Do Basic Drinking Water Services And Numbers of Nurses and Midwives Affect People's Life Expectancy?

Introduction

Research Question

Robin (2011) mentioned that access to safe water supplies have significantly positive effects on life expectancy and also fertility has a significant negative effect on it. Robin's consequences are convincing, based on his previous research and available data, this study will attempt to construct a Multiple Regression Model to explore the relationship between people's life expectancy and 2 potential variables(basic drinking water services and numbers of nurses and midwives) and verify whether the research results of previous scholar are correct.

Dataset

The data set used in this research comes from the 2017 World Bank IQM Data Set, which is made up with a sample of 180 countries. The dataset includes 29 variables – this includes 28 national development indicators and one variable indicating ‘country’ (categorical).

Variables and their Summary Statistics

Dependent Variable: `Lifeexpectancy`

The variable `Lifeexpectancy` in selected dataset represents the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life, which is consistent with the definition of Robin's research.

Use `summary()` function in R language, knowing its summary statistics are as follows(`Min.` and `Max.` represent its minimum and maximum value, `Median` and `Mean` represent its median and arithmetic mean):

```
> summary(data_NA.rm$Lifeexpectancy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  59.89   74.32   76.25   76.03   80.98   82.90 
```

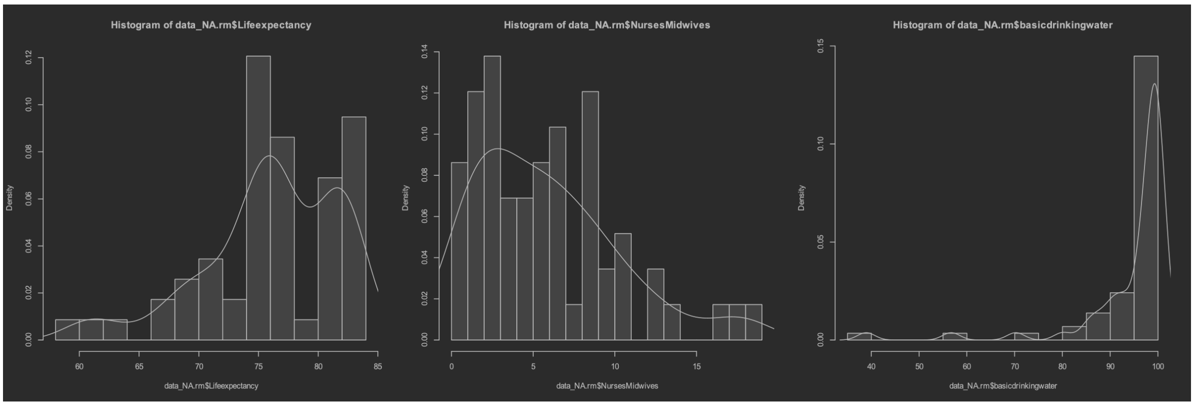
Explanatory Variables: `basicdrinkingwater` and `NursesMidwives`

The variable `basicdrinkingwater` indicates the percentage of people using at least basic water services and `NurseMidwives`includes the number of nearly all kinds of nurses and midwives. According to Wilson (2011), nurses and midwives play a important role in fertility journey, so this research select `NurseMidwives` as the alternative to solve the lack of fertility rates. Their summary statistics are as follows(the meanings are above):

```
> summary(data_NA.rm$NursesMidwives)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.240   2.518   5.198   5.841   8.389  18.230 
> summary(data_NA.rm$basicdrinkingwater)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  38.92   93.84   98.44   94.59   99.99  100.00 
```

Histograms

The following histograms and lines indicates the univariate distributions of variables mentioned above:



The Analysis

Correlation analysis

To verify if each explanatory variable has a liner relationship with dependent variable, this research analyzed their correlation separately, the results are following:

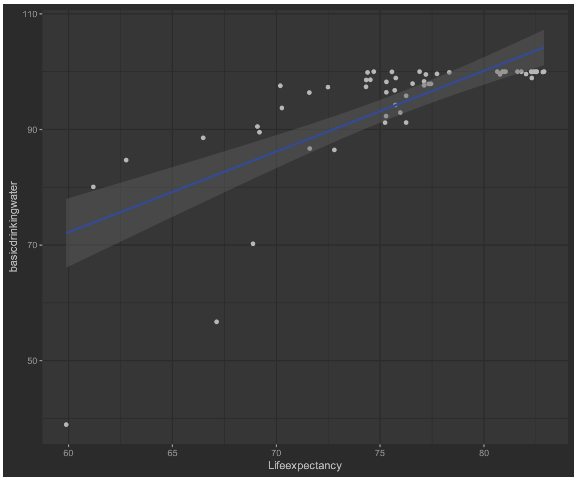
`Lifeexpectancy` and `basicdrinkingwater`

```
> cor.test(data_NA.rm$Lifeexpectancy,data_NA.rm$basicdrinkingwater)

	Pearson's product-moment correlation

data:  data_NA.rm$Lifeexpectancy and data_NA.rm$basicdrinkingwater
t = 8.0211, df = 56, p-value = 7.137e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5830096 0.8322993
sample estimates:
      cor 
0.7311954 
```

The pearson's correlation coefficient is 0.731 and the p-value is less than 0.001, which indicate that there is a significant linear positive correlation between these two variables, the scatterplot is following:



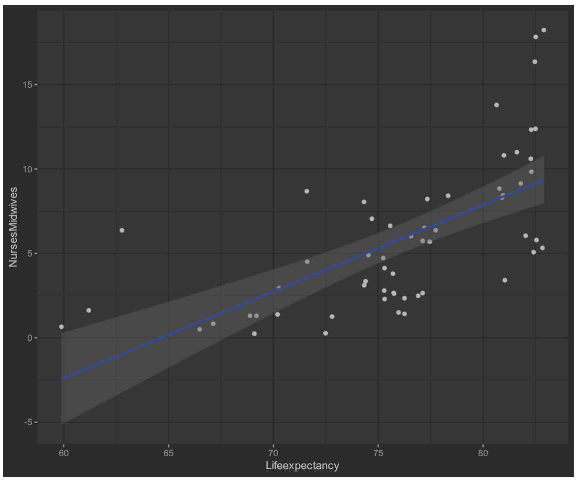
`Lifeexpectancy` and `NursesMidwives`

```
> cor.test(data_NA.rm$Lifeexpectancy,data_NA.rm$NursesMidwives)

    Pearson’s product-moment correlation

data:  data_NA.rm$Lifeexpectancy and data_NA.rm$NursesMidwives
t = 6.4726, df = 56, p-value = 2.563e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4763802 0.7805804
sample estimates:
     cor
0.654181
```

The pearson's correlation coefficient is 0.654 and the p-value is less than 0.001, which indicate that there is a significant linear positive correlation between these two variables, the scatterplot is following:



Multiple Linear Regression Model

The predict model is:

$$Lifeexpectancy = \beta_0 + \beta_1 basicdrinkingwater + \beta_2 NursesWidwives + \epsilon$$
$$\epsilon \sim (0, \sigma^2)$$

After removing the rows which contain `NA` values, the regression model results are as following:

```
> model=lm(Lifeexpectancy ~ NursesMidwives + basicdrinkingwater,data =
data_NA.rm)
> summary(model)

Call:
lm(formula = Lifeexpectancy ~ NursesMidwives + basicdrinkingwater,
    data = data_NA.rm)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6812  -1.4074   0.2257   1.9327   5.5329

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        45.69292    4.02129  11.363 4.72e-16 ***
NursesMidwives      0.52820    0.10912   4.840 1.09e-05 ***
basicdrinkingwater  0.28810    0.04477   6.435 3.16e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.264 on 55 degrees of freedom
Multiple R-squared:  0.6737,    Adjusted R-squared:  0.6618
F-statistic: 56.77 on 2 and 55 DF,  p-value: 4.224e-14
```

P-values of three coefficients are less than 0.001, and the adjusted R^2 is 0.6618, so the results are relatively significant.

The model is:

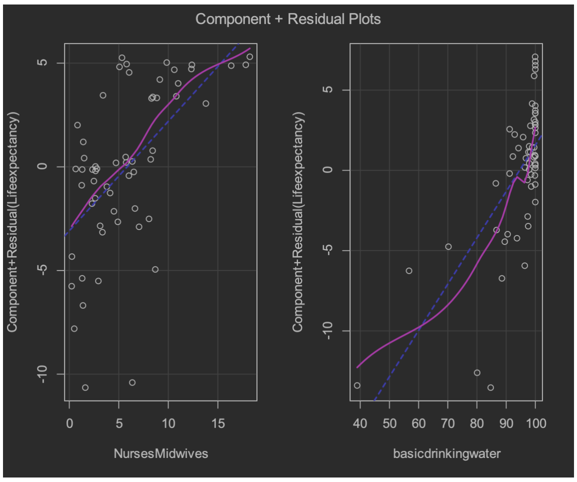
$$Lifeexpectancy = 45.7 + 0.29basicdrinkingwater + 0.53NursesWidwives + \epsilon$$

$$\epsilon \sim (0, \sigma^2)$$

Regression Model Analysis

Linearity

Using function `crPlots()` from car library to test the linearity of the model, the plot are following:



there are significant nonlinear relationships which indicate the non-linear transformation of explanatory variables is needed.

Homoscedasticity

Using function `ncvTest()` from car library to test the linearity of the model, the results are following:

```
> ncvTest(model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 4.810314, Df = 1, p = 0.02829
```

The p-value is less than 0.05, which rejects the hypothesis of homoscedasticity.

Independence

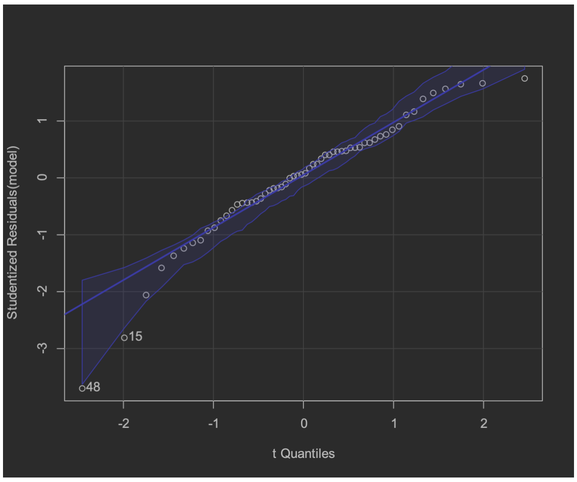
Using function `durbinWatsonTest()` from car library to test the linearity of the model, the results are following:

```
> durbinWatsonTest(model)
 lag Autocorrelation D-W Statistic p-value
   1      -0.1574487      2.289805   0.206
 Alternative hypothesis: rho != 0
```

The p-value is lager than 0.05, which indicates that two explanatory variables are independent.

Normality

Using fuction `qqPlot()` from car library to draw the Q-Q plot,the result is following:



Several points are away from the line, which rejects the hypothesis of normality.

Further using function `shapiro.test()` to test the normality, p-value is 0.022, which significantly rejects the hypothesis of normality.

```
> shapiro.test(model$residuals)

	Shapiro-Wilk normality test

data:  model$residuals
W = 0.95177, p-value = 0.02194
```

Multicollinearity

Using function `vif()` from car library to test the multicollinearity, the result is following:

```
> vif(model)
   NursesMidwives basicdrinkingwater
         1.234314           1.234314
```

The VIFs of two variables are relatively small, it can be considered that the problem of multicollinearity does not exist.

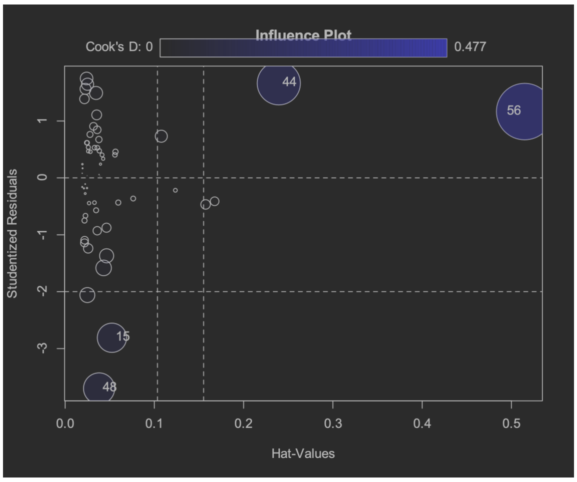
Outliers and Other Values

Using function `outlierTest()` from car library to test the outliers, find one outlier:

```
> outlierTest(model)
    rstudent unadjusted p-value Bonferroni p
48 -3.702036         0.00050337     0.029196
```

Draw an influence plot, and find there are some leverage point and influential point:

```
influencePlot(model,id.method="identity",main="Influence Plot")
```



Result

Conclusion from the Model

As Robin's research, basic water services have a significant positive effect on people's life expectancy, but in contrast of his conclusion, the number of nurses and midwives also have significant positive effects on people's life expectancy. The results indicate that if government wants to increase people's life expectancy, they shouldn't ignore the significance of drinking water supplies and the training of related medical stuff.

Testing of the Model

Disappointingly, the model is far away from an optimal model, most of the linear regression model testing failed, which means we need optimized the model a lot. Predicted direction of optimization is that the relationship between these variables are non-linear, we need to find the suitable non-linear transformation. Also the number of samples is too small, we need to include more data from more years and countries.

References

Barlow R, Vissandjee B. Determinants of national life expectancy[J]. Canadian Journal of Development Studies/Revue canadienne d'études du développement, 1999, 20(1): 9-29.

Wilson C, Leese B. Do nurses and midwives have a role in promoting the well-being of patients during their fertility journey? A review of the literature[J]. Human Fertility, 2013, 16(1): 2-7.