Group members: Joy Harjanto, Bella Lee, Bridget Lee, David Lu, Ciara Mandich, Selvam Sendhil
Professor Cha
Statistics 101A
Final Project

Significant Predictors of Happiness

Introduction

With an increase in empirical research on happiness, happiness is becoming a quantifiable phenomenon. However, determining which factors contribute to happiness is debatable. In this project, we are interested in developing a model that predicts happiness based on predictors such as age, gender, job satisfaction, and income. Using data from the National Opinion Research Center, we will perform various regression analyses to determine an optimal model for our study. We will also use power and square root transformation, inverse plot, and box-cox to optimize our model.

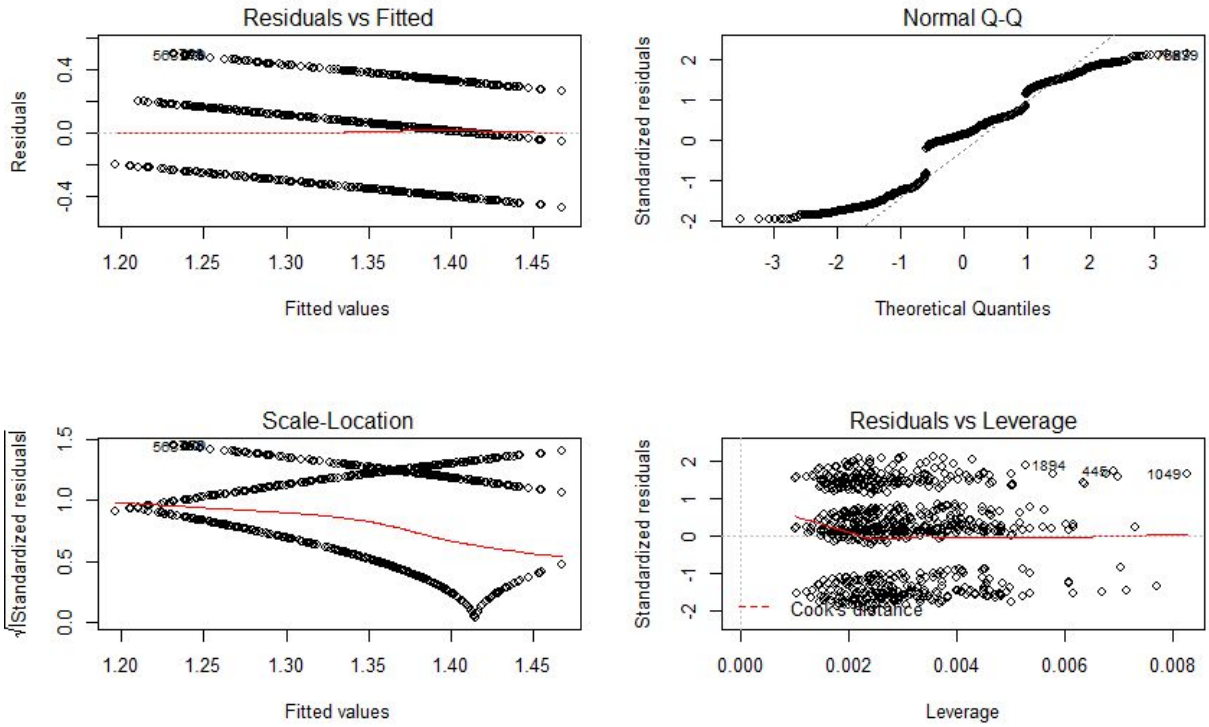
Methods and Procedures

We chose to use the X and Y Box Cox method to generate and choose our model. We compared the R-squared values as well as the plot diagnostics to the R-squared values and plot diagnostics of other models generated by different methods, and we came to the conclusion that the X and Y Box Cox was the best method to use. We also tried the Inverse Plot, the single variable Box Cox, and the square root of X, Y, and X and Y, but it did not give us a model that was as significant as the model determined by the X and Y Power Transform. The original model did not show very promising R-squared values or significant predictors, so we decided to use transformations. Before we continued with variable selection, we decided to remove OwnHome since every observation that we had contained the same number, so the values in the summary of the model were NA. Because Children and Education had zero values, we decided to add little value to the variable so the function could run. We did not want to remove the variables as they are potentially important predictors to explain Happiness. For variable selection, we tried the method with all possible subsets, forward selection, and backward selection to see which models and predictors had the lowest AIC and BIC scores. This was a necessary process due to the fact that the original model was not very significant. We did not use case omissions in order to keep the data as accurate as possible, but rather changed some values that were reported not answered as NA.

Results and Interpretation

The R results show that the number of dependent members in an individual's household, an individual's marital status, an individual's job satisfaction and an individual's health are all highly significant predictors of happiness levels. After exploring all possible subsets, we found the optimal model for all sizes. The AIC and AICc calculations suggested a model with 5 total variables (Household, Sex, Marital, JobSat and Health). The BIC results suggested a model with only 4 variables (Household, Marital, JobSat and Health) or one with 6 variables (Household, Sexf, Marital, JobSat, Workhrs and Health). The reason we chose the 5 variable model was because it had one of the highest adjusted R^2 values, all predictors besides Sex were significant, the F-statistic is significant, and forward and backward selection processes suggested that this was the best model.

The final model, one with 5 variables is shown below:

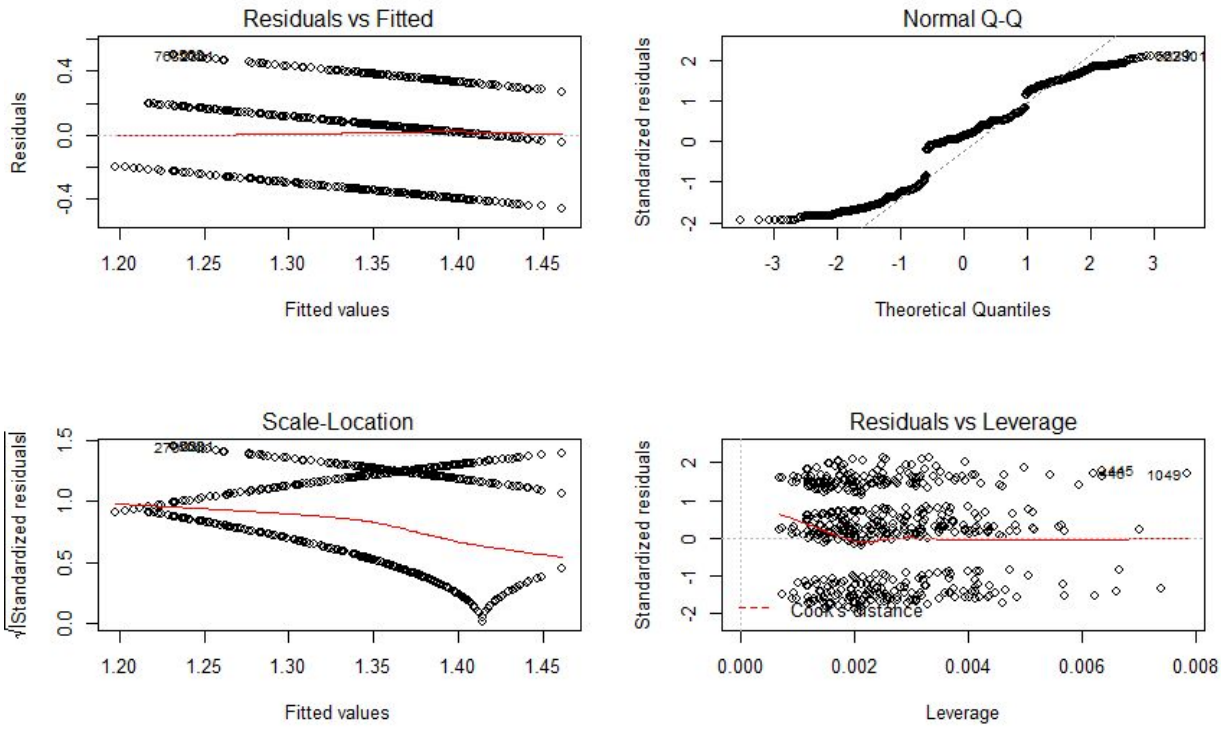



```
Call:
lm(formula = sqrt(Happiness$Happy) ~ log(Happiness$Household) +
    aMarital + log(Happiness$Sex) + aJobSat + sqrt(Happiness$Health))

Residuals:
     Min       1Q   Median       3Q      Max
-0.46722 -0.25254  0.02856  0.13023  0.50041

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.439146   0.035674  40.341  < 2e-16 ***
log(Happiness$Household) -0.037923   0.012023  -3.154  0.00163 **
aMarital                 -0.268903   0.028709  -9.367  < 2e-16 ***
log(Happiness$Sex)        0.019542   0.014224   1.374  0.16962
aJobSat                   0.053703   0.013367   4.017 6.07e-05 ***
sqrt(Happiness$Health)    0.020024   0.006277   3.190  0.00144 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2363 on 2348 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.06777,   Adjusted R-squared:  0.06578
F-statistic: 34.14 on 5 and 2348 DF,  p-value: < 2.2e-16
```

At a significance level of 0.05, we see that all variables are significant above with the exception of sex, which has a p-value significantly higher than 0.05. By far the most significant variable is job satisfaction, indicating that one's job satisfaction has a strong prediction on happiness. This model has an R^2 of

0.06578, higher than that of other models such as the 4-variable one. Additionally, both residuals graphs, Residuals vs Fitted and Residuals vs Leverage, show that the residuals have no apparent pattern. Additionally, the vast majority of points are within -2 and 2 standardized residuals from 0. These are all strong indicators of a healthy model. A potential limitation is the disjointed appearance of the Normal Q-Q plot, but the standardized residuals do somewhat generally follow a straight line. The Scale vs Location graph shows a strong horizontal line initially, and curves downwards towards the end, indicating another potential limitation.

For comparison, the 4 variable model is below:



```
Call:
lm(formula = sqrt(Happiness$Happy) ~ log(Happiness$Household) +
    aMarital + aJobSat + sqrt(Happiness$Health))

Residuals:
     Min       1Q   Median       3Q      Max
-0.46152 -0.24689  0.02844  0.13028  0.50082

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.444029   0.035503  40.673  < 2e-16 ***
log(Happiness$Household) -0.038637   0.012014  -3.216  0.00132 **
aMarital                 -0.267855   0.028704  -9.332  < 2e-16 ***
aJobSat                   0.055163   0.013328   4.139 3.61e-05 ***
sqrt(Happiness$Health)    0.019797   0.006276   3.155  0.00163 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2364 on 2349 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.06702,   Adjusted R-squared:  0.06543
F-statistic: 42.18 on 4 and 2349 DF,  p-value: < 2.2e-16
```

Here, with sex excluded, all variables are significant. This model shows a slightly lower adjusted R^2 of 0.06543. Job satisfaction is still the most significant predictor. Diagnostic plots for 4-variable model look almost identical to the 5-variable model. One slight difference is that the upward curve at the beginning of the Residuals vs Leverage plot starts slightly higher than it does for the 5-variable plot, showing that it could be slightly less ideal.

Discussion of Social Significance

A statistical analysis by two statisticians published in the MIT Technology Review analyzed the factors that affect an individual's happiness level. They utilized data from the 1972- 2012 General Social Survey as well and divided the variables into two groups: personal and macroeconomic indicators. Health was found to be the strongest determinant in the personal variables group, which was followed closely by marriage. Income, on the other hand, was found to have the least effect on happiness. The health variable was not included in our initial data. After running the model, we found both Marital and Income to be significant variables with Job Satisfaction as the most significant variable.

The results make sense in a real world context. We found income to be the least significant variable. Many social psychologists agree that income does not strongly predict happiness once an individual rises out of poverty. They also agree that spending money on experiences is more satisfying than spending on material items. Spending on leisure activities tends to make people less lonely, and builds social relationships. Additionally, spending money on others makes us happier than spending it on ourselves. Helping others enhances our social relationships, and correlates to a more positive view of ourselves.

Psychologists also agree that satisfying relationships and purposefulness are significant predictors of happiness levels (Gilovich et al., 219-225). Mortality, crime, and suicide rates are higher for divorced, unmarried, and widowed individuals. Married people have greater subjective well-being than unmarried people. Social support contributes to good health by strengthening the cardiovascular, immune, and endocrine systems. This has been shown in both correlational studies and in experiments that

manipulate perceived social isolation and assess immune system function. There appears to be a causal relationship between loneliness and compromised physical health (Gilovich et al., 385-390). In our data set, we found that marital status and job satisfaction were the single highest predictors of happiness.

Sociologically speaking, this project sheds light on important factors that contribute to one's happiness. As a social phenomenon, it is intriguing that happiness can be quantified as we see here from the data we are given. For future models that are designed to predict happiness, it would be unique and informative to observe cultural aspects of existence that may contribute to to our model as well. For instance, can we quantifiably assess empowerment from our cultural identities? How would such empowerment translate to happiness? These are interesting questions that perhaps could be later investigated by the National Opinion Research Center.

Limitations of the Project

However, in this project we faced several limitations. Many of the observations in the original dataset were converted into NAs, and were thus omitted from our analyses. Even though this was necessary in order to run any sort of analysis, this does mean that not all responses were accounted for. Also, we were unable to run a Power Transformation with the Children and Education variables, and because we were limited in techniques, we had to make the decision to take them out from our analyses. Since we developed an optimal model using the X and Y Box Cox, however, we are confident that the inclusion of these variables would not have made a significant difference, so they would have been excluded from our final model in the end. Nonetheless, statisticians with a fuller range of techniques should consider these variables for future models and attempt to account for a greater number of responses than we were able to in this project.

Works Cited

Gilovich, T. (2016). *Social Psychology.* New York, NY: W.W. Norton & Company.

MIT Technology Review. (2012, January 3). Statisticians Reveal What Makes America Happy. Retrieved from https://www.technologyreview.com/s/426529/statisticians-reveal-what-makes-america-happy/