Using Multiple Linear Regression to Predict Happiness

Amil Khan

December 27, 2017

Introduction

In this paper, we will explore the burning question of what makes people happy. The data we collected sampled 100 individuals with each rating their happiness on a 10-point scale (1 = very unhappy, 10 = very happy), their love relationship(s) on a 10-point scale (1 = very lonely, 10 = deeply in love), and the number of hours they work each week. We also noted their gender to determine whether men or women were significantly different in any way. Our main goal was to examine gender and determine whether a healthy relationship contributed more to a person's overall happiness.

Method

Examining Raw Data

We first examined the dataset to verify and, at the very least, convince ourselves that our theory could be plausible. There was a noticable influence on happiness when people rated their relationship higher.

We found that looking at *summary statistics* helped us better understand our data before creating any plots. This helped us determine if there were extreme values to look out for.

Exploratory Data Analysis

To visualize potential relationships, we produced *scatterplots.* This gave us the best look at the individual points of our dataset, as well as any trends that the points may be following. Naturally, we needed to plot the summary statistics, so we used boxplots.

Model Fitting

This was arguably the most fun and more artistic part of the study. Our first thought was to fit the full model and use it as a baseline for when we were evaluating higher order models with interaction terms. We told ourselves that we would use *Occam's Razor* and always choose the less complex model if the performed difference was not significantly justifiable.

Hence, we fit the full two-way interaction model and used p-values to evaluate model significance. Instead of using the contentious black box method first, we opted to add and drop terms ourselves. Once we were convinced that we found the best model—one that included only necessary interaction terms—we ran the stepwise regression algorithm in both directions.

Goodness of Fit Evaluation

We evaluated each model individually using the F-test and AIC (*Akaike's Information Criterion*). We did take into account the BIC (*Bayesian Information Criterion*), since our personal preference for model fitting was the least complex model.

To ensure the full model did not violate any assumptions, we examined diagnostic plots for non-constant variance, normality, and linearity. Similarly, we followed the same method with our best model—evaluate Residuals vs. Fitted, Normal QQ, Histogram of Residuals.

Results

We begin with providing summary statistics to guide us through the data exploration process. When examining the scatter plot, we will be able to approximate summary statistic values given the points on the graph.

	Happiness	Work Hours	Relationship
Min	3	13	1
Mean	5.42	23.76	5.69
Max	9	37	10
Variance	2.93	25.13	8.74

Table 1: Summary Statistics

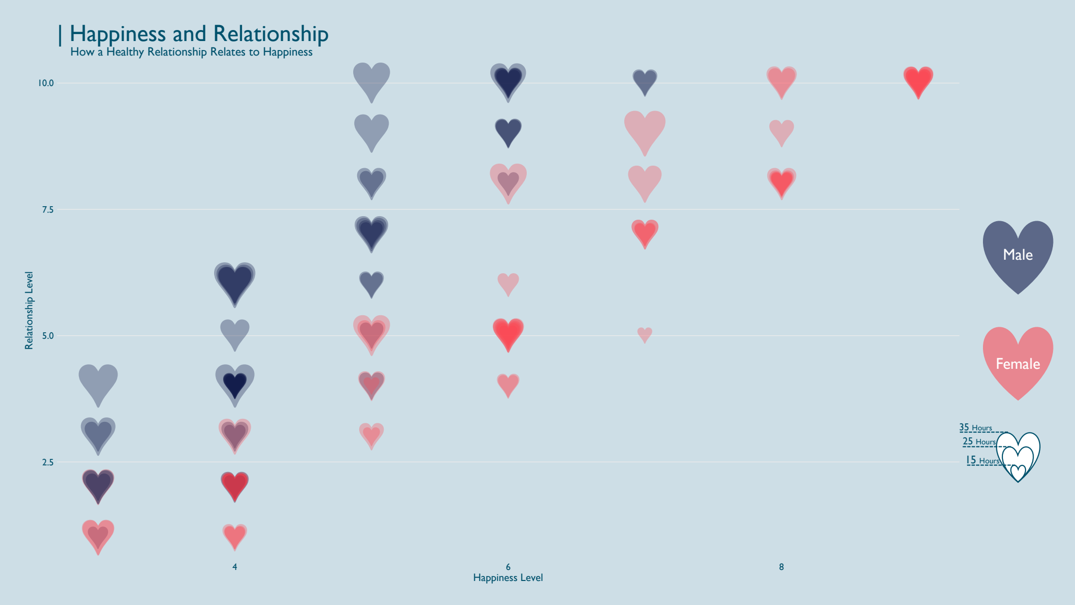

Figure 1: *This scatterplot shows the moderately strong relationship between happiness and relationship. Darker colors mean there are overlapping values, while the size of the hearts correspond to number of hours worked.*

Given our aforementioned summary statistics coupled with a scatterplot matrix, we were able to see that happiness and gender shared a moderately positive linear relationship. However, we noticed that relationship and happiness had a computed 79.6% positive correlation.

Full Model

$$\textbf{Happiness} = 3.54 + 1.55(\textit{Gender}) - 0.07(\textit{Work Hours}) + 0.48(\textit{Relationship})$$

After our exploratory data analysis, we fitted the full model and used it as a baseline for future model fitting. We were able to obtain a p-value of $<$ 2.2e-16, well below the 5% threshold, an AIC of -122.9669, and an R^2 able to explain 90.7% of the variance in the dataset. These are the numbers to exceed going forward.

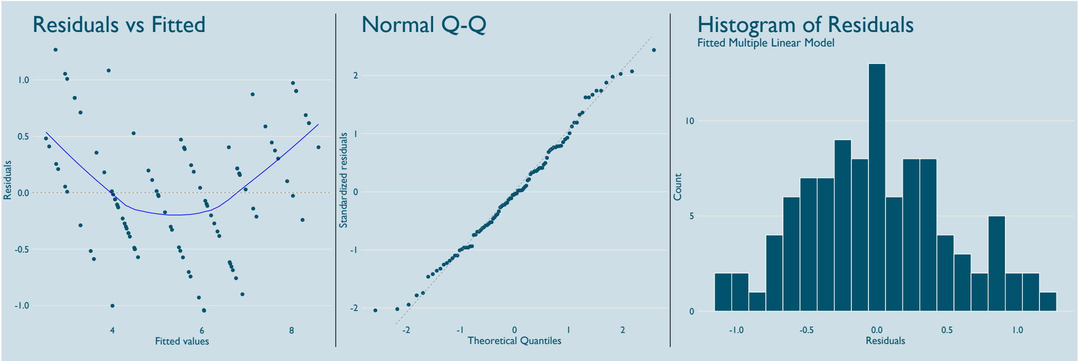

Figure 2: *In the first plot, we see some non-linearity. We argued that the QQ plot error terms are approximately normally distributed. And, finally, the third plot shows us the residuals are approximately normal, but with some signs of heavy tail because of the negative residuals.*

Our next step was to look for two way interactions between our predictor variables. Keeping our hypothesis in mind, we began by fitting all two way interactions and evaluated multiple models through the F-test and AIC. Our ultimate goal was to pick a model that was simple yet complex enough to predict new data.

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
$<$ *none* $>$			14.38	-179.90		
gender:workhrs	1	0.05	14.43	-181.57	0.31	0.5810
gender:relationship	1	12.41	26.80	-119.68	80.26	3.259e-14
workhrs:relationship	1	0.07	14.45	-181.42	0.45	0.5030

Table 2: Full Two-way Interaction Model

When we fit all of the two way interactions, we looked for small SSE, large p-values, and small AIC (most negative). We dropped the interaction between gender and hours worked since it had the largest p-value and smallest sum of squares. Next, we refit the model, ran another F-test, and decided to drop the interaction between hours worked and relationship. To ensure we arrived at the best model, we ran an ANOVA that included three models. The first model was our full model, and the other two were models with interactions.

```
Analysis of Variance Table

Model 1: happy ~ gender + work + relationship
Model 2: happy ~ relationship + gender:relationship + work:relationship
Model 3: happy ~ gender + work + relationship + gender * relationship
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     96 26.991
2     96 16.005  0   10.9860
3     95 14.506  1    1.4992 9.8188 0.002298 **
```

Best Fitting Model

$$\textbf{Happiness} =$$
$$4.29 + 0.18(\textit{Gender}) - 0.07(\textit{Work}) + 0.35(\textit{Relationship}) + 0.24(\textit{Gender:Relationship})$$

With this model, we predict that for every one unit increase in work hours, happiness goes down by 0.07 units. But for every one unit increase in relationship level, happiness goes up by 0.35 units. Now for our interaction term, we predict that an additional increase in relationship level will increase happiness for women by 0.24 more than the increase for men.

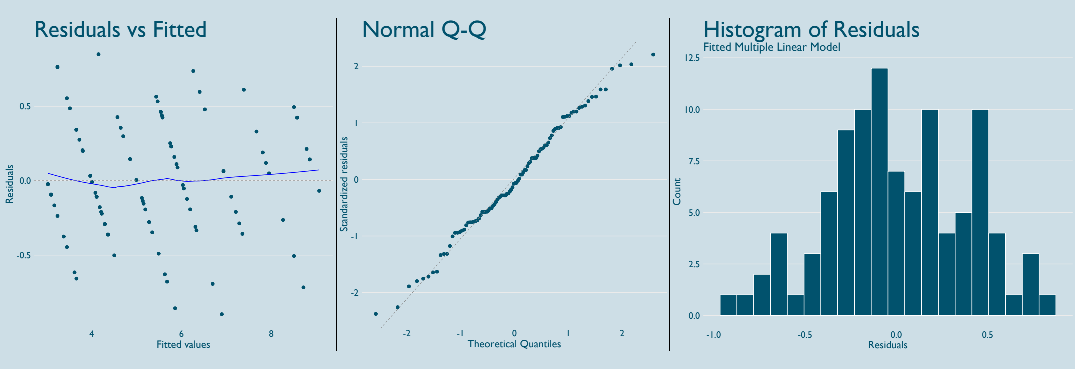

Figure 3: *In the first plot, we see an almost perfect example of linearity. The QQ plot error terms are approximately normally distributed. And, finally, the third plot shows us the residuals are approximately normal, but with some signs of heavy tail because of the positive residuals.*

Our diagnostic plots for our best model passes the tests for assumption violations—linearity, constant variance, normality. We will take a look at the summary output of the model for completeness.

```
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
workhrs            -0.070259   0.007978  -8.807 5.85e-14 ***
relationship        0.352098   0.019935  17.662  < 2e-16 ***
gender:relationship 0.241580   0.026716   9.043 1.84e-14 ***
---
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
```

Our best fit model had a p-value $<$ 2.2e-16, and an R^2 that enabled us to explain 95% of the variance in our data. All of the predictor variables were significant except gender. We were able to obtain an AIC of -183.06, which was the lowest. To verify our model was the cream of the crop, we used *stepwise regression* in both directions.

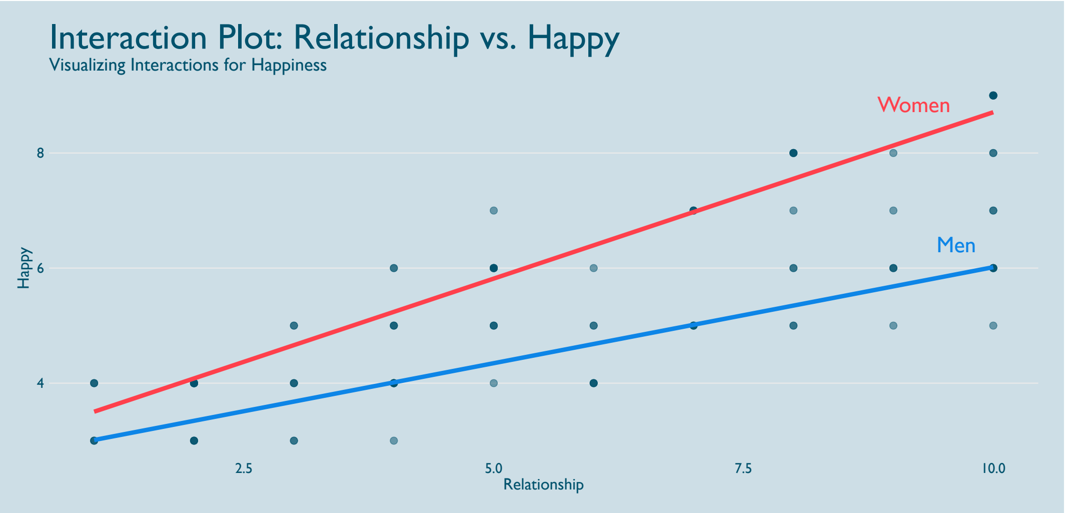

Figure 4: *Based on our interaction plot, women tend to be more happier than men when they are in a healthy relationship.*

Conclusion

Our research question began with whether men or women in a healthy relationship were happier. Given our best model, we discovered that women were happier when in a great relationship. We also discovered that both men and women had a decrease in happiness when they worked more. Although these conclusions may seem small, they lend themselves as a foundation for further research. Our next step will incorporate another dataset to help further quantify happiness. Predictor variables of interest are income, education level, location, and length of longest relationship.