Assignment

Computational Methods and Principles of Bayesian Inference - DATA.STAT.430-2021-2022-1-TAU

Duy Vu – H292118

1. Dataset

The dataset used for this assignment/coursework is the Maailmanpankin väestötilastoja (tentatively translated in English as World Bank population statistics) 2016 - 2018 from the Finnish Social Science Data Archive (FSD). The dataset contains basic information about the countries of the world, other territories and groups formed by countries. In total, there are 30 variables, most of which are from World Bank indicators. Some example variables are birth rate, morality, adult literacy rate by male and female, expenditure on education and military, and GDP per capita.

In this assignment, only two variables are taken into consideration: infant mortality rate (per 1000 live births) and life expectancy (at birth). By definition from the data source, infant mortality rate is the number of infants dying before reaching one year of age, per 1000 live births, and life expectancy is the number of years a new-born infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life. Also, the data used in this assignment is taken specifically from the year 2017.

According to The World Bank, these two variables, along with other variables like mortality rate are among the most important indicators of health status of a country, and high mortality in young age groups of a country tends to significantly lower the life expectancy of that country. Hence, a causal relationship may exist between these two variables.

Since the data has not yet been processed, before analysing the data, some pre-processing steps are needed to be carried out. The data is stored in a tabular format where all countries, territories, constituent countries, autonomous entities, regions, continents, and even the World itself have data in their own rows and data of 30 variables of World Bank indicators are presented in columns. Since the table consists of data region and groups of countries which aggregate data of separated countries and entities together, these data should not be used. Moreover, row do not have data for the two variables we concern about are also removed from the table.

After these steps, we end up with a table contains life expectancy and infant mortality rate of 174 countries. Since not all statistical parameters can be calculated easily from this amount of data, only 35 random countries are selected for further analysis.

2. Statistical basis of the method applied to the selected data

Before going deep into any sophisticated statistical methods, we first need to calculate some simple descriptive statistics and visualize the dataset.

	Infant morality	Life expectancy
Minimum	2.00	58.06
1st quantile	7.00	68.91
Median	15.40	72.64
Mean	21.58	72.33
3rd quantile	31.95	75.95
Maximum	64.30	82.95

Table 1: Summary of data used for analysis (obtained from R code)

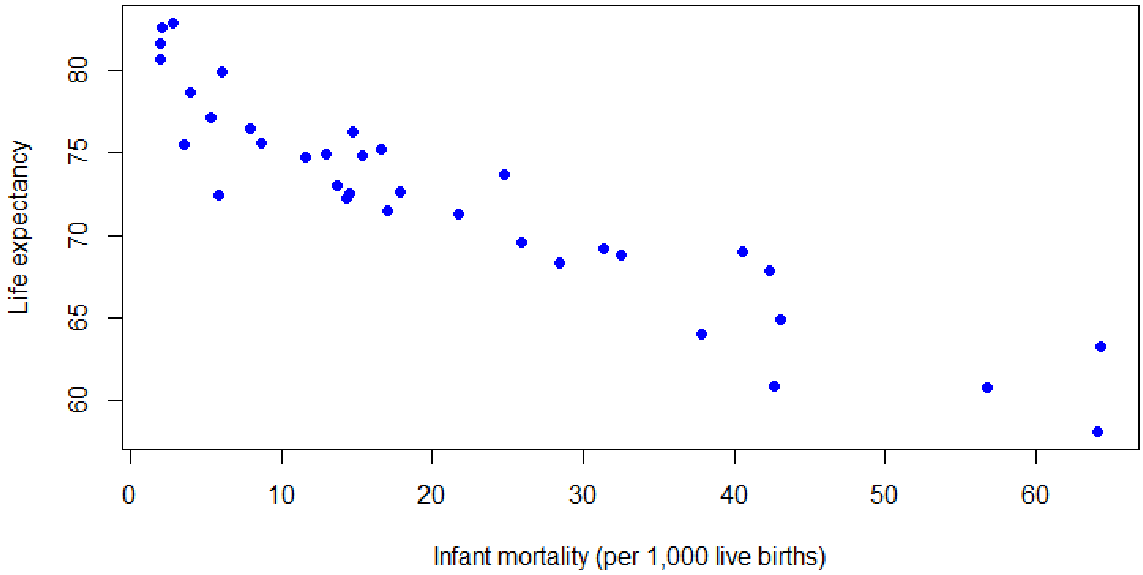

Figure 1: Visualization of infant mortality and life expectancy

The figure above does show a strong negative linear correlation between 2 variables, and as the matter of fact, the Pearson correlation coefficient of this sample dataset is around -0.92.

The numbers and figure above can lay a good foundation for some statistical methods to be applied to the dataset:

A. Use permutation test to conduct of hypothesis testing where the null hypothesis is that the correlation coefficient is equal to 0, and the alternative hypothesis is otherwise. The test is conducted by calculating all possible values of the test statistic, correlation coefficient in our case, under all possible rearrangements, permutations of the observed data points. The p-value of the test is obtained by calculating the probability that the correlation coefficient of permutation is larger than the absolute value (two-tailed test) of observed correlation coefficient from the initial sample which is -0.92. If the p-value is small enough (given some threshold we choose), we can say the observed alternative hypothesis is statistically significant.

B. Use ordinary least squares (OLS) to estimate linear regression parameter of the model.

   Since our dataset exhibits a strong linear correlation from the plot of dataset, Pearson correlation coefficient, and especially the null hypothesis that the correlation coefficient equal to zero is rejected using permutation test above (result and conclusion about this experiment is in the next section), it makes sense to try to find the parameter for linear regression model that fits the dataset the best. Basically, OLS tries to minimize the sum of squared distance from real values to the estimated values of the regression model. Mathematically, we have:

$$y = X\beta + \varepsilon,$$

   where y is the vector of observed output, X can be a multi-dimensional input matrix, but in our case, it is also just a vector, ε is the vector of error, and β is what we need to estimate. Normally, we also add a column vector of 1 on the right of X and add 1 more parameters to β to estimate. This parameter is also called the intercept.

   The objective function to this problem is:

$$\min(\|y - X\beta\|^2),$$

   and the unique solution is:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

C. Besides fitting regression model, another statistical method can be applied is estimating and evaluating the variance of an estimator of correlation coefficient by resampling method called Bootstrap and Jackknife. Even though both sharing those same purpose, there resampling methods are entirely different. While in Jackknife, *n* leave-one-out sample set (only containing *n-1* samples) where *n* is the sample size are collected, in Bootstrap, the resampling method is sampling with replacement (still containing *n* samples) and there is no limit on the number of times we can resample. Each with its own resampling method can then be used to form a distribution of parameters, evaluate, and give a 95% confidence interval (CI) of the estimator. Besides the difference in sampling method, Jackknife doesn't work well with on-smooth statistics like the median and nonlinear statistics like the correlation coefficient, and, in general, it produces larger estimated standard errors than Bootstrap does.

3. Results

A. Hypothesis testing

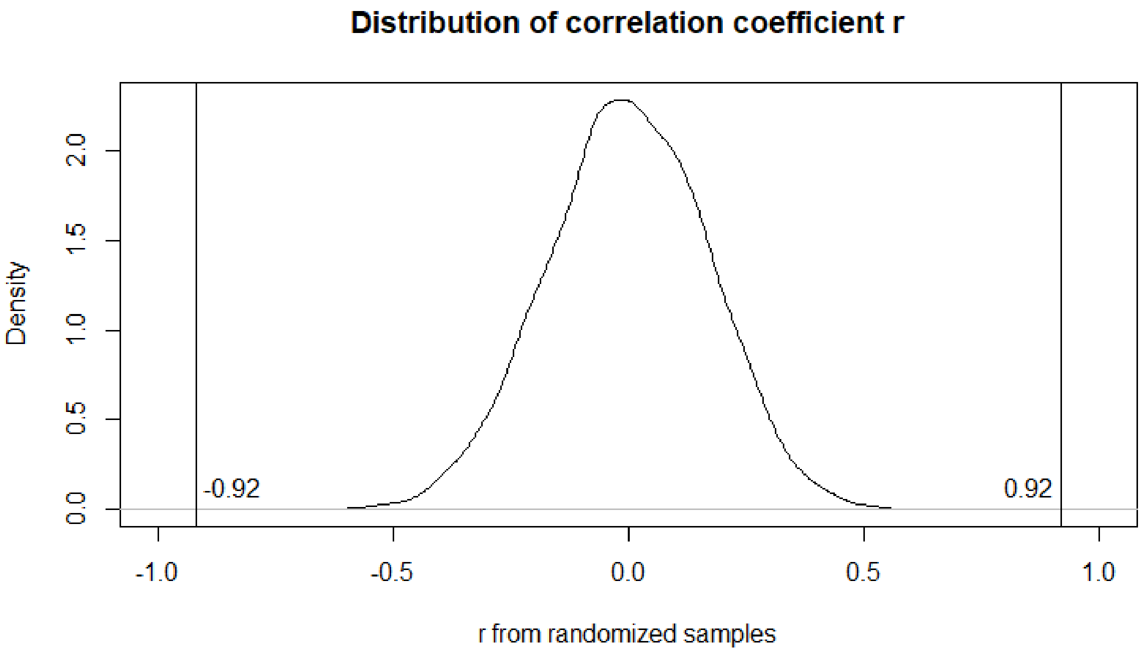

Figure 2: Distribution of correlation coefficient of permutations

Just by observing the figure above, we can confidently reject the null hypothesis. And as a matter of fact, with 10,000 permutations, the probability of having correlation coefficient outside the range of absolute value of 0.92 is 0. Therefore, we can reject the null hypothesis.

B. Estimation of linear regression parameter

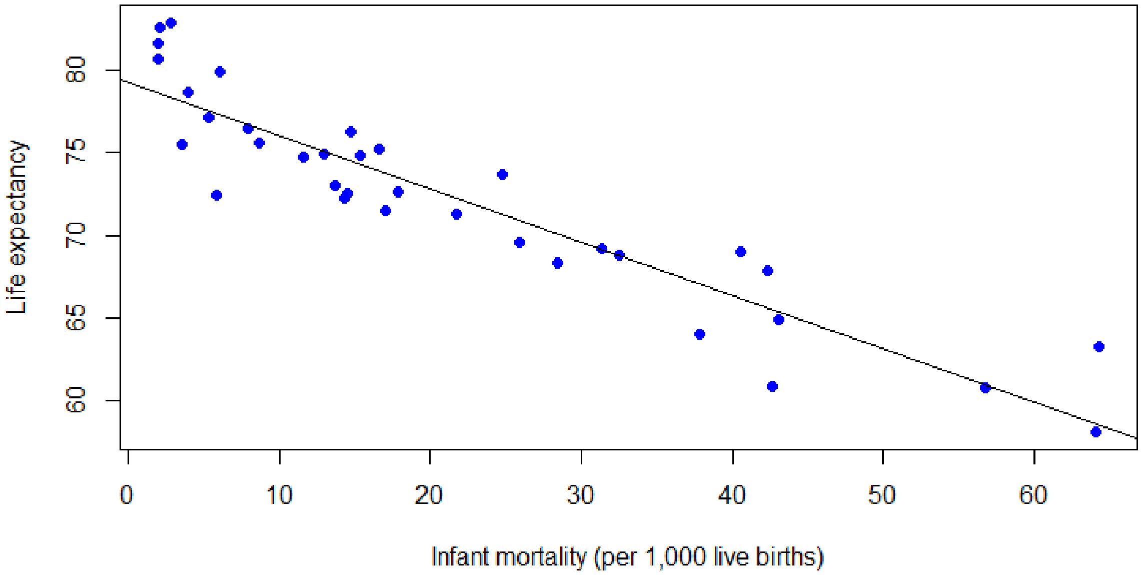

Figure 3: Best fit line $y = -0.32x + 79.33 + \varepsilon$

With the residual standard error of 2.422 on 33 degrees of freedom

C. Estimate correlation coefficient

For Bootstrap method, the number of bootstrap samples is 1000, and the CI method is the percentile CI.

	Jackknife	Bootstrap
Estimator value	-0.9237229	-0.9245287
95% confidence interval	(-0.9700611, -0.8773847)	(-0.9603, -0.8762)

In either case, the estimator is very close to each other, and we can confidently use -0.92 as the correlation coefficient of the whole population dataset.