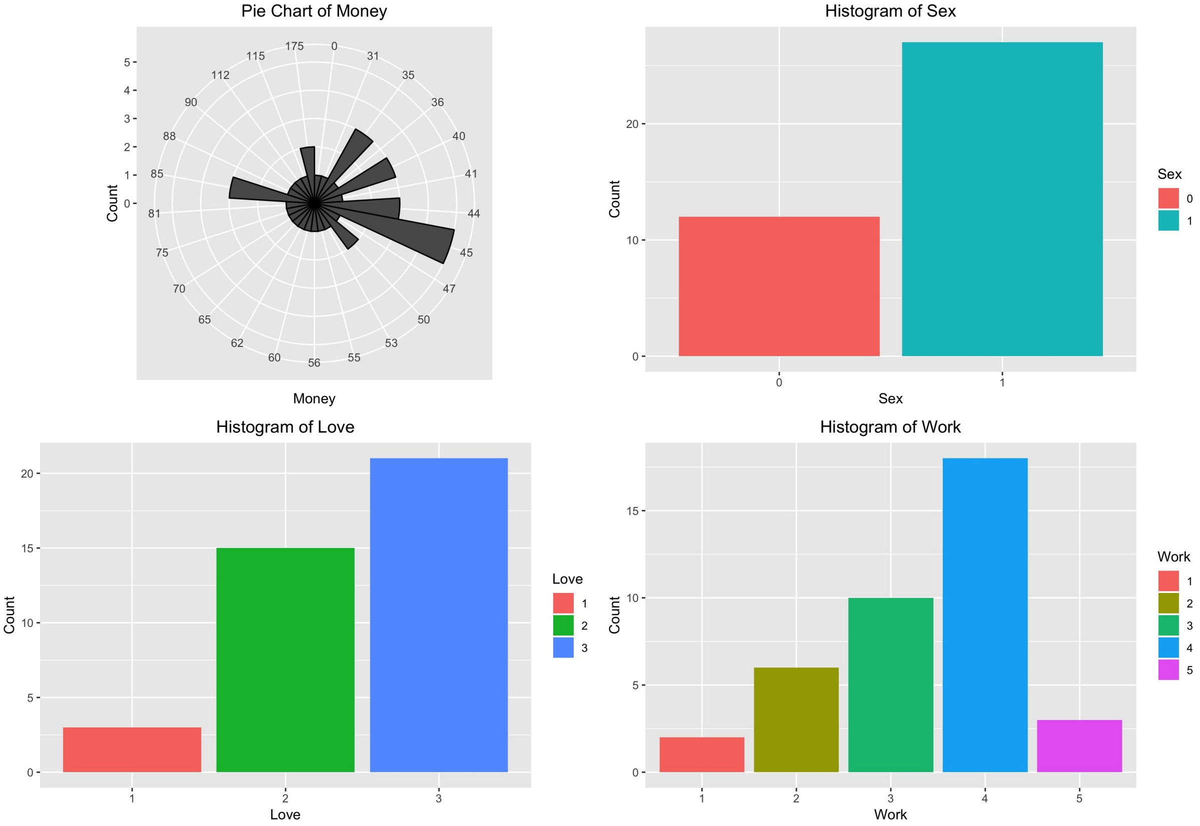


Figure 2: Histograms

Looking at Figure **??**. In the upper left histogram we can see most people have an income around $\$45k$.

In the upper right histogram, we can see that there are more people have satisfactory sexual activity, denoted by "1", than not, denoted by "0". In the lower left, we can see very few people consider themselves lonely, denote by "1", most people said that they have secure relationships, denoted by "2", or have a deep feeling of belonging and caring, denoted by "3". In the lower right we can see that very few people have no job, denoted by "1", most people are in the range of having an Ok job to having a great job, denoted by "3" and "5", respectively.

We then wanted to gain further insight into the relationship between the predictors themselves.

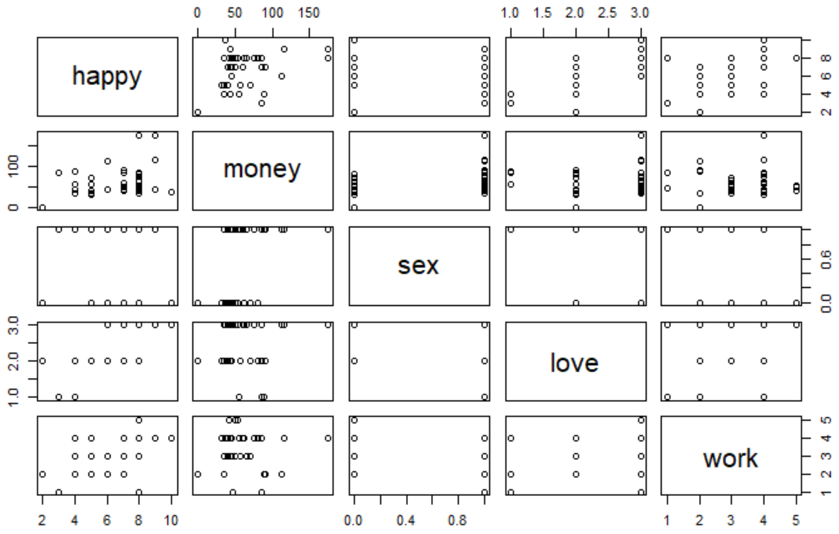


Figure 3: Pairwise Scatter plots

From the scatter plot, we see that there is no clear pairwise linear trend among the variables.

```
cor(happy)
      happy money    sex  love   work
happy  1.000 0.271 -0.033 0.784  0.539
money  0.271 1.000  0.307 0.126  0.068
sex   -0.033 0.307  1.000 0.047 -0.316
love   0.784 0.126  0.047 1.000  0.386
work   0.539 0.068 -0.316 0.386  1.000
```

```
vif(happy)
happy    money     sex      love     work
3.450936 1.257565 1.303891 2.740516 1.615127
```

The small covariance among the explanatory variables and the fact that all VIF values are less than 3 further implies that there is no multicollinearity among the variables.

Hence we starts our investigation with the full model that has *happy* as the response variable and *money, sex, love* and *work* as the explanatory variables. However, looking at the model,

```
lm(formula = happy ~ ., data = happy)
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.072081   0.852543  -0.085   0.9331
money        0.009578   0.005213   1.837   0.0749 .
sex         -0.149008   0.418525  -0.356   0.7240
love         1.919279   0.295451   6.496 1.97e-07 ***
work         0.476079   0.199389   2.388   0.0227 *

Multiple R-squared:  0.7102,Adjusted R-squared:  0.6761
```

we see that there is a small adjusted R^2 of 0.6761, and only 2 variables, *love* and *work*, are significant at a significance level of $\alpha = 0.05$. Thus, we have to modify our model. Based on the covariance matrix, we can see that the variable *sex* appears to have a very small correlation, close to 0, with the variable *happy.* We wanted to investigate the relationship between the two. Using an F-test, we tested the hypothesis that $\beta_{sex} = 0$. Our full model predicts happiness based on the remaining four variables, where as our reduced model predicts happiness without including *sex* in the model.

```
Model 1: happy ~ money + love + work
Model 2: happy ~ money + sex + love + work
Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     35 38.229
2     34 38.087  1     0.142 0.1268  0.724
```

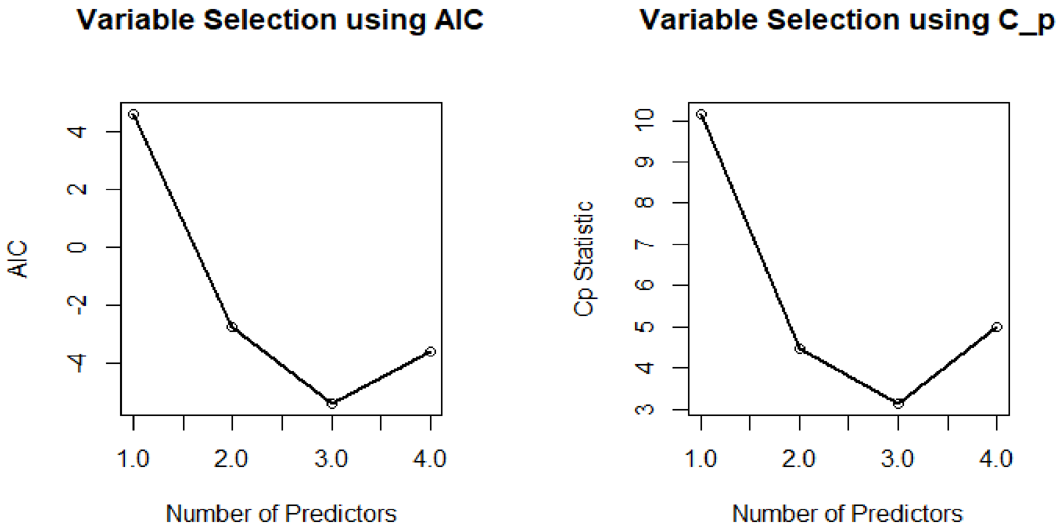
Based on the F-test, we got a P-value of 0.724, and thus, at an α level of 0.05, we failed to reject the null hypothesis that $\beta_{sex} = 0$. Hence, there is evidence to suggest that our rating for *sex* may not have an effect on our happiness. Further investigation can be done to determine whether or not *sex* is insignificant in our predictions for happiness.

We proceed by investigating whether or not we should include *sex* into our model using subset selection with the Akaike information Criterion (AIC) and Mallow's CP (C_p) (Dang 2018, slide 6-8).

We select the best model for each size from 1 to 4, and observed that for size $n = 3$, the model without *sex* is the best model.

```
regsubsets.formula(happy ~ ., data = happy)
1 subsets of each size up to 4
Selection Algorithm: exhaustive
money sex love work
1  ( 1 ) " "   " " "*"  " "
2  ( 1 ) " "   " " "*"  "*"
3  ( 1 ) "*"   " " "*"  "*"
4  ( 1 ) "*"   "*" "*"  "*"
```

It turns out that the model without *sex* minimizes both AIC and C_p (Dang 2018, slide 15).



Variable Selection using AIC

Variable Selection using C_p

Figure 4: Variable Selection using AIC and CP

Hence we should remove *sex* from our model.

Using *money, love and work* as our predictors for happiness we get the following model:

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.185936   0.780372  -0.238   0.8131
money        0.008959   0.004852   1.846   0.0733 .
love         1.901709   0.287644   6.611 1.22e-07 ***
work         0.503602   0.181486   2.775   0.0088 **
```

```
Multiple R-squared:  0.7091,Adjusted R-squared:  0.6842
```

$$Happy = 0.008959 * money + 1.901709 * love + 0.503602 * work - 0.185936 + \epsilon$$

The intercept term is -0.185936, which is the estimated value for *happy* when *money, love*, and *work* are all 0. But, *happy* takes value between 0 and 10, hence it does not make sense for us to have a negative intercept in our model. In fact, if *money, love* and *work* are all 0, assuming we have included all significant predictors, *happy* is expected to be 0. In other words, a model without the intercept term is more reasonable. To verify our hypothesis, we use nested F-test with hypothesis

$$H_0 : \beta_0 = 0 \qquad v.s. \qquad H_a : \beta_0 \neq 0.$$

```
Analysis of Variance Table
Model 1: happy ~ money + love + work - 1
Model 2: happy ~ money + love + work
Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     36 38.291
2     35 38.229  1  0.062009 0.0568 0.8131
```

With a p-value of $0.8131 \geq 0.05$, we fail to reject H_0, thus supporting the alternative hypothesis for a reduced model where $\beta_0 = 0$. Now we fit a model for *happy* based on *money, love* and *work*, without an intercept term.

```
lm(formula = happy ~ money + love + work - 1, data = happy)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
money 0.008644   0.004608   1.876  0.06879 .
love  1.862795   0.233659   7.972 1.82e-09 ***
work  0.485124   0.161919   2.996  0.00493 **
---

Multiple R-squared:  0.9799,Adjusted R-squared:  0.9782
```

We can see from above that at a significance level of $\alpha = 0.05$, *money* is not a significant predictor. We proceed by removing *money* from the model in which we previously removed the intercept and *sex* as a predictor.

```
lm(formula = happy ~ love + work - 1, data = happy)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
love   2.0081     0.2278   8.814 1.27e-10 ***
work   0.5330     0.1652   3.225  0.00263 **
---

Multiple R-squared:  0.9779,Adjusted R-squared:  0.9767
```

Removing *money* gives us a simpler model, but only leads to a small reduction in R^2. We proceed by carrying our further investigation and analysis to determine whether or not we should keep *money* in our model.

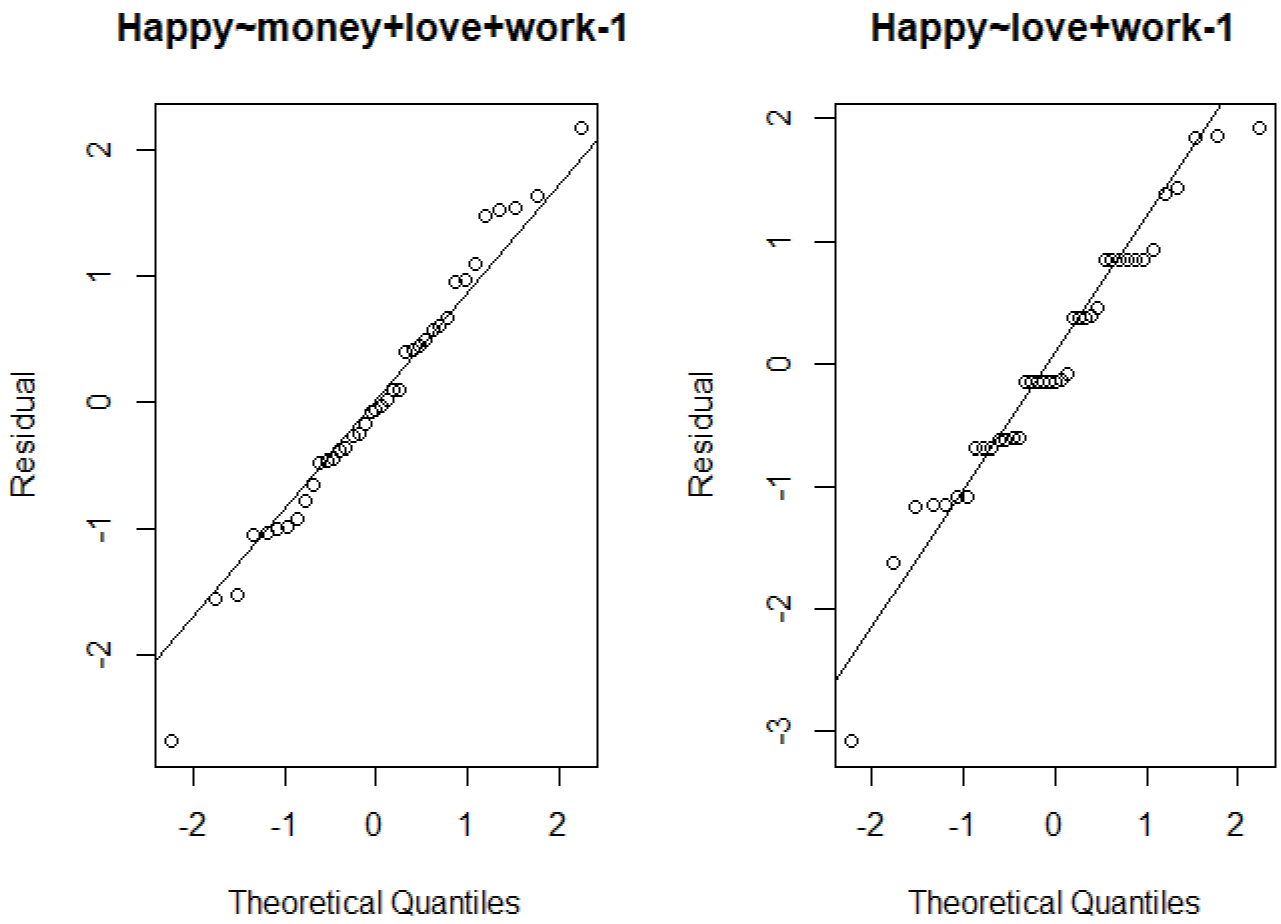


Figure 5: QQ plot for Happy love+work-1

Figure **??** indicates that the normality assumption for the error terms, $\epsilon_i's$, no longer holds if *money* is removed from the model. This supports that we should keep *money* in our model, but that a possible transformation is needed since *money* was found to be insignificant at the significance level $\alpha = 0.05$. *Money* needed to be included in our model as a predictor, but could be altered to create a model that better fit our data.

It is a well known fact from Economics that money has diminishing returns (Lane 2000). And the simple function $f(x) = \sqrt{x}$ is increasing and concave down on $[0, \infty)$, hence we may incorporate the idea of diminishing returns of money into our model by transforming the predictor *money* with the square root function.

```
lm(formula = happy ~ sqrt(money) + love + work - 1, data = happy)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
sqrt(money)  0.15263    0.06218   2.455   0.0191 *
love         1.73346    0.24128   7.184 1.88e-08 ***
work         0.40422    0.16369   2.469   0.0184 *

Residual standard error: 1 on 36 degrees of freedom
Multiple R-squared:  0.9811,Adjusted R-squared:  0.9795
F-statistic: 622.9 on 3 and 36 DF,  p-value: < 2.2e-16

AIC(all,nosex,nointsex,nointsexmoney,rootmoney)
df      AIC
all             6 121.7534
nosex           5 119.8985
nointsex        4 117.9617
nointsexmoney   3 119.5990
rootmoney       4 115.5633
```

In taking the square root of *money* we were able to create a simple model which has the smallest AIC (115.5) among all the models that we have tried. All the variables contained in this model are significant and the model gives an adjusted R^2 value of 0.9795.

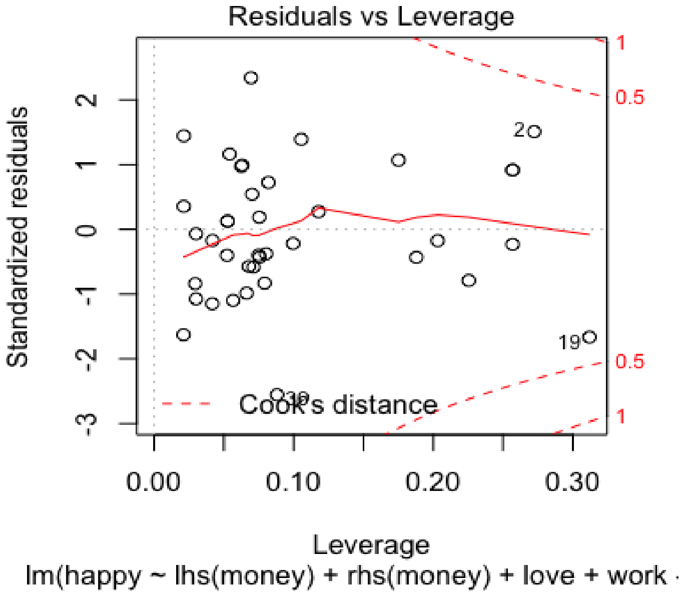


Figure 6: Residuals vs. Leverage in model with square root of money

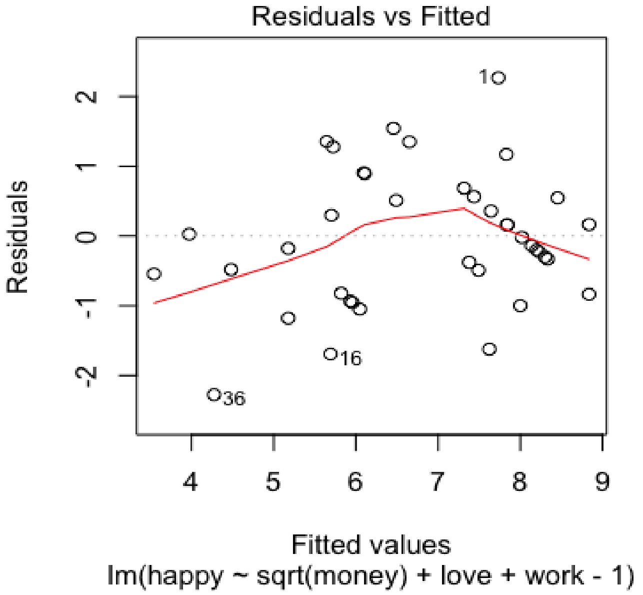


Figure 7: Residuals vs. Fitted in model with square root of money

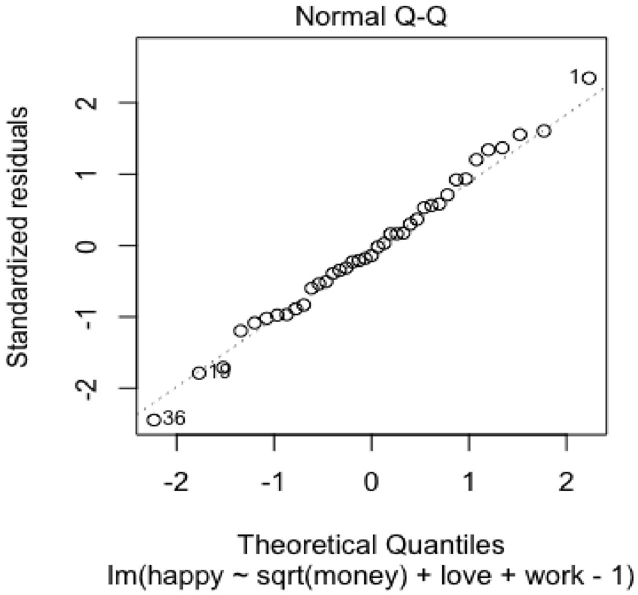


Figure 8: Normal QQ Plot for model with square root of money

Furthermore, in Figure ??, we can see that the residual vs. fitted values are randomly scattered, indicating that the constant variance assumption holds. In Figure ?? we can see that the Normal QQ plot is linear, indicating that the normality assumption of the residuals holds in the model that takes the square root of the predictor *money.* Thus

$$happy = 0.1526254 * \sqrt{money} + 1.7334612 * love + 0.4042174 * work + \epsilon$$

can be consider to be a good model.

In our model, *money* has a diminishing effect on happiness, and plays the least important role out of the included predictors. According to our model, earning $\$170k$ a year only contributes about 2 points to our overall rating of happiness. On the other hand, love plays the most important role in a person's happiness. Suppose a person has no job and makes no money. As long as that person has deep feeling of belonging and caring, our model indicates that this person can expect to be generally satisfied, with a rating for *happy* of 5.

Consider a person with median *money, love* and *work.* That is to say a person with an annual salary of $\$77.97k$, a person who has a deep feeling of belonging and caring (love=3) and a person having a more than ok job (work=4). Our model predicts that this person is going to have a happiness level of 7.89647 and we are 68% confident that his/her rating of happiness is in the interval

$$(6.947661, 8.845294)$$

and 95% confident that his/her happiness is in the interval

$$(5.821091, 9.971864).$$

We can do the same for a person whose statistics fall in the 0th, 25th, 75th or 100th quantiles.

```
quantile love work     money
0%       1    1      0.00000
25%      2    3      42.48676
50%      3    4      50.00000
75%      3    4      77.97114
100%     3    5      175.00000
predict(rootmoney,x0,interval="prediction",level=0.95)
fit        lwr       upr
0%   2.137679 0.07579235  4.199565
25%  5.674415 3.62628012  7.722550
50%  7.896478 5.82109083  9.971864
75%  8.164955 6.10069013 10.229219
100% 9.240514 7.10116518 11.3798641
```

Take, as an example, a person with stats in the 25% quantile (with secure relationships, an ok job, and a salary of \$42.48 k). Our model predicts with 95% confidence that this person will have a happiness rating in

$$(3.62628012, 7.722550).$$

This implies that even if a person makes only $\$42k$, there is still a chance that that person is feeling quiet happy, ($happy = 7$), given that he/she is in a secure relationship and has an ok job.

4 Conclusion

Based on the regression analysis that we have carried out, we have determined that our happiness can be predicted based on our *money, love* and *work*, but that our rating of *sex* in insignificant in making this prediction. The model that removes the intercept and the predictor *sex*, but takes the square root of *money* provides the best fit for predictions of *happy* for the observations given in the dataset.

5 Bibliography

1. Dang, S 2018, *Variable Selection*, lecture notes, Regression I MA531, Binghamton University, delivered 15, Oct 2018.

2. Lane, E 2000, 'Diminishing returns to income, companionship–and happiness', *Journal of Happiness Studies*, vol. 1, no.1, pp. 103-119.

6 Appendix A

1.
```
list.of.packages <- c("printr","Metrics","leaps","MASS","caret",
"gghalfnorm","glmnet","ModelMetrics","ISLR",
"ggplot2","dplyr","faraway","knitr","reshape2")

new.packages <- list.of.packages[!(
list.of.packages %in% installed.packages()[,"Package"])]

if(length(new.packages)) install.packages(new.packages)
```

2.
```
#Loading Libraies
library(faraway)
library(dplyr)
library(ggplot2)
library(ISLR)
library(knitr)
library(printr)
library(Metrics)
library(leaps)
library(MASS)
library(caret)
library(gghalfnorm)
library(glmnet)
library(ModelMetrics)
```

3.
```
#Loading Data
data(happy)
row = dim(happy)[1]
col = dim(happy)[2]

c("1"=0,table(happy$happy))
table(happy$money)
table(happy$sex)
table(happy$love)
table(happy$work)
```

4.
```
#Histogrram  for  Happinness
a <- ggplot(happy, aes(x = factor(happy),fill=factor(happy)))
a + geom_bar(aes(fill = factor(happy))) + xlab("Happy")+ylab("Count")+
labs(title = "Histogram of Happiness Level",fill = "Happy") +
theme(plot.title = element_text(hjust=0.5))
```

5.
```
#Pie Chart for money
cxc <- ggplot(happy, aes(x = factor(money))) +
geom_bar(width = 1, colour = "black")
cxc + coord_polar()+ xlab("Money")+ylab("Count")+
labs(title = "Pie Chart of Money") + theme(plot.title = element_text(hjust=0.5))
```

6.
```
#Histogram for  Sex
a <- ggplot(happy, aes(x=factor(sex),fill=factor(sex)))
a + geom_bar(aes(fill = factor(sex))) + xlab("Sex")+
ylab("Count")+labs(title = "Histogram of Sex",fill = "Sex") +
theme(plot.title = element_text(hjust=0.5))
```

7.
```
#Histogram for Love
```

```
a <- ggplot(happy, aes(x=factor(love),fill=factor(love)))
a + geom_bar(aes(fill = factor(love))) + xlab("Love")+ylab("Count")+
labs(title = "Histogram of Love",fill = "Love") +
theme(plot.title = element_text(hjust=0.5))
```

8.
```
#Histogram for  work
a <- ggplot(happy, aes(x=factor(work),fill=factor(work)))
a + geom_bar(aes(fill = factor(work))) + xlab("Work")+ylab("Count")+
labs(title = "Histogram of Work",fill = "Work") +
theme(plot.title = element_text(hjust=0.5))
```

9.
```
#pairwise scatter plot
pairs(happy,data=happy)
```

10.
```
#covariance matrix and VIF
round(cor(happy),3)
vif(all)
```

11.
```
#regression with all terms
all<-lm(happy˜.,data=happy) #all coef should be positive
summary(all)
```

12.
```
#regression without sex,
nosex<-lm(happy˜ money+love+work,data=happy)
summary(nosex)
anova(nosex,all)
```

13.
```
#subset selection
b <- regsubsets(happy˜.,data=happy)
summary(b)
rs <- summary(b)
```

14.
    ```
    #####Plot AIC
    par(mfrow=c(1,2))
    AIC <- 50*log(rs$rss/50) + (2:5)*2
    plot(AIC˜I(1:4),ylab="AIC",xlab="Number of Predictors",type="l",
    lwd=2,main = "Variable Selection using AIC")
    points(AIC)
    ```

15.
    ```
    ###Plot Cp
    plot(1:4,rs$cp,xlab="Number of Predictors",
    ylab="Cp Statistic",type="l",lwd=2,
    main = "Variable Selection using C_p")
    points(1:4,rs$cp)
    ```

16.
    ```
    #regression without sex,
    nosex<-lm(happy˜ money+love+work,data=happy)
    summary(nosex)
    ```

17.
    ```
    #regression without intercept and sex
    nointsex<-lm(happy˜ money+love+work-1,data=happy)
    summary(nointsex)
    anova(nointsex,nosex)
    ```

18.
    ```
    #regression without int,sex and money
    nointsexmoney<-lm(happy˜love+work-1,data=happy)
    summary(nointsexmoney)
    ```

19.
    ```
    #squareroot money
    rootmoney<-lm(happy˜ sqrt(money)+love+work-1,data=happy)
    summary(rootmoney))
    AIC(all,nosex,nointsex,nointsexmoney,rootmoney)
    ```

20.

```
#prediction
X <- model.matrix(rootmoney)
x0<-as.data.frame(apply(X,2,quantile))
x0$money<-x0$"sqrt(money)"^2
x0$"sqrt(money)"<-NULL
x0
betahat <- coef(rootmoney)
muhat<-x0%*%c(betahat)
muhat
n <- dim(X)[1]
p<-3
predict(rootmoney,x0,interval="prediction",level=0.95)
```