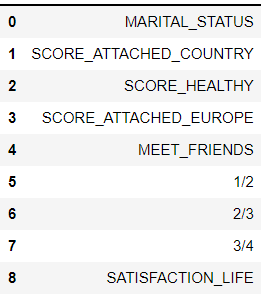
Ordinal Regression

Our main purpose of conducting an ordinal regression is that our dependent variable “SCORE\_HAPPY” is an ordered range of values between 0-10 that suggests how happy people are as a factor of many independent variables ranging from economic, societal, personal and governmental factors.

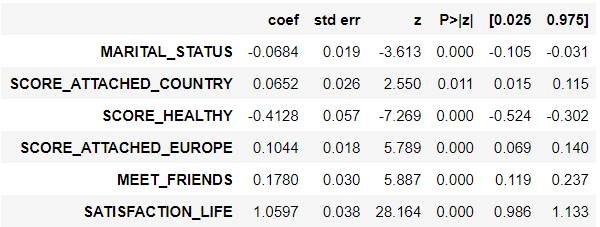
For better clarity and ease of understanding, we have divided our Happiness rating into four main categories – 1,2,3 and 4 as per the model that we run. Here 1 indicates a score of less than 6. 2 indicates a happiness score of between 6 and 7. 3 indicates a score of between 7 and 8 and 4 indicates a score of 8 or above on the happiness scale. We now have the required thresholds on the happiness scale and this **HAPPY** column that we have added now becomes our latent variable. The reason for choosing 3 thresholds is to split the available data evenly across the latency scale.

In the first iteration that we run, we first remove all the null values in our data and run an ordinal model with HAPPY as our dependent variable and the rest as our independent variables. We have 2745 observations at the beginning, after removing all the null values we still have 1273 observations left.

In order to check the performance of our ordinal models and make accurate predictions, we then split the cleaned dataset into two parts: train and test. The training dataset is used for fitting an ordinal model, and the testing dataset is used to check the accuracy of our predictions. By using all the variables, we have, we observe that not all of them are statistically significant. We only get a list of 6 significant variables.

 This picture shows all the significant variables we got after the 1st iteration. The threshold chosen for the significance is 10% as there is the chance that some variables being slightly past the 5% mark might stabilize when some of the noise added by the more openly irrelevant variables is removed. Besides, the variables ’1/2’, ’2/3’ and ’3/4’ are only the thresholds we set, and hence we remove them from this list. Finally, we have 6 significant variables and they will be used in the second iteration.

We now run a second iteration of the ordered response model with these significant variables on the original dataset with no null values. In this model, all the variables are statistically significant.



We now try to check the overall performance of it and try to make predictions based on this model. The metrics we will use include convergence, information criteria, odds ratio, association stat and PCC.

**What are the interpretations?**

1. **The Regression Model –** The final ordinal regression model that we have can be given as follows:

****

In this equation, we notice that ‘SATISFACTION\_LIFE’ and ‘SCORE\_HEALTHY’ are the most two important variables regarding the absolute values of their coefficients. Considering this equation describes the score or index of individual **i**, we try to explain them as follows:

* 1. The coefficients of ‘SCORE\_ATTACHED\_COUNTRY’, ‘SCORE\_ATTACHED\_EUROPE’, ‘MEET\_FRIENDS’ and of ‘SATISFACTION\_LIFE’ are positive, which means when these variables increase, the individual would have more chance to feel happier.
  2. The coefficients of ‘MARITAL\_STATUS’ and ‘SCORE\_HEALTHY’ are negative, which indicates when these variables decrease, the individual would have more chance to feel less happy.
  3. Individuals’ satisfaction level to their lives will most significantly increase the possibility of feeling happy.

1. **The Significant Variables –** The factors that affect the latent variable in our analysis include the following:
   1. **Marital Status** – It has a negative effect on the happiness of an individual indicating that a married individual might not have a higher index of happiness.
   2. **Health Score** – The health score impacts the happiness negatively according to our analysis but the reasoning is that the scale of how health is measured by the survey is that people facing health issues have a higher score (closer to 10) than the rest.
   3. **Life Satisfaction overall** – The life satisfaction of an individual positively impacts the health, indicating that people satisfied with the general quality of life are bound to be happier as compared to individuals not enjoying the same standards.
   4. **Health Satisfaction overall** – People enjoying a good standard of health tend to be happier.
   5. **Meeting with friends** – Meeting friends positively impacts the health, indicating that people who spend more time with friends could have a happier life.
   6. **Score of attachment with Europe** – Living in Italy and other parts of Europe and having an emotional attachment to Europe in general, could also result in people generally being happier than the rest.
   7. **Score of attachment with Country** – This variable speaks of an individual’s love for their country (Italy in this study) and our model predicts a positive relationship between the two, indicating that individuals who have a stronger connection to Italy are bound to have a higher happiness threshold in their lives based in Italy.
2. **The Latent Variable -** The latent variable represents the appetence score of the individuals to be happy. The larger the score, the larger the chance for an individual to be happy.
3. **Information Criteria –** The AIC and BIC are indicators of model performance. The lower the scores, the better a model is in terms of its explanatory powers. In our analysis, we notice that a model that has variables has an AIC value of 4687.101 and a BIC value of 4738.922, which are smaller than those of a model that only has intercepts. This result indicates that our model is much more informative than a model with only intercepts.
4. **Odd’s Ratio -** The odds ratio gives the ratio of the modified odd to the original odd and is essentially used to describe the change in the probability of an event given a unit change in the variables that the event is dependent on. Here, we notice that the odds ratio of each independent variable is statistically different from 1, indicating that the probability of experiencing happiness is impacted significantly by changes in independent variables.
   1. **For Marital\_Status** - Estimated odds ratio is 1.071. If this value increases by 1 unit, the modified odd would be 1.071 times the original odd for all the individuals.
   2. **For Score\_Attached\_Country** - Estimated odds ratio is 0.937. If this value increases by 1 unit, the odd decreases by 6.3% for all the individuals.
   3. **For Score\_Healthy** - Estimated odds ratio is 1.511. If healthy score increases by 1 unit, the odd increases by 51.1% for all the individuals.
   4. **For Score\_Attached\_Europe** - Estimated odds ratio value is 0.901. If this variable’s value increases by 1 unit, the odd would be multiplied by 0.901 for all the individuals.
   5. **For Meet\_Friends** - Estimated odds ratio is 0.837. If individuals meet 1 more friend, the odd would decrease by 16.3% for all the individuals.
   6. **For Satisfaction\_Life** - Estimated odds ratio is 0.347. If this variable’s value increases by 1 unit, the odd would decrease by more than half (65.3%) for all the individuals.

We still observe that ‘SCORE\_HEALTHY’ AND ‘SATISFACTION\_LIFE’ are the most predictive variables.

1. **Wald and LRT –** The Wald test shows us if the unrestricted model can be better able to highlight the probability of experiencing happiness as opposed to a restricted model. The conclusion that we seek is robust to both the Wald and the LRT where we notice that the p value is equal to 0 and we can heavily reject the hypothesis that the coefficients of our independent variables will be equal to 0.
2. **Measures of Association –** We conduct the association stat tests to understand the performance of the model and explain its predictive powers -- if we got a perfect model, all the 4 values (Somers D, Gamma, Tau-a and c) should be equal to 1. For our testing datasets in this project, we got the Somers D and the Gamma values of 0.676. These two measures are the same, which implies that we have 0 tied pairs in our classification. The Tau-a value is 0.501 and the c value is 0.838, both of them indicate a good model and a high predictive performance because they are significantly different from 0 and 0.5 respectively.
3. **Percentage Correct Classifications –** We finallyhave a PCC value of 57%. The PCC value is calculated to understand how many individuals were correctly predicted as being in the right thresholds of happiness. This performance shows that the model has been able to correctly classify 57% of the happiness threshold of the individuals, indicating that the model is much better than a random prediction model (the PCC value for a random prediction model in this case would be 25% since we have 4 groups in this case).