Name: Zheng Jing, Kendall Brown Section: Tuesday 6:00 p.m.

**Introduction**

In the project, the level of 'happy' is described by number from 0 to 10. For gender, 0 means male and 1 means female. The level of "relationship" is described by number from 0 to 10. Research question: Is there any relationship between happy and gender, work hours and relationship? If there is, what is the relationship, linear or non linear? Is there any interaction? We predict a positive linear relationship between happiness and relationship level because we believe the more people are in love, the happier they will be. We predict a negative linear relationship between happiness and work hours, because more work hours leave less time for joyful experiences. We predict that female is happier than male in general, because female has less work pressure than male. We guess there might be interactions.

**Method**

We read project data into R. First, we generate first order model and its scatterplot. Second, we generate summary output of the model to test its significance. Third, we generate a full second order model and we use F-test (anova in R) to test whether the three interactions are overall significant. Fourth, we use stepwise procedures to find our fittest model, but before that we generate a null model with no predictor. We also use partial, sequential extra sum of squares and adjusted R-square to select our best model. Fifth, after reaching our final model, we generate fitted residual plots, QQ normal plot and histogram to diagnose possible violations (linearity, independence, identically distributed, normality and constant variance). We also look at scatterplot matrix to check linearity. Sixth, we generate summary output of our final model. Lastly, we draw an interaction plot of gender and relationship.

**Result**



From the scatterplot matrix, there exists a moderate positive linear relationship between happiness and relationship. A female is happier than a male on average. There exists very weak negative linear relationship between happiness and work hours.

model1: happy = β0 + β1\*(gender) + β2\*(workhrs) + β3\*(relationship)

H0: β1=β2=β3=0 H1: not all β1, β2, β3 are zero.

From summary output of first order model, we observe a model F-statistics of 312.2 with degrees of freedom 3 and 96, with a correlating overall p-value of approximately zero (<2\*10-16), so we reject null hypothesis and conclude our first order model is overall significant. R-square is 0.907, which means 90.7 percent of the variance of happy could be explained by knowing gender, work hours and relationship level. We also observe that partial p-values of gender, work hours and relationship are all smaller than 0.05, which means all three predictors are significant given the rest are already in the model.

model1: happy = β0 + β1\*(gender) + β2\*(workhrs) + β3\*(relationship)

model2: happy = β0 + β1\*(gender) + β2\*(workhrs) + β3\*(relationship) + β4\*(relationship:gender) + β5\*(relationship:workhrs) + β6\*(workhrs:gender)

H0: β4=β5=β6=0 H1: not all β4, β5, β6 are zero.

By comparing model1 and model2 in anova, we observe a p-value of 1.047\*10-12 < 0.05, so we reject null hypothesis and conclude that three interactions are overall significant in second order model. Then we did stepwise regression in 'forward', 'backward' and 'both' direction. All of them give us the same model as following:

happy = β0 + β1\*(gender) + β2\*(workhrs) + β3\*(relationship) + β4\*(relationship:gender)

Moreover, by using partial and sequential extra sum of squares method, we get the same model as above. By comparing adjusted R-squared of different models, we determine the model above is our final best fittest model with highest R-square.

 

Then we check violations of our final model. From fitted residual plot, variances look constant along x-axis. Moreover, there is no pattern on fitted residual plot, so the model is linear. Also, from the scatterplot matrix above, there is a weak linear negative relationship between happy and work hours, there is a moderate linear positive relationship between happy and relationship, and on average women are happier than men.

 

In qq normal plot, points basically follow the line. Besides, the shape of histogram is close to the normal curve. Thus it is normal. Additionally, we observe no outlier from the plots above, so the data are identical distributed. We can't check independence since the order of project data is not given.

happy = β0 + β1\*(gender) + β2\*(workhrs) + β3\*(relationship) + β4\*(relationship:gender)

H0: β1=β2=β3=β4=0 H1: not all β1, β2, β3, β4 are zero.

The model has a F-statistic of 451.7 on 4 and 95 degrees of freedom with a correlating p-value smaller than 2.2\*10-16, which is smaller than 0.05, so we reject null hypothesis and conclude that our final model is overall significant. Moreover, partial p-values of workhrs, relationship and the intersection between gender and relationship are all smaller than 0.05, which means they are all significant given the rest of predictors already in the model. However, p-value of gender is 0.301 > 0.05, which means gender is not significant given all other predictors already in the model, but we cannot drop it since gender appears in the intersection which is significant. R-square of our final model is 0.95, which means 95 percent of the variance of 'happy' could be explained by knowing gender, work hours, relationship and the intersection between gender and relationship. β0 is 4.287745, meaning a male with 0 work hours and no relationship will have 4.287745 level of happy. β1 is 0.178353, meaning a female will be 0.178353 happier than male with the same work hours and relationship level. β2 is 0.070259, meaning that with each additional hour of work, both male and female will have a decrease of 0.070259 in happy, given other variables constant. β3 is 0.352098, meaning that with each additional unit increase in relationship, both male and female will have an increase of 0.352098 in happy, given other variables constant. β4 is 0.24158, meaning that with each additional unit increase in relationship, a female will increase more than male by 0.24158 in happy.



We generate an interaction plot between gender and relationship. We observe that there is a moderate positive linear relationship between happy and relationship for both male and female, But the slopes of two lines are different, which means with each additional unit increase in relationship, the amount of increase in female happiness is larger than that of male. Two lines are not parallel, so there does exist an interaction between gender and relationship.

**Discussion**

we find that a female is happier than men in general. More work hours make people unhappy. Deeper relationship makes people happier. Deeper relationship has a larger positive impact on female than male in happiness. These results confirm my predictions in the introduction. Model limitations: other predictors, such as income, appearance, might also affect happy. Further research question: what are the other predictors that also have a great impact on happiness?

**Appendix**







