Project Report

*Estimation and Data Analysis*

Sara Montese

**Broad Exploratory Data Analysis - Part B**

Dataset description

For this part, I chose a dataset from www.kaggle.com/datasets. This dataset was created by combining 5 heart datasets (over 11 common features) already available independently but not combined before. This dataset contains 11 features that can be used to predict a possible heart disease: Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST-Slope, HeartDisease.

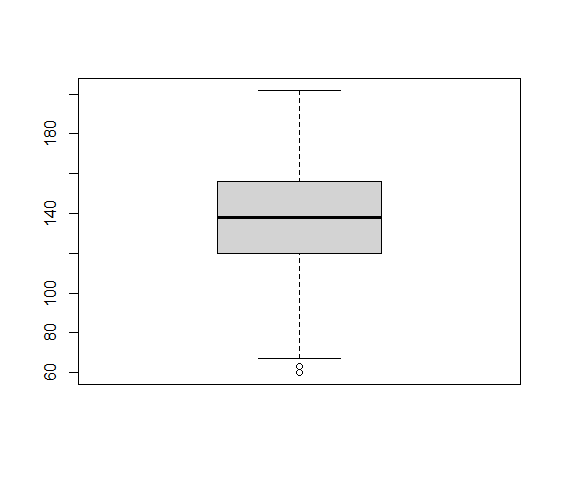
I decided that the role of population data is taken by one numeric field: MaxHR (maximum heart rate achieved).

Total observations: 918

Kaggle link: <https://www.kaggle.com/fedesoriano/heart-failure-prediction>

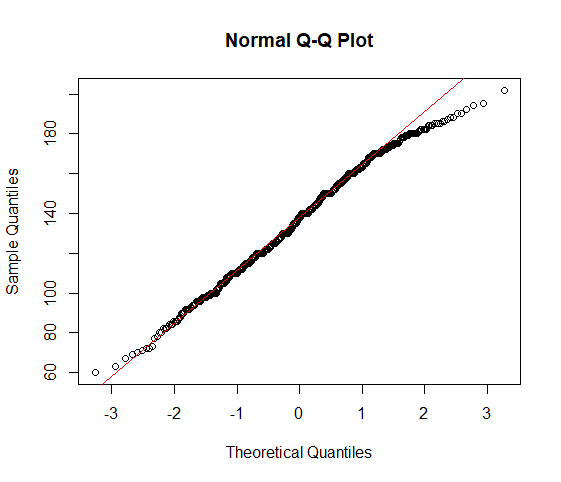
Repository link: <https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>

Distribution of the data

To investigate whether the data is normally distributed or not, I applied 3 different methods: 2 visual and 1 statistical.

1. Visual -> I use a visualization of the data like a boxplot

As visible from the image, although the box-plot is not perfectly symmetric, there is no clear violation of normality.



1. Visual -> I use a quantile-quantile plot

From the plot we notice that the data falls along the dotted line (expected quantiles of a normal distribution). Since there are no strong deviations from the dotted line, the assumption of normality is not violated.

1. Statistical -> I perfom the most widely used statistical test for normality: the Shapiro-Wilk test.

I found a p-value of 0.0001683 < 0.05: the test is significant. I may reject that null hypothesis and so accept the alternative hypothesis that the data is not sampled from a Gaussian population. Even after removing the outliers of the set, the result of the test doesn’t change.

The visual and statistical methods give different results. However, since the dataset is rather large, the Shapiro-Wilks normality test is no longer very useful, because for large amounts of data even very small deviations from normality are detected. This will lead to the rejection of the null hypothesis, even though for practical purposes I can accept this level of normality.

I conclude that the data is approximately normally distributed.

|  |  |  |  |
| --- | --- | --- | --- |
| Measures of central tendency | Mean | Median | Mode |
| Population data | 137 | 138 | 150 |
| Sample data | 138.8 | 140 | 170 |

The population distribution is negatively skewed or left-skewed: the values fall towards the higher side of the scale. [ mean < median < mode ]

This can be said also looking at the histogram: the tail on the left side of the distribution is longer than the right side.

Even the sample distribution, as shown by the histogram and boxplot, is not symmetrical.

[ mean < median < mode ]

|  |  |  |  |
| --- | --- | --- | --- |
| Measures of dispersion | Variance | Standard deviation | Coefficient of Variation |
| Population data | 636.5211 | 25.22937 | 18.41912 % |
| Sample data | 643.1667 | 25.36073 | 18.27142 % |

Variance and standard deviation of the population data are slightly lower than the sample values.

As the sample size increases, indeed, the sampling distributions approach a normal distribution, and the mean of the sampling distribution approaches the population mean (µ).

**Hypothesis testing - Part C**

Dataset description

For this part, I chose two datasets from ourworldindata.org. They represent the total government expenditure on education, from 1970 to 2019, for Italy and United Kingdom. The amount is given as a share of GDP.

Data published by: World Development Indicators - World Bank (2021.07.30)

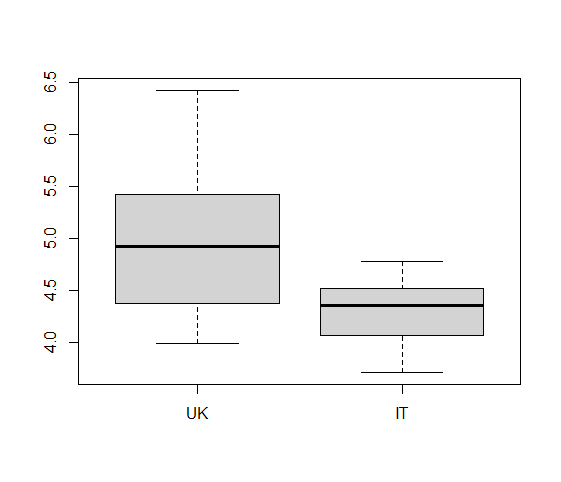
Data publisher's source: UNESCO

Total observations for Italy: 36

Total observations for UK: 42

Link: <https://ourworldindata.org/grapher/total-government-expenditure-on-education-gdp>

Hypothesis tests

By observing the boxplots, I can already say that the median expenditure in education of the UK is greater than in Italy.

Besides, the interquartile range and the range of the data set is greater for the UK than Italy (as shown by the length of the boxes and the distances between the ends of the two whiskers for each boxplot).

By displaying the numerical summary of each dataset, I can see that also the mean expenditure in education of the UK is greater than in Italy.

I then perform a hypothesis test in R.

H0: IT mean is equal to UK mean

H1: IT mean is different from UK mean

Result of the test: p-value=3.353e-07 ( <0.05)

A small p-value means that there is strong evidence in favor of the alternative hypothesis. As expected, the two means are different.

**Linear Regression - Part D**

Dataset description

For this part, I chose a dataset from [www.infodata.ilsole24ore.com](http://www.infodata.ilsole24ore.com) (Italian newspaper). It contains, for each Italian region, two features:

* % Population with at least 1 dose of vaccine (at 13/12/21)
* % Population with a degree

Total observations: 21

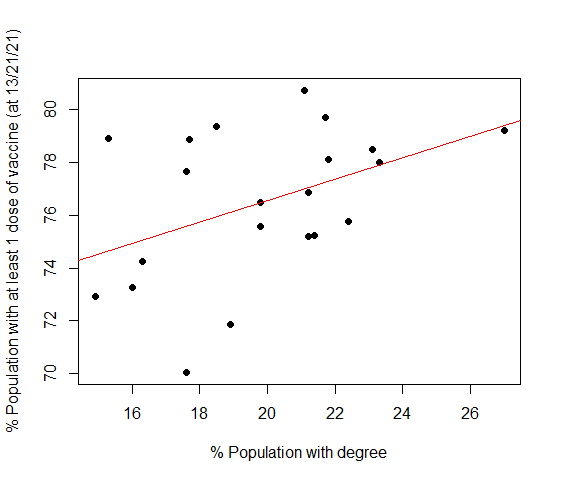
Source: Eurostat, Special Commissioner for Covid-19 emergency in Italy

Link: <https://www.infodata.ilsole24ore.com/2021/12/27/perche-alcune-persone-decidono-non-vaccinarsi-la-correlazione-titolo-studio/>

*Hypothesis*: the % of vaccinated people in a region depends on the education level of the region itself.

*Predictor variable*: % population with a degree

*Response variable*: % Population with at least 1 dose of vaccine (at 13/21/21)

****

Linear Regression Model

Coefficient of correlation: 0.4336502

The coefficient indicates a weak positive relationship between the variables.

The regression line's equation:

**% Vaccinated population =**

**0.4052854 \* % Population with degree + 68.45962**

From the analysis of the linear regression model:

* The p value for the slope tells us that the independent variable doesn't have strong statistical predictive capability (\*)
* The R-squared indicates that only about 20% of the dependent variable is predicted by the independent variable - which is weak;

Although R-squared can gives an idea of how strongly associated the predictor variables are with the response variable, it doesn't provide a formal statistical test for this relationship.

This is why the F-Test is useful since it is a formal statistical test.

Since the p-value is almost equal to 0.05, the F-test is considered on the borderline of statistical significance, and so the correlation between the independent variable and dependent variable.

The fitted vs residuals plot is very useful to visualize the suitability of the linear regression model.

The residuals do not really "bounce randomly" around the 0 line. This may suggest that the assumption that the relationship is linear is not very reasonable.

The residuals do not form an approximate "horizontal band" around the 0 line. This suggests that the variances of the error terms are not equal.

There are some residuals that "stand out" from the basic random pattern of residuals. This suggests that there could be some outliers.

In conclusion, by analyzing the residual plot I can't validate the regression model. It may be necessary to have a bigger dataset.