World happiness

Group 21

2021/6/24

Introduction

Exploratory Data Analysis

Visualization of the data

Formal Data Analysis

To begin to analysis the world happiness dataset, we need to check the correlation between the six explanatory variables to avoid the problem of multicollinearity.

Table 1: Correlation coefficient table of explanatory variables.

	LoggedGDP	Social	expectancy	Freedom	Generosity	corruption
LoggedGDP	1.0000000	0.7852987	0.8594606	0.4323235	-0.1992864	-0.3423374
Social	0.7852987	1.0000000	0.7232561	0.4829298	-0.1149459	-0.2032070
expectancy	0.8594606	0.7232561	1.0000000	0.4614939	-0.1617503	-0.3643735
Freedom	0.4323235	0.4829298	0.4614939	1.0000000	0.1694374	-0.4013630
Generosity	-0.1992864	-0.1149459	-0.1617503	0.1694374	1.0000000	-0.1639617
corruption	-0.3423374	-0.2032070	-0.3643735	-0.4013630	-0.1639617	1.0000000

From our correlation table we can see that the correlation between our Logged GDP and Healthy life expectancy is 0.859, which is a strong positive linear relationship.And the Logged GDP and Social support also have the high degree of collinearity, the correlation between this two variables is 0.785.So we remove Healthy life expectancy and Logged GDP.Then, using the remaining 4 explanatory variables to perform stepwise regression, and observe whether the remaining variables need to be eliminated.

```
Start:  AIC=-141.88
score ~ Social + Freedom + Generosity + corruption

             Df Sum of Sq    RSS      AIC
- Generosity  1     0.091 53.856 -143.628
<none>                    53.765 -141.879
- corruption  1     6.078 59.843 -127.920
- Freedom     1     6.364 60.129 -127.210
- Social      1    43.953 97.718  -54.857

Step:  AIC=-143.63
```

```
score ~ Social + Freedom + corruption

             Df Sum of Sq     RSS      AIC
<none>                     53.856 -143.628
- corruption  1     5.988  59.844 -129.919
- Freedom     1     6.325  60.181 -129.082
- Social      1    47.398 101.254  -51.561


Call:
lm(formula = score ~ Social + Freedom + corruption, data = happiness)

Coefficients:
(Intercept)       Social      Freedom   corruption
     0.0779       5.6256       2.2271      -1.2254
```

According to the results of stepwise regression, we choose the model with the smallest AIC as the final model.Then,we fit the following linear model to the data.

$$\widehat{\text{score}}_\text{i} = \widehat{\alpha} + \widehat{\beta} * \text{Social}_\text{i} + \widehat{\gamma} * \text{Freedom}_\text{i} + \widehat{\delta} * \text{corruption}_\text{i}$$

where

- the $\widehat{\text{score}}_\text{i}$: the happiness score of the ith country.
- the $\widehat{\alpha}$: the intercept of the regression line.
- the $\widehat{\beta}$: the coefficient for the first explanatory variable Social.
- the $\widehat{\gamma}$: the coefficient for the second explanatory variable Freedom.
- the $\widehat{\delta}$: the coefficient for the second explanatory variable corruption.

When this model is fitted to the data, the following estimates of α (intercept) and β,γ and δ are returned:

Table 2: Estimates of the parameters from the fitted linear regression model.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.078	0.559	0.139	0.889	-1.028	1.184
Social	5.626	0.498	11.297	0.000	4.641	6.610
Freedom	2.227	0.540	4.127	0.000	1.160	3.294
corruption	-1.225	0.305	-4.015	0.000	-1.829	-0.622

According to this table, the coefficient for social support tells us that, taking all other variables in the model into account and holding them constant, there is an associated increase, on average,every increase of 1 unit in the social support score increases the happiness index score by approximately 5.63 units.In the same way, when the freedom score of life choice increases by 1 unit, the happiness index score also increases by approximately 2.23 units.On the contrary, for every increase of 1 unit in the score for corruption, the total score of happiness index decreases by 1.23 units.

Before we can continue to use the fitted model, we must check the model's assumptions. It is best to consider these according to the residual plot in Figure 2.

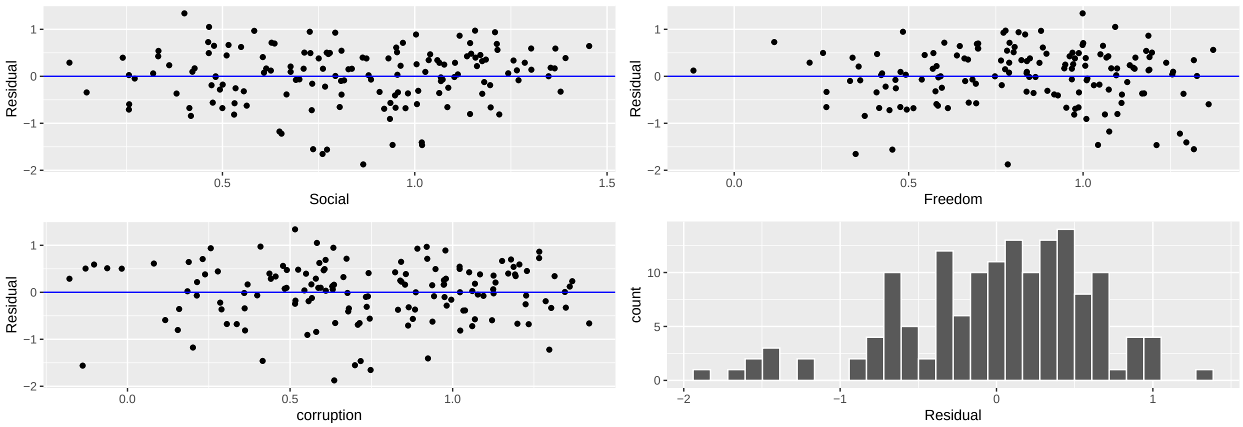

Figure 1: Scatterplots of the residuals by Social,Freedom,corruption and a histogram of the residuals

The assumptions of the residuals having mean zero and constant variability across all values of the explanatory variable appear to be valid in this case.According to the three different explanatory variables scatter plots, it can be concluded that the residuals are uniformly distributed above and below the zero line, so the mean is 0. The residuals are randomly distributed around the zero line, and the distribution of the residuals is constant across all fitted values, so there is no obvious pattern or change in the variant.And also the histogram supports the assumption of normal distribution error.

Conclusions