**Stats 101A Final Project**

By: Justin Kim, Xiaoying Huang, Yae Hyeon Kyeong, Arvin Monajemi, Avash Monajemi, Jaehyeong Lee

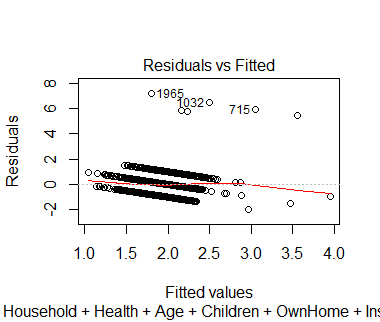
**Introduction**

In this project, we are interested in finding what kinds of factors are affecting the happiness level of Americans. The purpose of this project is to find the significant predictors that explain one’s happiness level, and model their relationships. In this project, we have 12 predictor variables, which are standard core of demographic, behavioral, attitudinal questions with special interest topics such as civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events, and 1 response variable, happy. We expect that not all of the predictor variables would be significant predictors. Hence, we plan on transforming the full multilinear regression model, differentiating the transformed predictor variables to see which are significant and which are not, and then running the model selection based on the transformed model. We expect to keep on doing these procedures until we can get a multiple linear regression model that best predicts the response variable.

**Methods and Procedures**

We started by studying and analyzing the codebook to understand each predictor variable and the codes that represent the responses. We changed some of the data points like ‘Inapplicable’, ‘Don’t know’, and ‘Not answered’ to NAs or other appropriate responses. In addition, we changed the 0’s in the response to 0.01 to prevent the data from upsetting the log function when we use power transformation. Next, we used the lm function to find a multiple linear regression model. To check whether this full model was good, we examined the summary of the model to check the significance of each variable, the adjusted R^2, the diagnostic plots, and the standardized residual plots for each predictor variable. We found in the summary that there were some insignificant predictor variables. In addition, in the diagnostic plots and normal Q-Q plot, there were many points that deviated far from the main line which indicated that the normality assumption was violated and that the model needed to be fixed. Also, the standardized residual plot showed clusters of data points instead of random scattered points which revealed that there was not constant variance and an average error of 0. Another red flag was that the residual vs leverage plot showed many leverage points and outliers which meant that some of the data points needed to be examined more closely. Also, most of the individual standardized residual plots showed clustered points which implied that this linear regression model was not appropriate for the data.

With the results of the analysis, we concluded that we had to transform the current model to find a better model that would explain the data and happy variable well. Using the powertransform function, we found values of lambda, which told us the appropriate transformation for each variable. With the appropriate transformation, we created a new multiple linear regression model and repeated the same analyses procedures that we did for the original model. The summary of the new model showed fewer significant variables, but the diagnostic plots showed favorable results. For instance, the Residual vs. Fitted plot had more linear relationship compared to the previous model and its average of errors was 0. Another favorable result was that the Scale-location Plot showed constant variance and random spread and the Residual Plot vs. Leverage Plot showed a significantly decreased number of influential points that might cause problems. Moreover, the individual standardized residual plots showed constant variance for almost all the predictor variables which implied that they were good.

Since this newer model was better based on our analyses, we checked the multicollinearity with vif function and it showed that all of them had vif less than 5 which indicated low correlation among the predictor variables. However, some of the added variable plots did not show significant linear relationship which suggested that there could be some possible multicollinearity between the variables. Therefore, variable selection was carried out through possible subset method, forward selection method and backward elimination method to find the best possible model. Based on these methods, we found the best model that excluded some variables that had limited prediction power. Then we repeated the same procedure of analysis on this reduced model to double check how well this model fitted the data. Based on the summary, which showed that the predictor variables were all significant, the diagnostic plots, which were all very good, and the standardized residual plots, which showed constant variance in all the predictor variables and average error of 0, we concluded that the model was the best fit. 

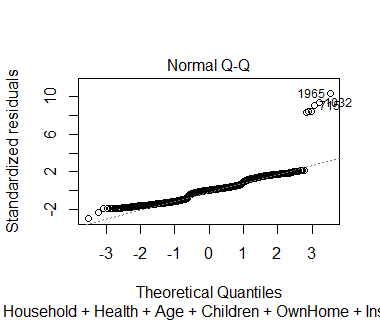
**Results & Interpretation**

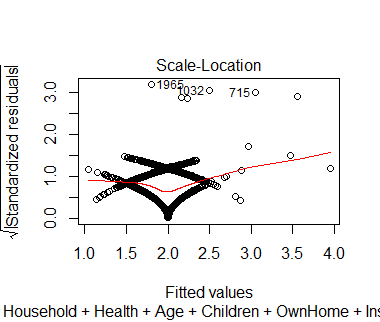
Untransformed, Full Model

We examined the untransformed multiple linear regression model as follows:

Predicted(Happy) = Intercept + (Household)X1 + (Health)X2 + (Age)X3 + (Children)X4 + (OwnHome)X5 + (Instagram)X6 + (Martial)X7 + (Sex)X8 + (Education)X9 + (JobSat)X10 + (Income)X11 + (WorkHrs)X12

Firstly we paid great attention to the basic regression summary results of p-values, R^2 and slope coefficients. From our summary output, we did see a number of significant predictors with low p-values and a low R^2 (however this is trivial as regression models relating to the behavioral sciences often turn out in this manner). These summaries however are not sufficient enough as our choice of model must be supported with model diagnostics checking various model assumptions and further analysis of our variables.

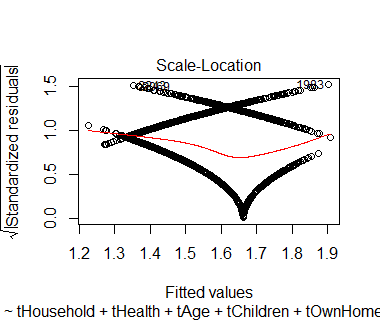
We decided to take a look at various plots that can determine if the model is appropriate for the Happiness data or not. These variables that describe American Society are related to the areas of one’s background of demographics, family characteristics, and other social variables. There are 12 total of these variables and we seek to determine if a multiple regression model containing these and the predicted Happiness score is appropriate. In doing so, we used a series of diagnostic plots of residuals, leverages, normal probability plot, and others to examine if this full model is an appropriate fit for our data.  Some of these were the 4 plots in a diagnostic plot output, as well as identification of bad leverage points.

* Our Residuals vs. Fitted Values plot did not have random scatter about the horizontal line, so the assumptions that the variance of the error terms being equal is not satisfied. There is definitely a downward pattern that we see. This indicates that our model is not appropriate for the data that we have.
* In addition, the Q-Q Plot does not seem to be that linear in form, and there are outliers present which are an issue for our model. These can distort the accuracy of predicting Happiness (Y).
* Our Fitted Values vs. sqrt(Standardized Residuals) plot also does not look good. We see that our variance of the error terms changes quite a lot as scanning from left to right. Nonconstant variance is one of the assumptions we make when dealing with both simple linear regression and multiple linear regression and this has been violated.
* We then examined leverage points. Leverage points will exercise considerable influence on the Happiness model. We decided to use the cutoff point of 2 X (P + 1) / N for considering a point as a leverage value (where N is the number of observations and P is the number of predictors).
* The observations that were found to be leverages were (just a few of them here):
* 5  12 26   46 50 74   98 102 127 136  137 169 186 250 260  273 274
* In total, there were 105 leverages for this model.
* Specifically, another way that we analyzed the model was by determining if our leverage points were bad leverage points. We examined the standardized residuals using rstandard() and checked to see which leverage points are OUTSIDE the interval (-2,2). This is because a leverage point that is an outlier is a bad leverage point.
* Our outliers were (just a few of them):

9   12 55  389 398 448  684 715 923 1001 1032 1253 1286 1518 1629 1813 1965

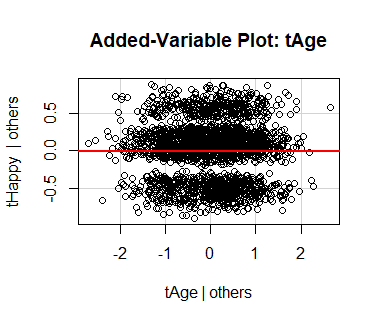
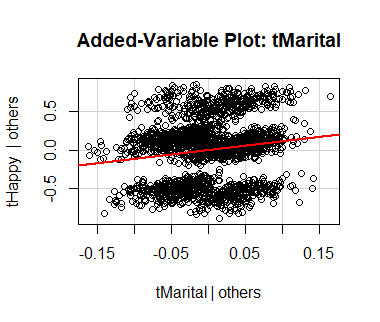
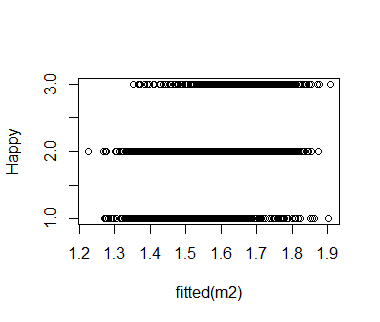
* We found that 6 of the leverage points were outside the interval, and so were bad leverages. These do not follow the trend set by other points, so it distorts the regression line for Happiness. Those bad leverage points were:

2  7 8  9 11 13 (therefore, 6 total)

* We would like to determine if multicollinearity is present in our data, there could be pitfalls if two or more predictors in the reg model are correlated. Specifically, we do not want to see a linear dependence between variables (eg: X4 = 3X2 + 2X5). Variance Inflation factor outputs (**TABLE 2)** indicate that all are less than 5. Not a lot of the variance of each estimated coefficient of regression is “inflated”  by the other predictor variables. We can say that multicollinearity is not an issue here.
* In light of our analysis for the diagnostic plots (eg: to aim to reduce bad leverage points, correct the variances of the error terms), overall producing fallacies, we attempt to take transformations of our data for better results.

Transformed, Full Model

Our transformation using PowerTransform in R gave us the powers that we need to apply for each X value (determined by maximum likelihood estimates). The first column in the output gave us the best powers to apply for the transformation of the multiple linear regression model. We then applied these to a new model, and examined the diagnostic plots.

* We notice that the Residuals Vs. Fitted plot does not show random scatter. There is a definite linear relationship (negative) being seen which shows that the model is not appropriate.
* We notice that the Normal Q-Q plot is relatively linear which implies a normal distribution of the errors. However, it can be significantly improved. There are large departures from normality in the middle of the plot.
* The Fitted Values Vs. Standardized Residual plot definitely shows nonconstant variance here. As we scan from left to right, we see that the variability changes greatly. The issue of non constant variance still exists, which means that something about this transformed model is not good.
* However, this model had less bad leverage points. The untransformed, full model had 6 bad leverage points whereas this new transformed model had only 2 bad leverage points. Therefore, we are on the right track to producing a much better model.
* We decided to investigate the added variable plots to see if each individual.
* l predictor can explain Happiness well (taking into account all the other predictors).
* We notice that some variables, such as Age, did not explain Happiness very well when the others are considered. The linear seeme horizontal and there was almost strong relationship.
* However, Marital status seemed to explain Happiness quite well and there is a much stronger relationship when the other predictors are considered. We then realize that we should remove certain variables that might be problematic and do not predict Happiness well. Supporting evidence is that the P-value in the F-test for Age is high whereas the P-value for Marital is less than 0.05, implying significance as a predictor (see below!)
* In addition, we have less bad leverages! Now, only observations 8 and 27 are bad leverages. Therefore, this model is better than the original model.
* We do not see a linear relationship between Y and Y(hat)

In the linear regression, you want the predicted values to be close to the actual values. So to have a good fit, that plot should resemble a straight line at 45 degrees.

* We conclude that the model does not seem to provide an adequate fit and should be revised

Transformed, Subsetted Model

* We resorted to AIC, AICc, and BIC for determining the appropriateness of the 12 subset models for the Happiness data. These were determined using forward selection, backward selection, and all possible subsets (diagnostic plots are in appendix)

Size       Radj2     AIC AICc        BIC

1     1 0.04614797 -4333.069 -4333.059 -4321.536

2     2 0.05548366 -4357.213 -4357.196 -4339.914

3     3 0.06705359 -4375.114 -4375.089 -4352.049

4     4 0.07949525 -4405.635 -4405.599 -4376.803

5     5 0.08460030 -4409.011 -4408.963 -4374.412

6     6 0.08817370 -4414.842 -4414.781 -4374.477

7     7 0.08954797 -4417.608 -4417.532 -4371.477

8     8 0.09007296 -4413.043 -4412.949 -4361.145

9     9 0.09037201 -4402.565 -4402.453 -4344.901

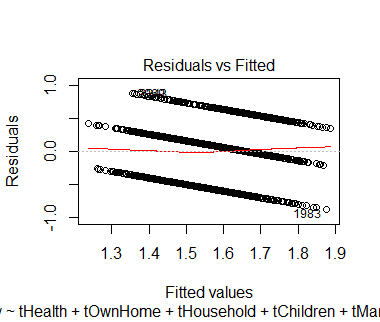
10   10 0.09044800 -4401.501 -4401.368 -4338.070

11   11 0.09055417 -4389.162 -4389.007 -4319.965

12   12 0.09018496 -4372.692 -4372.513 -4297.728

we notice that for subset n = 4, BIC is the lowest

we notice that for subset n = 7, AIC is the lowest, and AICC is also the lowest

we notice that for subset n = 11, R^2 adjusted is the highest

We chose n = 7 as the best subset for the model, since both AIC and AICc were the lowest.

Looking at the **Appendix, (Table 1),**the summary of the best model can further support our conclusion, especially due to all the predictors being highly significant.

* i

Health, OwnHome, Household, Children, Marital, JobSat, and Income were the 7 predictors that we used for the subset

* We noticed that the diagnostic plots all seem relatively the same as the previous model (transformed, full model). The Residuals Vs. Fitted plot showed less outliers than the previous plots. Also, we did not have any bad leverages at all (**Table 3 in Appendix)** which signifies that this was another step in the right direction. This means that our model is much more appropriate in predicting happiness than before.
* In addition, all our predictors have an extremely low p-value, implying great significance.
* In addition, the added variable plots seem to be much better than the previous model. They all show a relatively positive (or negative) linear relationship with Happy, controlling for the other predictors in the same model.

**Discussion**

In our project, we examined the variables that affect happiness level. By transforming and doing variable selection, we found that the best model to predict happiness included the variables Health, OwnHome, Household, Children, Marital, JobSat and Income.

According to some other research publications, our model does make sense in the real world. In the article “Factors Affecting Happiness”, the researchers studied the factors that influenced the happiness of Iranian Youth. Results of the cross-sectional study showed that “age groups”, “type of occupation”, “physical activity” and “place of residence” were factors associated with happiness in young persons. (Mehrdadi et al.) This supports our results as JobSat and Income (variables related to type of occupation), Health (variable related to physical activity) and OwnHome (variable related to place of residence) are included in our final model. The scientific study also demonstrated that “gender”, “marital status” and “educational level” were not significant in predicting happiness. Our analysis also provided similar results as we eliminated the variables “gender” and “education” from our final model.

Having said this, there are in fact some limitations that should be addressed. The data is vulnerable to unpredictable social and economic events that could occur. For instance, during an economic recession, people lose their homes or become unemployed. In these circumstances, predictors such as OwnHome, JobSat, and Income could all end up becoming much more significant in explaining happiness. In an economic crisis, a stronger relationship might be seen between factors related to one’s career, job, socioeconomic status and happiness.  In a real-life setting, we typically get a large percentage of nonresponse in surveys or observational studies. People who refuse to respond to a survey could be systematically different from the rest of the population, and so the data could produce dramatically different results if they were to be included. This could lead our inferences to be based on a biased model. Methods on reducing nonresponse in order to minimize bias should be examined to create a more generalizable model. One way of handling this issue could be assigning weights based on the proportion of non response for each observed individual. Therefore, observations that have nonresponse values would be taken into less consideration in the calculations.

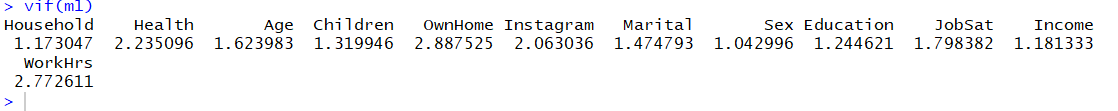
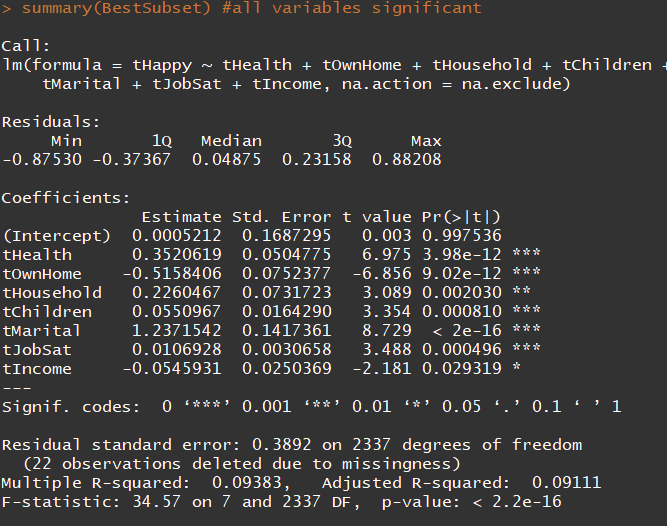
In addition, we must be careful when making claims that certain variables cause a low happiness index. Because the data originated from an observational study, we can only state there might be an association between our explanatory and response variables.

**APPENDIX**

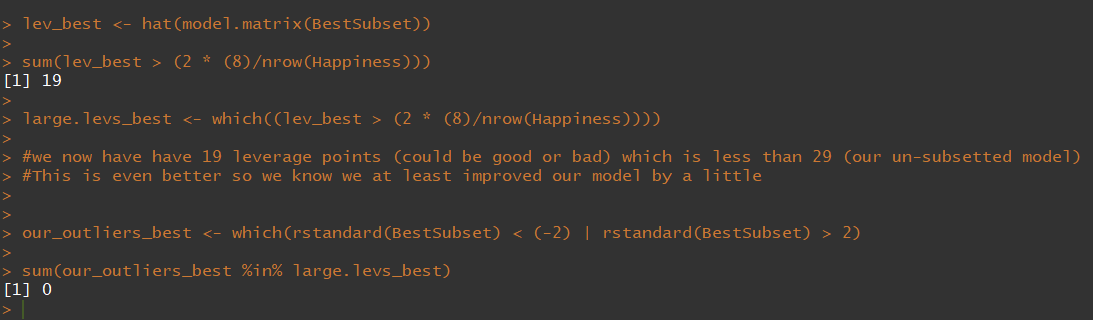
Work Cited:

**Mehrdadi, Amir et al. “Factors Affecting Happiness: A Cross-Sectional Study in the Iranian Youth.” *Journal of Clinical and Diagnostic Research : JCDR* 10.5 (2016): VC01–VC03. *PMC*. Web. 17 Mar. 2018.**

**TABLE 1**

**Table 2**

**Table 3**

****