PSTAT 126 Final Project

Lauren Wong and Alison Do
Jinwen Qiu, Monday 3pm
June 6, 2018

Introduction

The main question of this report is: does a model with more predictors fit the data significantly better than a model with less predictors regarding an individual's happiness? The data set used was gathered from a survey of 100 volunteers to rate their happiness on a 10-point scale based on: gender, number of hours they work each week, and the quality of their love relationship. Each contributor's happiness was calculated in the following point system: 1 = very unhappy, 10 = very happy; each contributor's gender was categorized as: 0 = male, 1 = female; and the quality of the volunteer's love relationship was valued in a point-system as 1 = very lonely, 10 = deeply in love. We predict that a model with more predictors will fit the data better and play a significant role in one's happiness.

Method

In order to find the most appropriate regression model, we first identified the full linear regression model by observing the scatterplot matrix between the outcome variable and all the predictors. From there, we observed the basic relationships between happiness versus gender, happiness versus work hours, and happiness versus relationship. Then we determined the null hypothesis, where each predictor has a slope of zero and is therefore insignificant. The alternative hypothesis on the other hand, describes how there is at least one predictor that does not have a slope of zero. Through the summary function, we focused on the p-value of each predictor variable to write up the first order linear model and decide the conclusions of the hypotheses. Then, we performed the extra sums squares test by generating two-way interaction plots to further test our hypotheses. Our plots compared gender versus work hours, and gender versus relationship. If the lines crossed, we concluded that there was an interaction between the two predictors. Otherwise, if the lines were parallel, then we could not conclude that there was an interaction present.

Utilizing the data from the ANOVA output comparing two different fits, we were able to select the better fit through analysis of p-values. Then we proceeded to use the better fit of the two to execute stepwise regression, which consisted of: forward addition, where the model starts with no predictors, but continues to add them one by one if they hold a significant effect on the overall regression; backward elimination, where the model starts with all possible predictors, but works backwards to drop any predictors that may not be necessarily important from the model; and finally both directions of each predictor variable, where the model starts with no predictors, but as it continues to add predictors one by one, it also goes backwards through the model to check for any unnecessary predictors that can be dropped each time, before proceeding to add more predictors. From this, we end up with a best-fit model that is consistent throughout all three stepwise regression methods. Following all of this, we tested for possible violations of assumptions of our final model through analysis of the QQ Plot, Histogram, and Residuals Plot. We focused on the residuals plot to diagnose linearity (if the data points are randomly scattered), constant variance (the data points are not in a cone-shape that fans outwards to one side), level of independence and possible outliers. From the QQ Plot and Histogram, we diagnosed normality. Upon finding any violations, we will remedy them by transforming them accordingly. Through all of this, we are able to arrive at what we believe to be the best fitting linear model.

Results

From the scatterplot matrix, we first noticed that between happiness and gender, males tend to be less happy on average, while females tend to be more so; between happiness and work hours, there was a spectrum where the happier volunteers resided in the low middle; between happiness and relationships, the volunteers were happier the more high quality their relationships were.

Next, we performed our hypothesis test: Null: $B1 = B2 = B3 = 0$. Alternative: $B1 \neq B2 \neq B3 \neq 0$. When the slopes of gender, work hours, and quality of relationship all equal zero, then happiness is at 3.54123 points. When work hours and relationship are held constant, for every additional unit in gender, there is a 1.55447 point increase in happiness. When gender and relationship are held constant, for every additional unit in work hours, there is a 0.07118 decrease in points for happiness. When gender and work hours are held constant, for every additional unit in quality of relationship, there is a 0.48538 point increase in happiness. From this, we concluded our initial model to be: $Y = 3.54123 + 1.55447X_1 - 0.07118X_2 + 0.48538X_3$. Then looking at the p-value, we noticed that $2.2 * 10^{-16} < 0.05$, so we can reject the null hypothesis. So we can reject the null hypothesis and conclude that there is a relationship between happiness, gender, work hours, and relationship. Finally, we find that the coefficient of determination $R^2 = 0.907$, which means that 90.7% of the variation in happiness can be explained by knowing all three predictors.

In order to perform stepwise regression, we first utilized the ANOVA function to compare two different fits: fit #1 that included all three predictors, versus fit #2 that included all 2-way interactions. Because Fit #2 had a p-value of $1.047 * 10^{-12} < 0.05$, we concluded that Fit #2 was a better fit to use in the stepwise regression tests. From here, we analyzed the partial p-values of each two-way interaction to determine which are significant. While gender:workhours and workhours:relationship are both greater than 0.05, gender:relationship has a p-value of $3.26 * 10^{-14}$, which is less than 0.05 and is therefore significant and should be kept in the final fit model. Proceeding to forward addition, we started with a model with no predictors, but ended up with the model: happy ~ relationship + gender + workhours + relationship*gender, which had the smallest AIC value of -183.06. Doing the same for backward elimination, we started with a model with all two-way interaction predictors, but ended up with the same model and AIC value as that of forward addition. Finally, for both directions, the same model and AIC value resulted. In conclusion, we can say that the best-fit

model includes all three original predictors, plus the two-way interaction between gender and relationship.

These results were further supported by our extra sums squared test and our two-way interaction plots. The plots showed that the regression lines between gender and work hours were parallel, signifying that there is no interaction. On the other hand, between gender and relationship, there is an interaction.

Now, through analysis of residuals for potential violations of assumptions, we were able to observe linearity due to the random scattering of the data points, and constant variance because there was no trumpeting/outwards fanning to one side on the residuals plot. Through the same plot, we also noticed independence and a lack of outliers. Then normality was verified through the QQ plot due to no points significantly straying from the line, and through the Histogram which exhibited a bell-shaped, symmetric curve. As a result, there were no violations that we needed to remedy.

In conclusion, through all of our plots, ANOVA, and sum of squares test, we were able to determine that all three variables, gender, relationship and work hours were all significant, along with the two-way interaction between gender and relationship. Therefore our final best-fitted regression model is $Y = 4.28774 + 0.17835X_1 - 0.07026X_2 + 0.35210X_3 + 0.24158(X_1 * X_3)$. If no parameters are included, the intercept is 4.28774 level of happiness. Considering the other predictors are held constant, the difference in mean happiness between males and females is 0.17835. For each additional work hour considering the rest remain constant, happiness decreases 0.07026. For each additional relationship score considering the rest remain constant, happiness increases 0.35210. And thus, for each additional interaction between gender and relationship, happiness level increases 0.24158. The p-value of the overall model is $2.2 * 10^{-16}$ which is less than 0.05, meaning that the model is significant. The coefficient of determination R^2 is 0.95, which implies that 95% of the variance can be explained by knowing gender, relationship, work hours, and the interaction between gender and relationship. Compared to

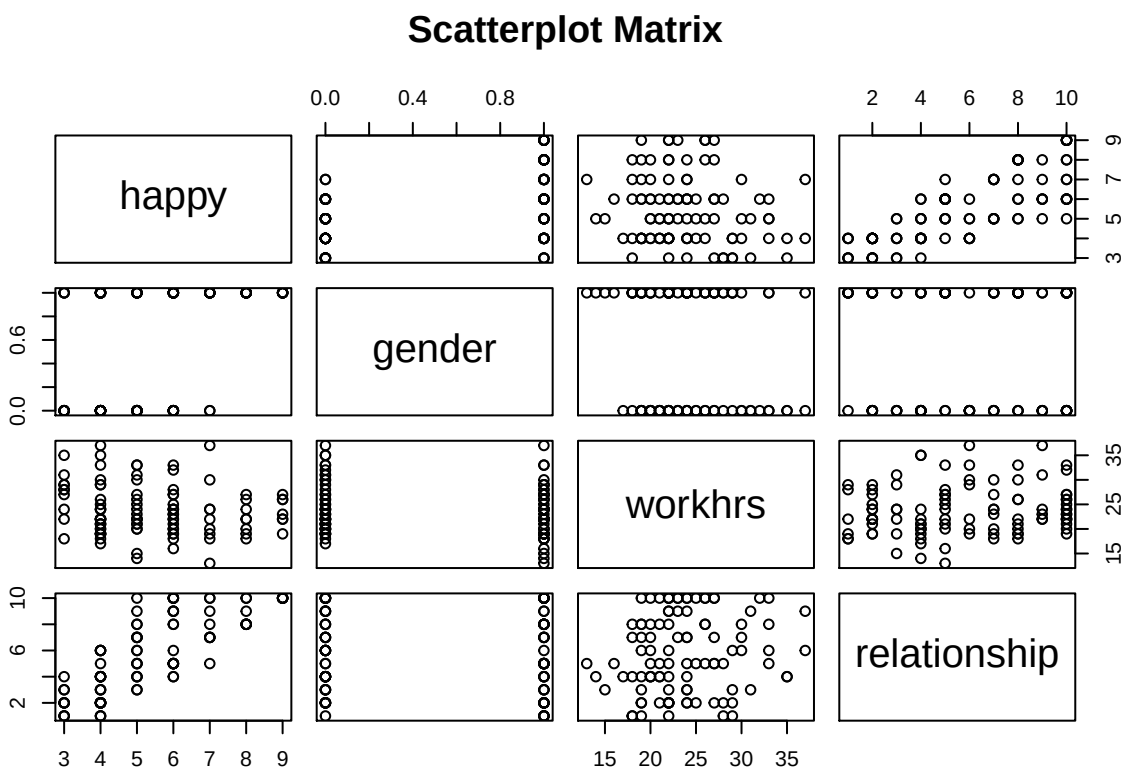
the initial fitted model value of $R^2 = 0.907$, the final model fit R^2 increased by 5%, indicating that the additional predictors make a statistically significant contribution to the regression model. The slopes of gender and relationship were both increasing but non-parallel, therefore there was an interaction (see two-way interaction plots in Appendix). The partial p-value of relationship is $2.2 * 10^{-16}$, work hours is $5.85 * 10^{-14}$, and relationship times gender is $1.84 * 10^{-14}$. Since all these values are less than 0.05, those variables are all significant in the model. The partial p-value of gender is 0.301, which is not significant, however, the overall conclusion is not affected because when rejecting the null, it is still possible for at least one predictor to be insignificant.

Discussion

Overall, we can conclude that happiness is heavily dependent on gender, quality of relationship, and number of work hours, supporting our initial prediction. The only thing missing from our prediction that we did not account for is the significance of the two-way interaction between gender and relationship. A limitation to this data set/experiment is the the lack of other possible predictors that could help us determine a more precise conclusion related to an individual's happiness level. Some questions that we would ask for future research would include the location where the individual resides, wealth status, age, and other similar variables that could affect the data set.

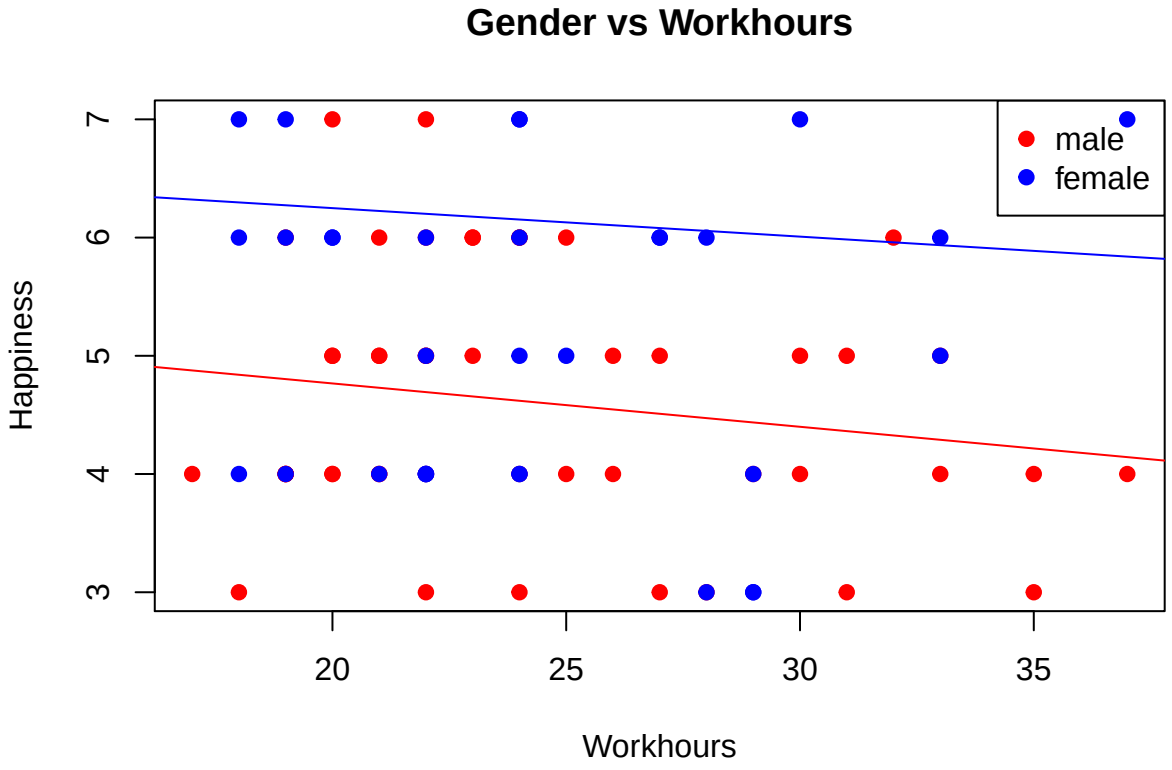
Appendix

```
projdata = read.table("/Users/laurenwong/Downloads/projdata.txt",header=T)
attach(projdata)
pairs(happy~gender+workhrs+relationship, main = "Scatterplot Matrix")
```

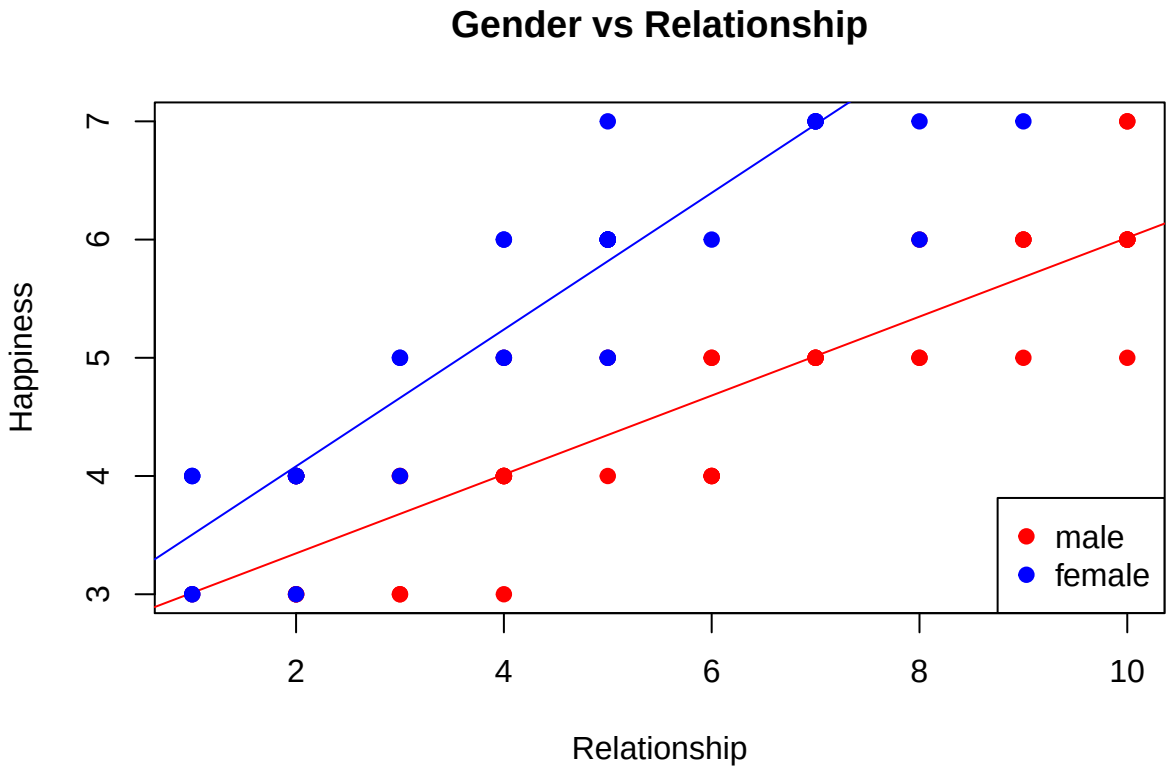



```
model <- lm(happy~gender+workhrs+relationship)
summary(model)
#the fit of the overall model/the overall p-value is 2.2e-16
```

```
plot(workhrs[gender==0],happy[gender==0],col="red",pch=19, xlab = "Workhours",
     ylab = "Happiness", main = "Gender vs Workhours")
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")
legend("topright",c("male","female"),col=c("red","blue"),pch=c(19,19))
```

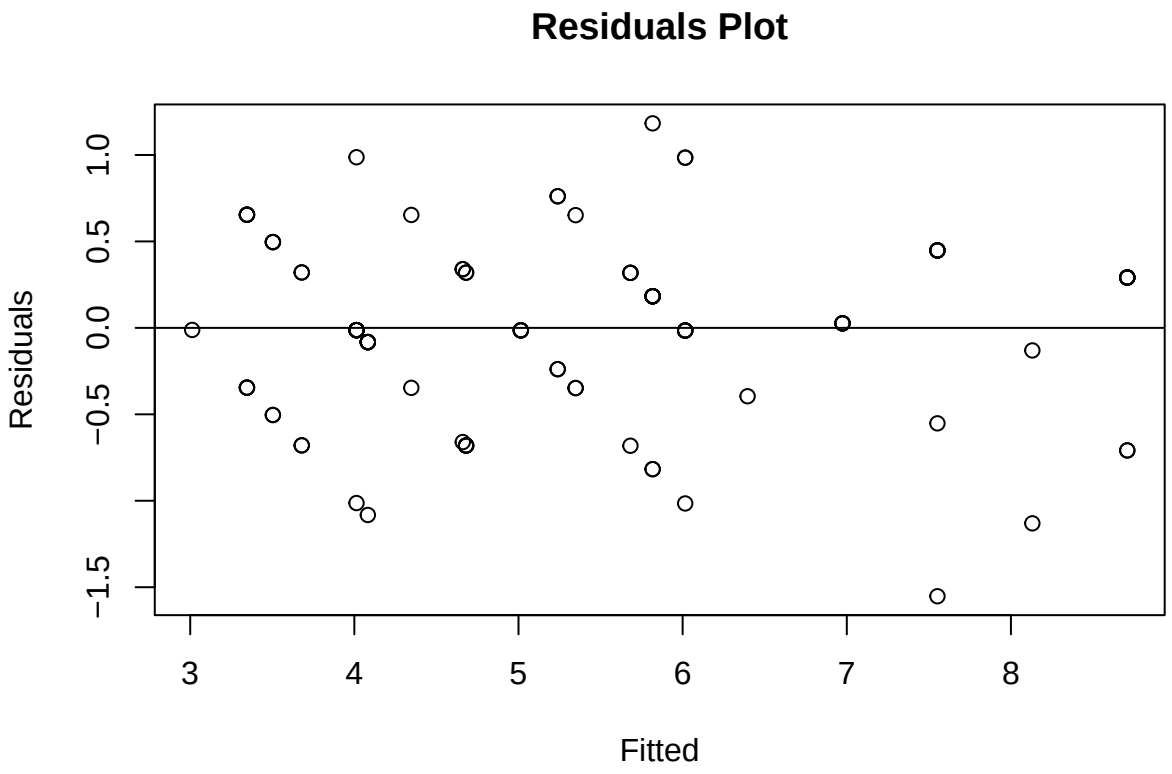



```
plot(relationship[gender==0],happy[gender==0],col="red", pch = 19, xlab = "Relationship",
     ylab = "Happiness", main = "Gender vs Relationship")
abline(lm(happy[gender==0]~relationship[gender==0]),col="red")
points(relationship[gender==1],happy[gender==1],col="blue",pch=19)
abline(lm(happy[gender==1]~relationship[gender==1]),col="blue")
legend("bottomright",c("male","female"),col=c("red","blue"),pch=c(19,19))
```

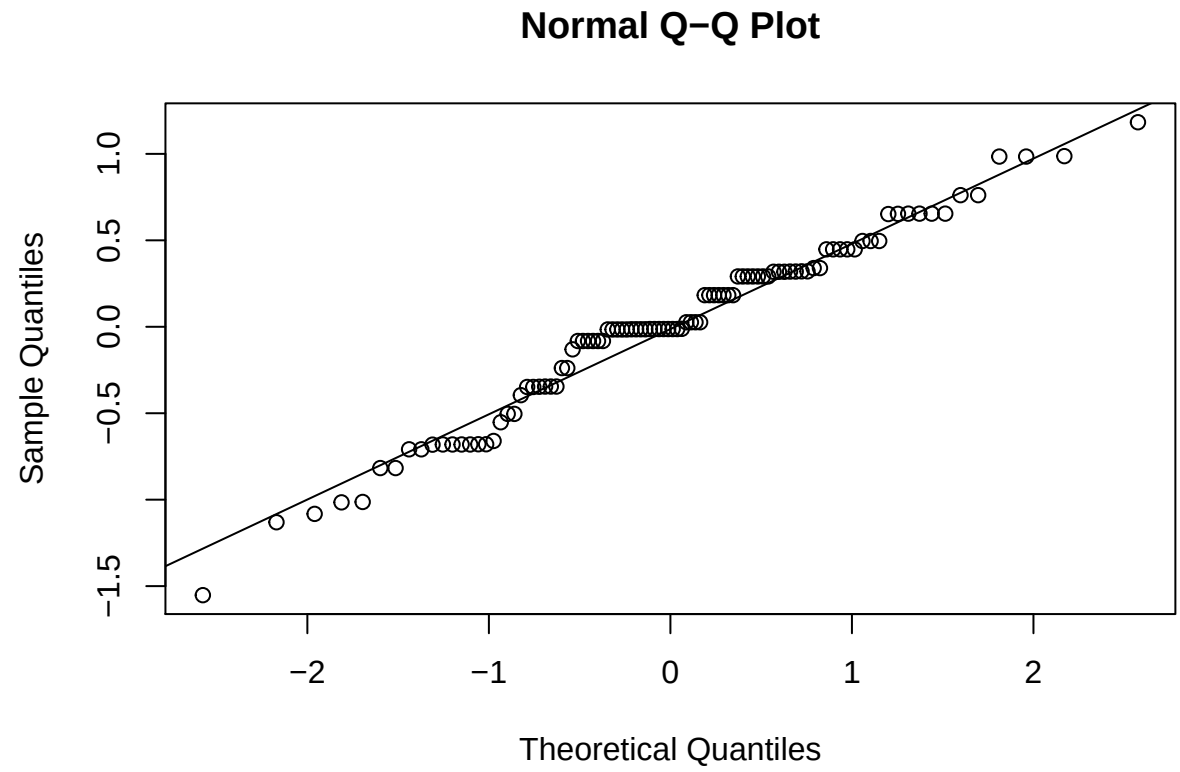



```
fit1 <- lm(happy~.,data=projdata)
summary(fit1)
fit2 <- lm(happy~.^2,data=projdata)
summary(fit2)
anova(fit1,fit2)
#fit2 is better so use full in stepwise regression tests
null=lm(happy~1,data=projdata)
#empty, no predictors - estimates the mean for every variable
full=lm(happy~.^2,data=projdata)
step(null,scope = list(lower=null,upper=full),direction='forward')
#looks at all possible predictors (all 2-way interactions) determines out of those
#which one makes the biggest improvement in R2 and adds it to the model
#step 1: relationship is the most important predictor,
#of the 5 remain, which one makes the biggest improvement so on through AIC
#adds in order of importance, contribution
#step 2: keeps going until it gets to
#happy ~ gender + relationship + workhours + gender * relationship
step(full, direction='backward')
step(null, scope=list(upper=full),direction='both')
```

```
rel = lm(happy~relationship*gender)
#Final model = happy ~ g + r + w + g*r
#constant variance (because graph doesn't fan out)
#no outliers, no transformations
plot(fitted(rel), residuals(rel),xlab = "Fitted", ylab = "Residuals", main = "Residuals Plot")
abline(h=0)
```




```
#normal
qqnorm(residuals(rel))
qqline(residuals(rel))
```



```r
hist(residuals(rel))
```