Project Report:

Happiness Score Analysis Study

By

Linwan Xu 301243591

Yutao Shen 301251469

December 2, 2019

Simon Fraser University

STAT 350 Fall 2019

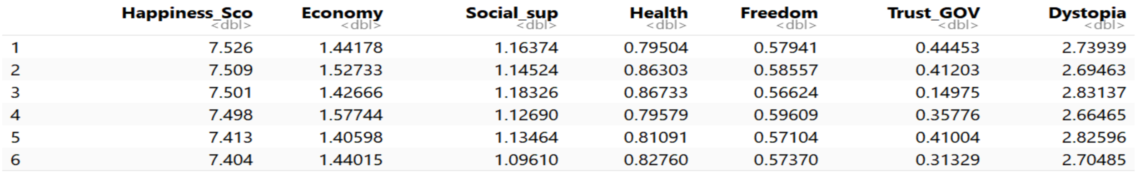
Instructor: Harsha Perera

Index:

Introduction

Our project topic is happiness score analysis. The world happiness report ranks 157 countries by the survey of how happy their citizens perceive themselves to be ("World Happiness Report 2016", 2016). Our main idea analysis the happiness level that was affected by influences like economy, social support, health, freedom, trust, and dystopia. Most people use happiness score as the basic evaluation to analyze global human developing, also many governments increase the citizens' satisfaction by different results of influences. So we want to find out what relationship between those influences and happiness.

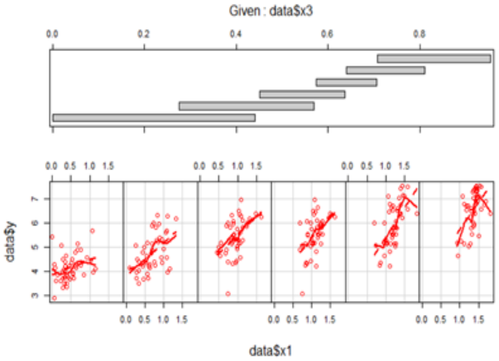
We use the data of World happiness score report from kaggle, Base on our dataset we have 157 countries as observation, the dataset like following figure. We choose Happiness Score as the dependent, the economy, and social support, health, freedom, trust, dystopia as multiple independent variables. We use x_1 to x_6 to replace those independent variables.

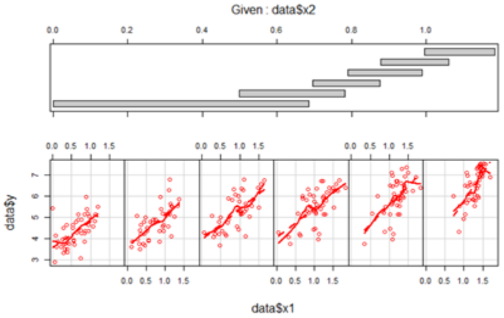


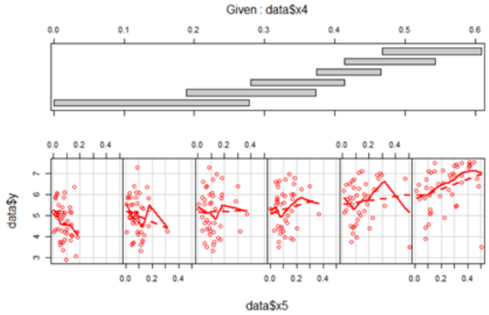
	Happiness_Sco <dbl>	Economy <dbl>	Social_sup <dbl>	Health <dbl>	Freedom <dbl>	Trust_GOV <dbl>	Dystopia <dbl>
1	7.526	1.44178	1.16374	0.79504	0.57941	0.44453	2.73939
2	7.509	1.52733	1.14524	0.86303	0.58557	0.41203	2.69463
3	7.501	1.42666	1.18326	0.86733	0.56624	0.14975	2.83137
4	7.498	1.57744	1.12690	0.79579	0.59609	0.35776	2.66465
5	7.413	1.40598	1.13464	0.81091	0.57104	0.41004	2.82596
6	7.404	1.44015	1.09610	0.82760	0.57370	0.31329	2.70485

Build Model

Regression analysis is a modeling way to investigate the relationship between response variable and explanatory variables, but several interaction effect exit when the dependent variable changing followed by value of more independent variables. The conditioning scatter plots is a plot of two variables conditional on the value of another variable. In R, we use coplot() to construct the figure. This R command separate conditional variable range to 6 parts. For each part, we can see other 2 variables linear line with corresponding range. If each part line is not parallel, there represents an interaction. The following three figures are not parallel. There exist interaction term. So we add x_1x_2, x_1x_3, $x_4x_5$as the interaction effect, x_1 to x_6 as main effect. The model is $y \sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_1x_2 + x_1x_3 + x_4x_5$.



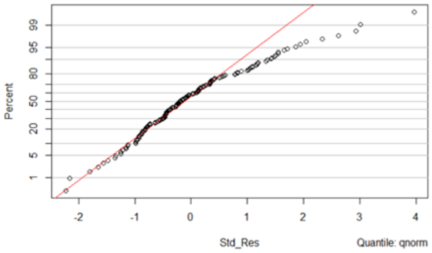


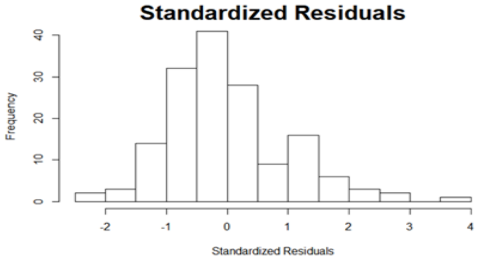


Model Adequacy Checking

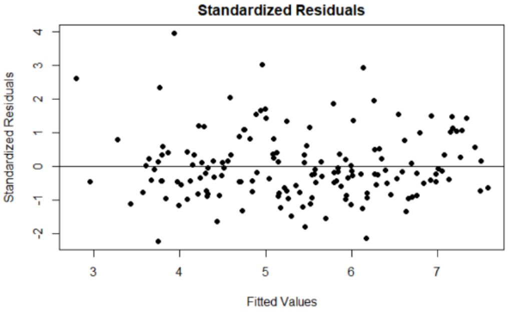
There are some model assumptions must be satisfied in the regression model; first, the relationship between the response variable and the regressors is linear or at least approximately linear; then the error term should have zero mean, constant variance and are uncorrelated; last, residuals should follow normal distribution. If one of those assumptions is not met, we should do a transformation for the data. Scatter plots of $x_i$versus y in the Appendix A, each regressor has approximately linear trend with the response variable.

The normal probability plot of residual can checks the normality assumption. With the left side figure, the plot looks main points around red line, but has several negative residual as outlier. The relationship is approximately linear with the exception of some data points. To do more tests, the following histogram of residuals suggests that the residuals are normally distributed, but there is one extreme outlier (with a value larger than 3).





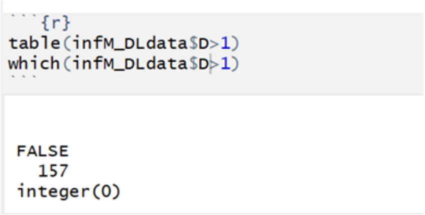
The residuals vs. fitted value can test the last assumption that exits equally variance. With the following figure, the residuals around the 0 line, this suggests that the relationship is linear is reasonable. It roughly forms a "horizontal belt" around 0 to indicate error terms are equal variance. But left corner has some outlier points away from the random pattern. Sometimes they are bad data and should be excluded. Under the $|y_i - \hat{y}_i| < 3$, we get the conclusion those outliers are ineffective. Our model fit the regression model assumptions.



Base on the model checking adequacy, the box-cox method check the $\lambda = 1$ at the lowest point, we don't need any transformation. The box-cox method figure are showed in Appendix A.

Influence point

On the regression analysis, it is possible to have a great influence on the result. So we detect influential observations and consider how it affects. Cook's distance measure how well the model fits the ith observation y_i and how far that point is from the remaining dataset (Harsha, 2019). If the D value smaller than 1, it indicates no influential point. Under the cook's distance method to check we can see the result like right. All values are less than one, there are no influence points. We can the model fit well under the current condition.



~~~
```{r}
table(infM_DLdata$D>1)
which(infM_DLdata$D>1)
```

FALSE
  157
integer(0)
~~~

Methods

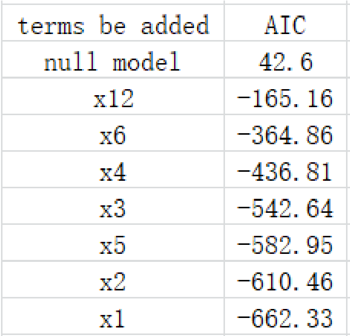
Stepwise Regression Methods

After we do the regression model assumptions checking, we would like to select an appropriate subset regression model; at the beginning, we tried stepwise regression methods; there are three types of stepwise regression methods; each method can gives us one subset regression model.

Forward selection

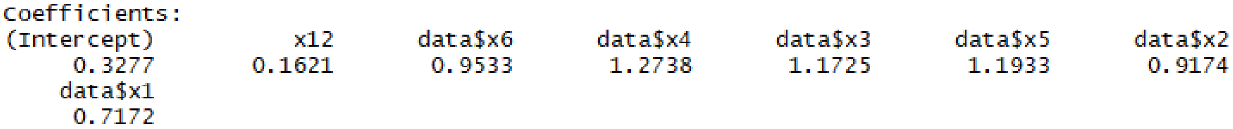
The forward selection is starting from a null model $y = \beta_0$, based on AIC criteria, each time add one regressor or do not add any regressor to make new subset model has no greater AIC value than the previous subset model.

The initial null model has AIC = 42.6; the first step add interaction term $x_1 x_2$would make AIC decrease to -165.16; the second step



terms be added	AIC
null model	42.6
x12	−165.16
x6	−364.86
x4	−436.81
x3	−542.64
x5	−582.95
x2	−610.46
x1	−662.33

add x_6 term, AIC = -364.86; then x_4, x_3, x_5, x_2, x_1 have been added successively; finally the appropriate subset model has minimum AIC value, that is -662.33. The subset model is:

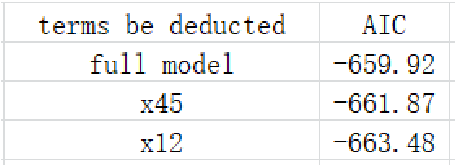


```
Coefficients:
(Intercept)          x12     data$x6     data$x4     data$x3     data$x5     data$x2
     0.3277       0.1621      0.9533      1.2738      1.1725      1.1933      0.9174
    data$x1
     0.7172
```

$y \sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_1x_2$; Only one interaction have been added to the model.

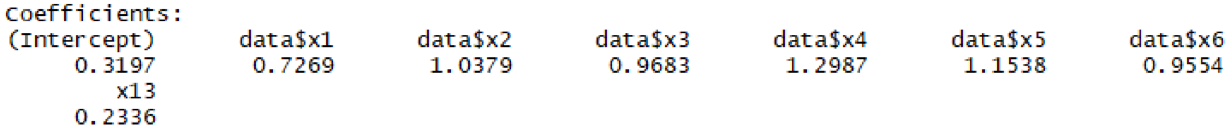
Backward elimination

The backward elimination method is starting from a full model; in our data, we have 9 regressors: x_1, x_2, x_3, x_4, x_5, x_6, x_1x_2, x_1x_3, x_4x_5; each time we deduct one regressor to decrease AIC of the model.



terms be deducted	AIC
full model	-659.92
x45	-661.87
x12	-663.48

The initial full model begin with AIC = -659.92; the first step, eliminate interaction term x_4x_5 can minimize AIC to -661.87; the last step, deduct x_1x_2 term can obtain minimum AIC of the model that is AIC = -663.48 for the final subset model selected by backward elimination method,



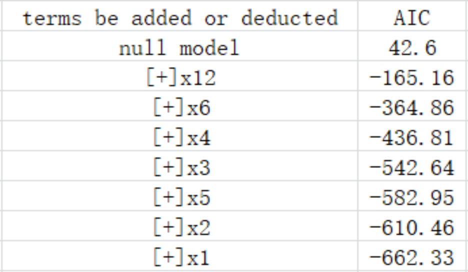
```
Coefficients:
(Intercept)      data$x1      data$x2      data$x3      data$x4      data$x5      data$x6
     0.3197       0.7269       1.0379       0.9683       1.2987       1.1538       0.9554
        x13
     0.2336
```

the appropriate subset model is: $y \sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_1x_3$

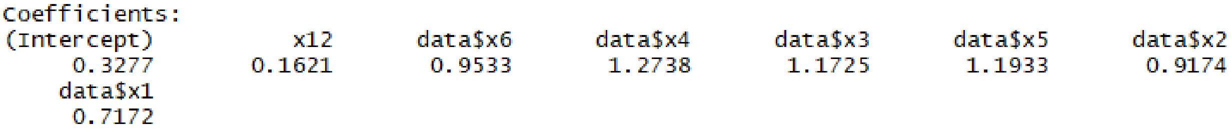
Stepwise Regression

The stepwise regression method involves both forward selection and backward elimination; we start from a null model $y = \beta_0$; the first time we can only add one term to make AIC of the new subset model smaller than the null model, after that, we can choose to add a new term to the current subset model or deduct a term from the current subset model.

Same as forward selection, it starts from null model, the first step add interaction term would decrease the AIC from 42.6 to -165.16; then second step could choose to add one more term or eliminate x_1x_2 term; in our dataset, keep adding more terms can decrease AIC continually; therefore, the final subset model that selected by stepwise regression is: $y\sim x_1+x_2+x_3+x_4+x_5+x_6+x_1x_2$. This subset model is the same as the one that selected by forward selection method.



terms be added or deducted	AIC
null model	42.6
[+]x12	-165.16
[+]x6	-364.86
[+]x4	-436.81
[+]x3	-542.64
[+]x5	-582.95
[+]x2	-610.46
[+]x1	-662.33

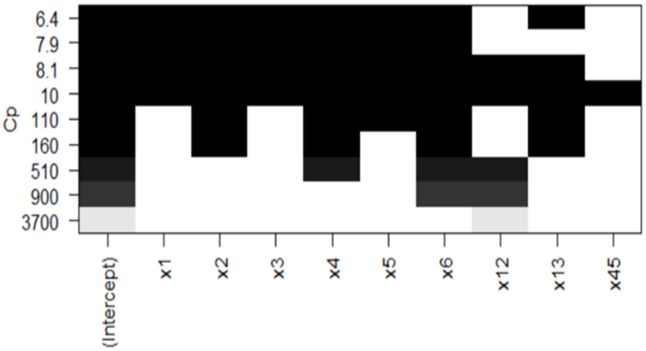


```
Coefficients:
(Intercept)          x12       data$x6       data$x4       data$x3       data$x5       data$x2
     0.3277       0.1621        0.9533        1.2738        1.1725        1.1933        0.9174
    data$x1
     0.7172
```

Criteria for Evaluating Subset Regression models

Mallow’s Cp Statistic

Mallow’s Cp criteria is to measure both variance and $bias^2$ of a model in order to get rid of model’s overfitting problem. When the number of regressors increase, the model would have smaller bias and larger variance; thus Mallow’s Cp is a way to balance variance and $bias^2$ to choose a better subset model. Usually the smaller Cp means the model is more accuracy; in our data set we perform Mallow’s Cp statistic, the minimum Cp = 6.4, which indicate that we want regressors in our model, the subset model $y\sim x_1+x_2+x_3+x_4+x_5+x_6+x_1x_3$ is the appropriate model we want by using Mallow’s Cp statistic.



Shrinkage Methods

Same as the Mallow's Cp statistic, both of the Ridge Regression and LASSO involve Bias-Variance Tradeoff as well, since we select a subset model according to the value of the criterion function; we would like to find a model with the minimum value of the criterion function. Both Ridge Regression and LASSO criterion functions have Mean-Squared error term and penalty term; the λ is a tuning parameter, it controls the strength of the penalty term in ridge regression and lasso regression. When $\lambda = 0$, no parameters are eliminated; as λ increase, more coefficients are set to zero and eliminated; when $\lambda = \infty$, all coefficients are eliminated. Also, as λ increase, bias increases and as λ decreases, variance increase. (Stephanie, 2017)

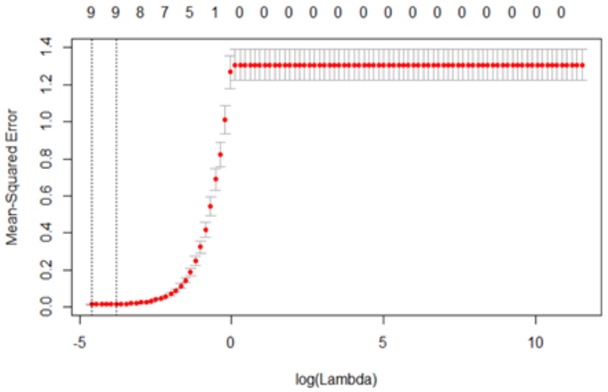
$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

The first formula is for Ridge Regression criterion function; the penalty term has tuning parameter and the sum of squared coefficients. The second formula is for LASSO criterion function; the penalty term has tuning parameter and the sum of absolute coefficients. Since LASSO has penalty term sum of absolute coefficients, parameters can equal to 0 faster; however ridge regression can only let parameters getting closer to 0; the additional figures of Lasso and Ridge regression plots are shown in the Appendix A. Thus LASSO is better for doing model selection.

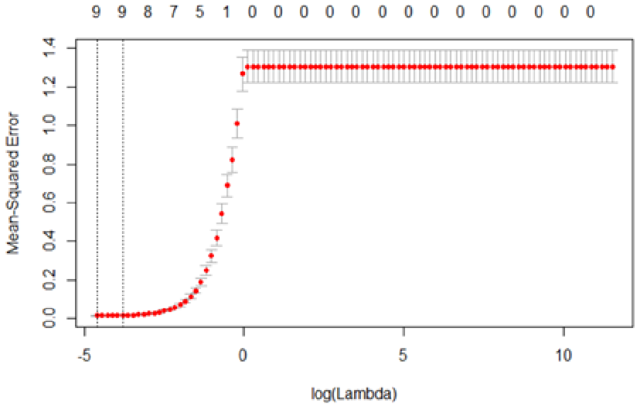
Ridge Regression

The minimum mean squared error is pointed out by the left dotted line; this line indicates that 9 regressors are selected; even if using the most parsimonious model, the right dotted line, we still get full model as the best model. The coefficients of the model that selected by Ridge regression is shown in Appendix A.



LASSO

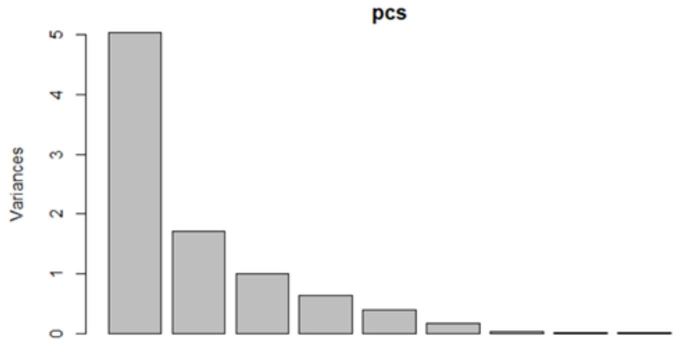
LASSO standard for least absolute shrinkage and selection operator; with the λ increase, coefficients set to 0 easier than Ridge regression. In this figure, the minimum MSE indicates the model we want is a full model, same as the most parsimonious model. Therefore we get the same result by using Ridge regression and LASSO; the coefficients of the model that selected by LASSO is shown in Appendix A.

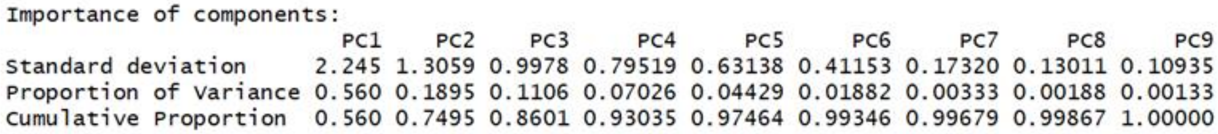


Other Method

Principal Component Analysis

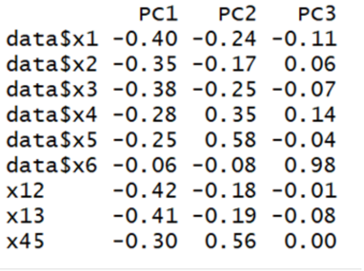
The principal Component Analysis is a method to reduce the dimension of the model, use less variables to represent 80% - 90% of the total variation of the original model. Our original model has 9 regressors, including 6 main effect terms and 3 interaction terms x_1x_2, x_1x_3, x_4x_5; thus the original model is $y\sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_1x_2 + x_1x_3 + x_4x_5$. Then we bring to lower dimension model to represent the less information of the original model; we use $Z_1 \ldots Z_m$ to represent m principal components; this number of m is less than 9; each principal component is taken to be a linear combination of all the regressors. We get the figure of each pc versus its variance; the first pc will always represent the greatest variance.





```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     2.245 1.3059 0.9978 0.79519 0.63138 0.41153 0.17320 0.13011 0.10935
Proportion of Variance 0.560 0.1895 0.1106 0.07026 0.04429 0.01882 0.00333 0.00188 0.00133
Cumulative Proportion  0.560 0.7495 0.8601 0.93035 0.97464 0.99346 0.99679 0.99867 1.00000
```

We have 9 principal components in total, from the summary output of the principal components, we can find that the first 3 principal components can explain 86% of the total variation; add one more pc can only increase 7% of the total variation; since 7% is not a significant increase, for getting a simpler model, we would like to use 3 principal components to build a new model.



```
           PC1   PC2   PC3
data$x1 -0.40 -0.24 -0.11
data$x2 -0.35 -0.17  0.06
data$x3 -0.38 -0.25 -0.07
data$x4 -0.28  0.35  0.14
data$x5 -0.25  0.58 -0.04
data$x6 -0.06 -0.08  0.98
x12     -0.42 -0.18 -0.01
x13     -0.41 -0.19 -0.08
x45     -0.30  0.56  0.00
```

So the loadings of the first 3 principal components is shown at the left side, then we can get principal components Z_1, Z_2, Z_3; each of the principal component represent some of the information from all the 9 regressors. Now our new model becomes to $y\sim Z_1 + Z_2 + Z_3$.

Conclusion

Reference

1. World Happiness Report 2016. (2016, March 20). Retrieved from https://worldhappiness.report/ed/2016/.

2. McNeney, B. (2018). Shrinkage Methods, Chapter 6 part 2. Retrieved from https://github.com/SFUStatgen/SFUStat452/tree/master/Notes2018/Ch06

3. Stephanie(July 29, 2017). Statistics How To. Retrieved from https://www.statisticshowto.datasciencecentral.com/tuning-parameter/

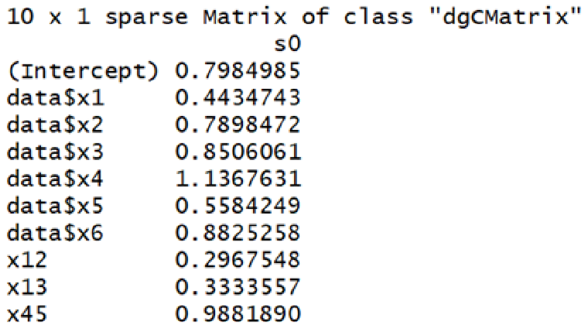
4. Perera, H. (November 7, 2019). Diagnostics for leverage and Influence. Retrieved from https://canvas.sfu.ca/courses/48266/files/folder/unfiled?preview=10843398

5. McNeney, B. (2018). Dimension Reduction Methods, Chaper6 part 3. Retrieved from https://github.com/SFUStatgen/SFUStat452/tree/master/Notes2018/Ch06

Appendix A

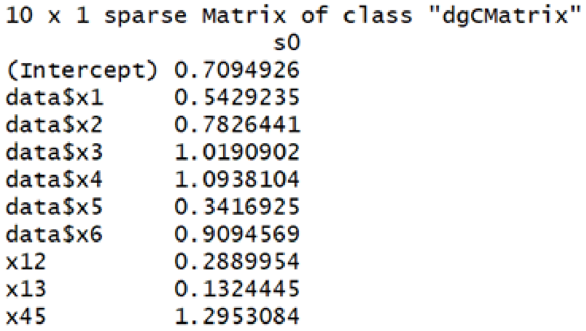
Additional tables

Coefficients of ridge regression



```
10 x 1 sparse Matrix of class "dgCMatrix"
                    s0
(Intercept) 0.7984985
data$x1     0.4434743
data$x2     0.7898472
data$x3     0.8506061
data$x4     1.1367631
data$x5     0.5584249
data$x6     0.8825258
x12         0.2967548
x13         0.3333557
x45         0.9881890
```

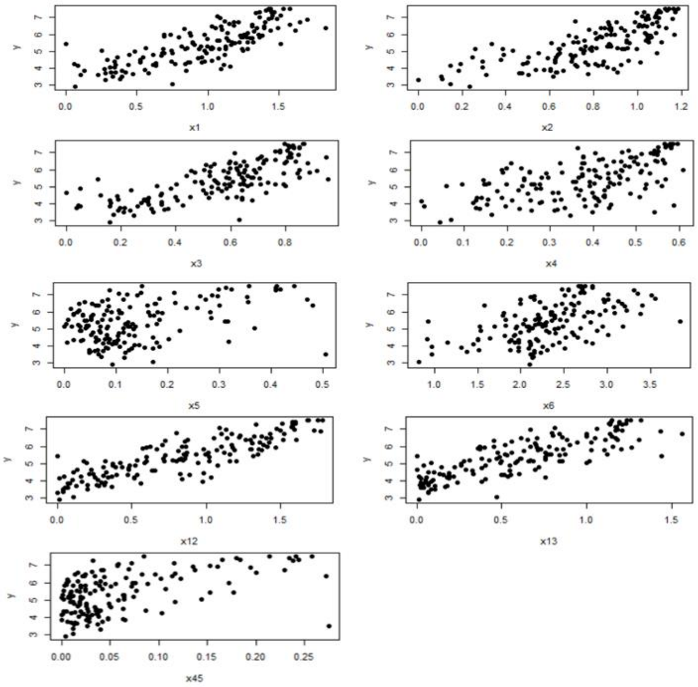
Coefficients of LASSO



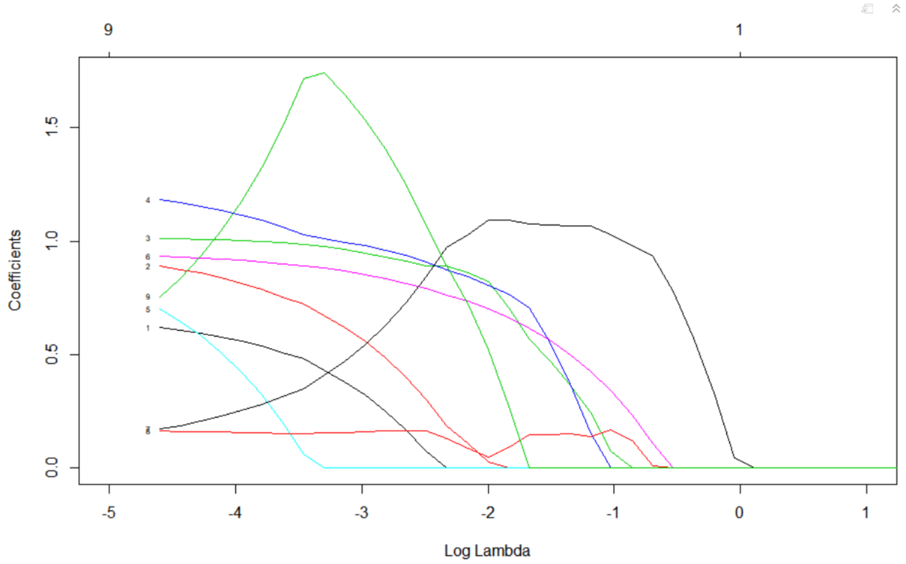
```
10 x 1 sparse Matrix of class "dgCMatrix"
                    s0
(Intercept) 0.7094926
data$x1     0.5429235
data$x2     0.7826441
data$x3     1.0190902
data$x4     1.0938104
data$x5     0.3416925
data$x6     0.9094569
x12         0.2889954
x13         0.1324445
x45         1.2953084
```

Additional figures

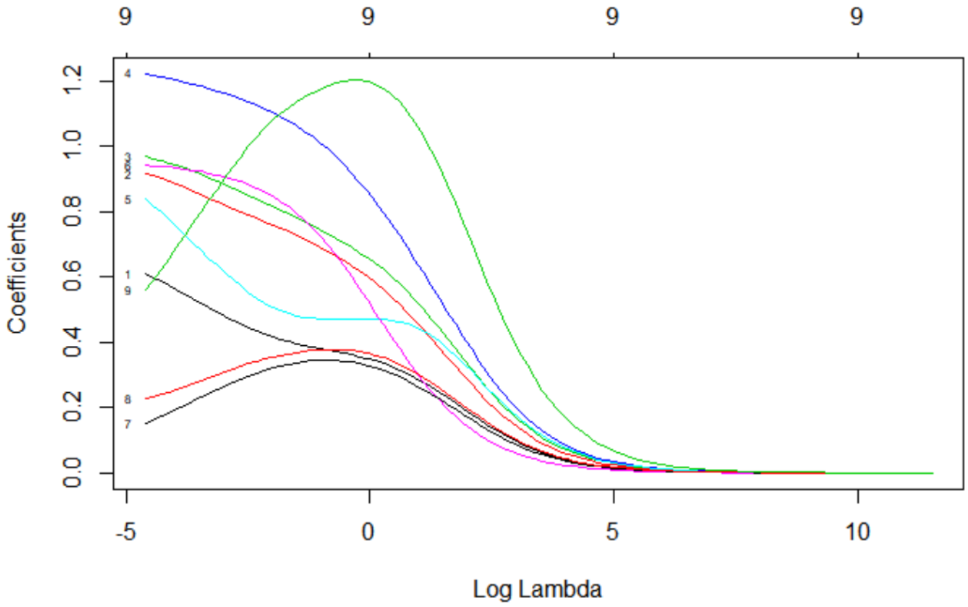
Linear trend plot



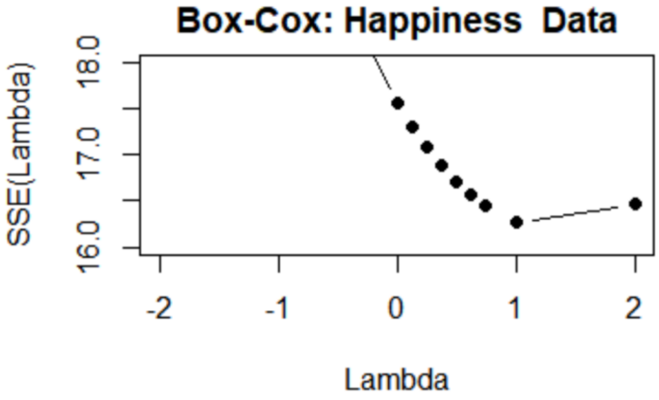
Lasso plot



Ridge plot



Box-Cox Method figure :


Box-Cox: Happiness Data
SSE(Lambda)
Lambda