ISM 6137: Statistical Data Mining

Analysis of Health Quality and Health Expenditure

Submitted by

Anusha Seshagiri

Nivedita Mangal

Vinay Kumar Reddy Baradi

Sai Suraj Argula

Submission Date

5th May 2020

Project Guide

Dr. Anol Bhattacherjee

Table of Contents

Executive Summary

As we all are well versed with the saying “Health is wealth”, a healthy nation is a wealthy nation. The quality of healthcare system is considered as one of the crucial factors that define a country’s state and is also highly recognized as a great contributor to the country’s economy. Statistical modelling helps us to determine where our research needs to be focused and realize the factors that impact the healthcare quality and expenses. Also, there are many challenges to be considered while maintaining the health quality of a country. One such challenges is the health expenditure. The main objective of our research and analysis is to understand how to minimize the health expenses while improving or maintaining the same health quality.
Collection of data and data source play a paramount role in the analysis. Since analysis relies on data, the data considered must be legitimate, else the entire analysis would go in vain. To prevent such an event in this report, we have built our own dataset by taking the data from reliable resources such as OECD and WHO after analyzing the literature.

After collecting the data and analyzing it, we aim to understand the factors impacting the life expectancy and health expenditure. We also aim to provide recommendations to reduce the health expenses and improving the health quality of the country. For this, we performed an exploratory analysis, built hypothesis and regression models to analyze the data. Two independent models have been built to understand and analyze the factors that are affecting healthcare quality and expenses. The statistical data modelling involved building up of several models and to find out the unbiased model that generalizes and provide feasible recommendations that could reduce the health expenses and improve the health quality.

The key insights from our models is that the life expectancy of a country can be improved by increasing the health resources such as hospitals, medical graduates. The country’s healthcare expenses can have the maximum benefit when we bring more % of population under public insurance.

Problem Definition & Significance

We would like to present our findings and suggestions to the policy makers and top health officials of the OECD countries. Although, all countries under consideration are developed countries, these countries failed to formulate good strategic health policies, these resulted in high expenditure over the years with minimal improvement in Quality.

Our model provides recommendations to these countries of the factors they need to consider while formulating policies that improve Health care quality and keeping the expenses low. This is an interesting problem that might explain the peculiar case of USA which spends around 10000$ per person as compared to other OECD countries that spend an average of $3000, while Health quality measures such as life expectancy is not significantly different.

Prior Literature

Health Quality is important measure of accessing nation health. Health Quality can be measured on different scales, most important metrics being mortality- Infant, Child or Adult, Life Expectancy. Life Expectancy encompass mortality along the life cycle as compared to Adult or Child which accounts for a section of life cycle.

Background studies indicate that Life Expectancy has improved a lot in the 20th centuries. Studies show different factors like advancements in Technology, Increase in Income levels and nations expenditure on Healthcare are major determinants of upsurge. Increased demand and need for Healthcare reachability led to spike in healthcare employment and improved health quality. Reforms in education furthered sanitation and personal health. However, deaths by different causes like diseases decreased the mean number of years in a person life. In this model we aggregated the above factors in available format to access Life Expectancy.

Besides quality, the health expenditure also plays a key role in formulating health care policies and proper allocation of funds.

Availability of health resources is one of key indicators influencing the health expenses of a country. Health Economics suggest that medical practices and equipment are the major determinants in the growth of health expenditure. Population age structure is often considered as a distinguisher of the health characteristics of a country. From the analysis of WHO health financing, the high-income countries have a positive effect of 0.477 on the expenses. The use of primary care gatekeepers seemed to result in lower health expenditure. Public sector provision of health services was associated with lower health expenditure

Data Source/Preparation

The data we have considered is taken from OECD[5] and WHO[6] websites. Data has been collected from 2010 to 2015 for 39 different countries for 20 different variables. We have done some literature survey to come up with variables that explains variation in both health expenditure and health expenses. We took life expectancy of a country as a measure of health quality and health expenditure in USD as our dependent variables for analyzing the health quality and health expenditure, respectively.

Our dataset includes variables such as number of hospitals, hospital employment, medical graduates, nurse graduates, total equipment, social factors such as mean schooling years, the fatalities which included the deaths caused by respiratory, circulatory diseases, cancer and accidents and also economic factor such as expenditure per capita of the country to measure the Life expectancy i.e., health quality model. To analyze the health expenses model, Insurance type, social factors such as population structure (above 65 years), medical procedures and the health

resources available such as total equipment, hospitals, hospital employment are considered as independent variables to analyze the health expenses.

Once we collected the data, we have imputed null values with the countries which have equal expenditure and by searching for the relevant country websites. Our main assumption here is that most countries with equal expenses could have similar values for the other factors. All the data has been scaled with respect to the population of the respective country. Below are the variables and their scales.

Variable	Units
Diagnostic_Exams	per 1000 population
ExpenditurePerCapita	$ / capita
Hospitals	per million population
Life_Expectancy_at_Birth	years
Percpopulationulationabo	percent above 65
Physicians	per 1000 population
Private_Insurance	% of total population
Public_Insurance	% of total population
death_by_cancer	per 1000 population
death_by_circular	per 1000 population
death_by_accident	per 1000 population
Mean_Schooling_Years	years
NationalIncome	$ / capita
hospital_employment	per 1000 population
tot_equipment	per 1000 population
medical_grads	per 100,000
nurse_grads	per 100,000

The data has been cleaned using Microsoft Excel and R, while the data from different sources is combined using different libraries in R studio. Visualizations have been made with the help of R and Power BI.

Variable Selection and Hypothesis

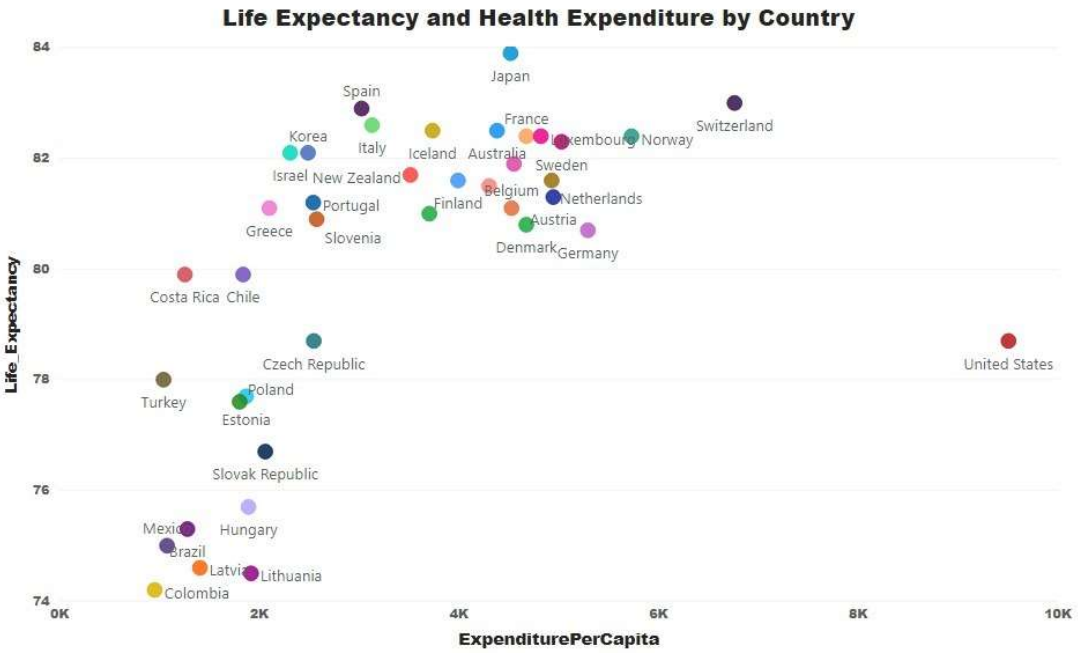
In this section we discuss the factors that influence the target variables - “Life Expectancy” – is a statistical measure of the average time an organism is expected to live, “Health Expense”- costs incurred for the medical expense of the country per capita, and how these factors impact the target variables.

We are interested in focusing on the actionable variables while forming the hypothesis. The actionable variables are the variables on which government have a control and can make strategic policies to improve the life expectancy or to reduce the health expenses.

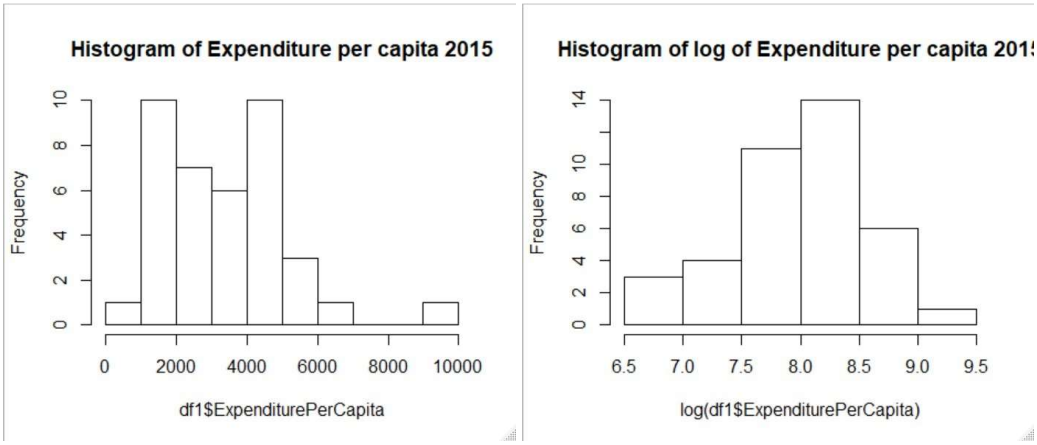
Based on our domain knowledge and literature survey, we figured out that the following variables can be considered for the prediction of our target variables i.e. “Life Expectancy” & “Healthcare Expense”

Variables		Hypothetical Description
1. Health Resources	a. Hospitals	With a greater number of hospitals in a country, more people can be treated with low waiting time, hence the life expectancy might be better for a country. With a greater number of hospitals in a country, the health expense can be increased initially, but might decrease in longer run.
	b. Hospital Employment	The greater number of employees in hospitals, the better and instant care to the patients, hence it might increase the life expectancy. The greater number of employees, their payment will increase and hence the health expense might increase.
	c. Total Equipment	With more equipment, the people can be diagnosed or treated with less waiting time and hence might increase the life expectancy. The greater number of equipment, might increase the health expense, as need to pay the initial cost and the maintenance cost for those.
	d. Medical Graduates	The increase in medical graduates, will increase the upcoming medical force of a country, so better and instant assistance to the people and hence it might increase the life expectancy.
	e. Nurse Graduates	The increase in nurse graduates, will increase the upcoming medical care force of a country, so people can get better and instant care and hence it might increase the life expectancy.
2. Health Quality	a. Death by respiratory disease	Now a days, most of the people are suffering from respiratory diseases, because of increase in smoking practices & pollution, hence the more death count by these diseases can result in decrease in life expectancy of a country.
	b. Death by circulatory disease	Many people are suffering from circulatory diseases because of unhealthy food habits & lack of agile practices, hence the more death count by circulatory diseases, can result in decrease in life expectancy of a country.
	c. Death by Cancer	As the death count increases by cancer, because of unhealthy food practices and decrease in immune systems, hence the life expectancy might decrease.
	d. Death by accidents	As the usage of vehicles are increasing, the deaths by accidents is also increasing in each country and hence it can result in decrease of life expectancy of a country.
3. Social Factors	a. Mean Schooling years	With the increase in literate people, the hygienic practices and primary care increases and hence the life expectancy might increase.
	b. Population Structure (Age>65)	With the higher number of aging populations, the health resources should increase, as they need more assistance and frequent care and hence it might increase the health care expenditure.
4. Economic Factors	a. Expenditure per capita	With increase of health care expenses, life expectancy might increase, as the health resources per capita increases, but excessive spending can result in wastage of money and might not help in improving life expectancy.
5. Medical Procedures	a. Diagnostic Exams	With the increase in number of diagnostic exams, the payment for these exams increases and hence the health expenses might also increase.
6. Insurance	a. Public Insurance	The premium cost of public insurance is comparatively low and hence increase in public insurances, might decrease healthcare expenses.
	b.Private Insurance	The premium cost of private insurance is very high and hence increase in number of private insurance holders, might increase healthcare expenses.

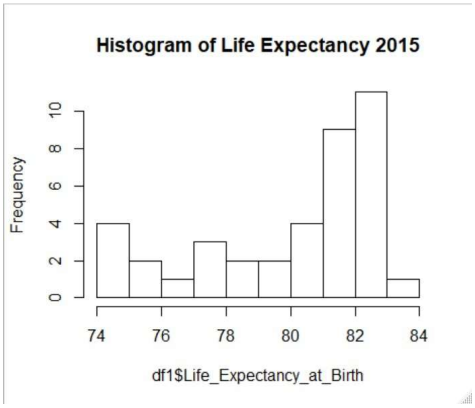
Exploratory Data Analysis & Visualizations



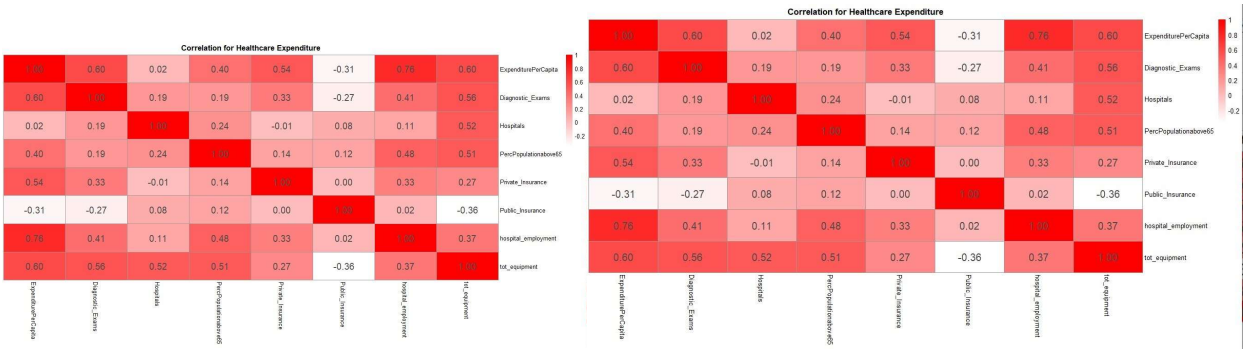
- The above graph shows the relationship between dependent variables (Life Expectancy and Expenditure per capita) for 2015. We can see United states has high expenses(~10K$) compared to other countries, while Life Expectancy ~80 years.



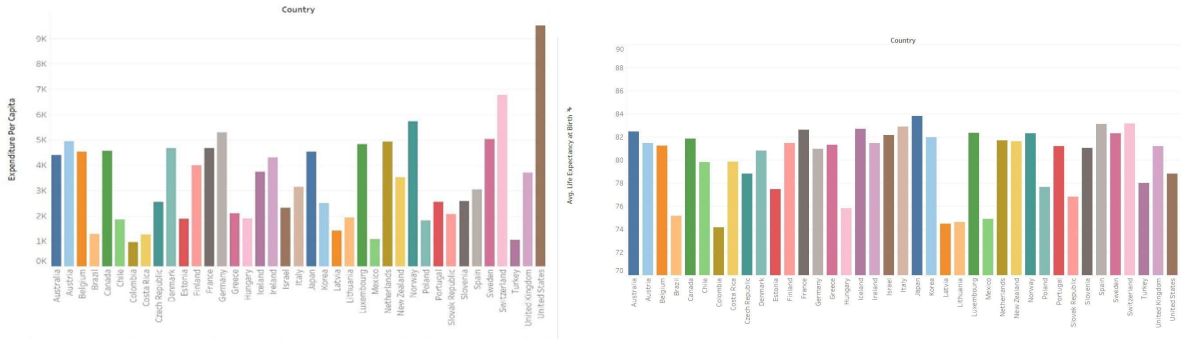
We can see from the above graph that our dependent variable “Expenditure per capita” is skewed toward right, we applied Log transformation on this variable, which not only helped removing skewness but also increased the interpretability of the variable in the model.



From the above plot, we can see the data is left skewed. Applying transformations will not improve skewness.



Correlation between dependent variables for Model 1 and model2. We did not have significant correlation(0.8) between the Variables under consideration, ruling out the possibility of multicollinearity between dependent variables.



Above diagram a) show the expenditure for different countries in USD per year for the year 2015 and b) show the Life expectancy of different countries for year 2015.

Models

Once the data has been cleaned and merged with the different columns, we have built the models in R.

Life Expectancy:

As we have data from 2010-2015(time variant) and for different countries (Cross sectional), panel statistical models would fit appropriate. So, we have built panel models, taking life expectancy as dependent variable and 11 predictors variables, model 1 is the basic OLS model/pooling, model 2 is fixed effects panel model and model3 is random effects panel model.

Model 1 :- Pooling Model:

```
panel_he <- plm((ExpenditurePerCapita) ~ Diagnostic_Exams + Hospitals +
        PercPopulationabove65 + Private Insurance + Public Insurance+ hospital employment +
            tot_equipment  ,data=d, model = "pooling")
```

Model 2 :- Fixed Model:

```
panel_hef <- plm((ExpenditurePerCapita) ~ Diagnostic_Exams + Hospitals +
        PercPopulationabove65 + Private Insurance + Public Insurance+ hospital employment +
            tot_equipment  ,data=d, model = "within")
```

Model 3:- Random Model:

```
panel_her <- plm((ExpenditurePerCapita) ~ Diagnostic_Exams + Hospitals +
        PercPopulationabove65 + Private Insurance + Public Insurance+ hospital employment +
            tot_equipment  ,data=d, model = "random")
```

Summary of Models:

From the below summary of the models, we can see the fixed effect model and random effects model's coefficients are stable and R2 value is comparatively good compared to other two models and fixed effects models explains the precise nature of relationship to our initial hypothesis and also Hausman test suggests Fixed model is good.

```
stargazer(pool_LEr,pool_LEf,pool_LEr,type="text")
```

```
==========================================================================
                                  Dependent variable:
                     ---------------------------------------------
                                 Life_Expectancy_at_Birth
                           (1)                 (2)                 (3)
--------------------------------------------------------------------------
Hospitals                 0.002               0.067**             0.002
                         (0.022)             (0.029)             (0.022)

ExpenditurePerCapita      0.0002              0.0001              0.0002
                         (0.0001)            (0.0001)            (0.0001)

Mean_Schooling_Years     0.267***            0.265**             0.267***
                         (0.094)             (0.107)             (0.094)
```

```
hospital_employment    0.162***          0.288***          0.162***
                       (0.049)           (0.069)           (0.049)

tot_equipment           0.012            0.100***           0.012
                       (0.008)           (0.022)           (0.008)

medical_grads          0.064***           0.042*           0.064***
                       (0.024)           (0.024)           (0.024)

nurse_grads             0.008*           0.009**            0.008*
                       (0.005)           (0.004)           (0.005)

death_by_cancer        0.606***          0.663***          0.606***
                       (0.160)           (0.223)           (0.160)

death_by_circular     -0.477***          -0.200*          -0.477***
                       (0.082)           (0.109)           (0.082)

death_by_accident     -1.869***          -1.266*          -1.869***
                       (0.717)           (0.709)           (0.717)

death_by_respirat       0.256            -0.295            0.256
                       (0.315)           (0.348)           (0.315)

Constant               71.997***                          71.997***
                       (1.197)                            (1.197)

-----------------------------------------------------------------------
Observations             234               234               234
R2                      0.509             0.541             0.509
Adjusted R2             0.485             0.419             0.485
F Statistic          230.248*** 19.727*** (df = 11; 184) 230.248***
=======================================================================
Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

From the fixed effect model, we can infer the impact of these variables on the life expectancy as below

	Variables	Impact
2	Hospitals	1 hospital per ten thousand persons,life Expectancy will increase by 6.7 years.
3	Expenditure per capita	increase in 10000$ expenditure per capita, Life Expectancy will increase by 1 years.
4	Mean_Schooling_Years	increase in 1 years of mean schooling per person, life expectancy will increase by 0.265 years
5	hospital_employment	incorporating 10 more employees in health sector per 1000 people. We can increase the life expectancy by 2.8
6	tot_equipment	increase of 100 equipment's per 1000 population, we can increase the life expectancy by 10
7	medical_grads	1 medical graduate per 1000 population, can increase the life expectancy by 4.2 years
8	nurse_grads	1 nurse graduate per 100 population, can increase the life expectancy by 9 years
9	death_by_cancer	increase of 10 deaths by cancer per 1000 population, life expectancy will decrease by 6.6 years
10	death_by_circular	increase of 10 deaths by circular per 1000 population, life expectancy will decrease by 2 years
11	death_by_accident	increase of 1 death by accidents per 1000 population, life expectancy will decrease by 1.2 years
12	death_by_respirate	increase of 1 death by respiratory per 1000 population, life expectancy will decrease by 0.2 years

From the above interpretation of the fixed effect model, it is evident that increase in number of hospitals, medical graduates, nurse graduates have the highest impact on life expectancy, while the deaths by any diseases has a severe negative effect on the life expectancy.

Also, here are the fixed effect coefficients across different countries for Fixed level model on Life expectancy model.

```
> fixef(pool_LEf)
      Australia         Austria         Belgium          Brazil
         54.955          60.575          59.111          65.917
         Canada           Chile        Colombia      Costa Rica
         61.710          69.052          64.498          69.824
 Czech Republic         Denmark         Estonia         Finland
         61.781          55.263          61.648          58.483
         France         Germany          Greece         Hungary
         62.181          57.472          57.212          60.484
        Iceland         Ireland          Israel           Italy
         59.454          64.119          69.116          60.538
          Japan           Korea          Latvia       Lithuania
         47.907          55.938          56.349          58.031
     Luxembourg          Mexico     Netherlands     New Zealand
         64.924          66.486          61.672          63.526
         Norway          Poland        Portugal Slovak Republic
         60.622          63.521          64.227          60.928
       Slovenia           Spain          Sweden     Switzerland
         64.477          66.345          59.999          57.907
         Turkey  United Kingdom   United States
         65.407          63.374          45.803
```

Health Expense Model:

We have taken an exponential model for the health expense, as it is evident from the figure 3 of visualization section, that the log(Health_Expenditure_perCapita) has a normal distribution. Like Life expectancy model, we did panel model to Health expense as the data is available from 2010-2015(Time variant) and for different countries(Cross-sectional).

Model 1:- Pooling Model

```
pool_HE<-plm(log(ExpenditurePerCapita)~Diagnostic_Exams+Hospitals+
        PercPopulationabove65 + Private_Insurance + Public_Insurance    +
        hospital_employment+tot_equipment , data=d, model = "pooling")
```

Model 2:- Fixed Model:

```
pool_HEf<-plm(log(ExpenditurePerCapita)~Diagnostic_Exams+Hospitals+
        PercPopulationabove65+      Private_Insurance+    Public_Insurance      +
        hospital_employment+tot_equipment,data=d,model = "within")
```

Model 3:- Random Model:

```
pool_Her<-plm(log(ExpenditurePerCapita)~Diagnostic_Exams+Hospitals+
        PercPopulationabove65+      Private_Insurance+    Public_Insurance      +
        hospital_employment+tot_equipment,data=d,model = "random")
```

Models Summary:

```
> stargazer(pool_HE,pool_HEf,pool_HEr,type="text")

================================================================================
                                        Dependent variable:
                    ---------------------------------------------------------
                                  log(ExpenditurePerCapita)
                            (1)                 (2)                 (3)
--------------------------------------------------------------------------------
Diagnostic_Exams          0.0004              0.001***            0.001***
                         (0.0003)            (0.0003)            (0.0002)

Hospitals                -0.014***            -0.006              -0.009***
                          (0.002)             (0.005)             (0.003)

PercPopulationabove65     0.015***            0.047***            0.045***
                          (0.005)             (0.010)             (0.007)
```

Private_Insurance	0.005***	0.001	0.004***
	(0.001)	(0.002)	(0.001)
Public_Insurance	0.002	-0.010**	-0.005*
	(0.002)	(0.004)	(0.003)
hospital_employment	0.058***	0.044***	0.046***
	(0.004)	(0.010)	(0.007)
tot_equipment	0.005***	0.009***	0.003**
	(0.001)	(0.003)	(0.001)
Constant	6.335***		6.716***
	(0.176)		(0.310)
Observations	234	234	234
R2	0.820	0.603	0.641
Adjusted R2	0.814	0.509	0.630
F Statistic	146.752*** (df = 7; 226)	40.876*** (df = 7; 188)	402.883***
Note:			*p<0.1; **p<0.05; ***p<0.01

From the above summary of the models, we can see the fixed effect model and random effects model's coefficients are stable and fixed effects models explains the precise nature of relationship to our initial hypothesis and also Hausman test suggests Fixed model is good.

From the fixed effect model, we can infer the impact of these variables on the life expectancy as below

	VARIABLES	IMPACT ON HEALTH EXPENSES
2	Private_Insurance	with increase in 10% percentage of population under private insurance coverage. The expenses increases by 1%.
3	Public_Insurance	with increase in 1% percentage of population under public insurance coverage. The expenditure decreases by 1%.
4	Hospitals	increasing hospitals by 1 per 1000 person, The healthcare expenses decrease by 6%.
5	Diagnostic_Exams	with increase in 1 diagnostic exams per 1000 population, the healthcare expenses increases by 1%.
6	PercPopulationabove65	with increase in 1% percentage of population above 65 years of age. The cost increases by 4.7%.
7	hospital_employment	with increase in 1 employee in health sector per 1000 population. The healthcare expenses increases by 4.4%.
8	tot_equipment	with increase in 10 equipment's per 1000 population. The healthcare expenses increases by 0.9%.

From the above interpretation of the fixed effect model, it is evident that increase in percentage of people receiving Public insurance and hospitals have the highest impact on decreasing the Health expenditure while encouraging the percentage of people in private insurance increases the Health expenditure.

Also, here are the fixed effect coefficients across different countries for Fixed level model.

```
> fixef(pool_HEf)
      Australia         Austria         Belgium          Brazil
         6.7648          7.0183          6.5954          6.9271
         Canada           Chile        Colombia      Costa Rica
         7.1103          7.0566          6.6079          7.0270
 Czech Republic         Denmark         Estonia         Finland
         6.7640          6.6771          6.3779          6.6838
         France         Germany          Greece         Hungary
         6.9934          6.6631          5.7537          6.7247
        Iceland         Ireland          Israel           Italy
         6.7414          7.5081          7.1463          6.4483
          Japan           Korea          Latvia       Lithuania
         5.6104          6.5405          5.8369          6.2874
    Luxembourg          Mexico     Netherlands     New Zealand
         7.4512          6.9477          7.1255          7.0122
          Norway          Poland        Portugal Slovak Republic
         7.0195          7.0404          6.7550          6.9316
       Slovenia           Spain          Sweden     Switzerland
         6.8948          6.8801          6.8133          6.7124
          Turkey  United Kingdom   United States
         6.6869          7.0089          5.7546
```

Quality Checks

Life Expectancy Model

```
> plmtest(pool_LE)

        Lagrange Multiplier Test - (Honda) for balanced panels

data:  Life_Expectancy_at_Birth ~ Hospitals + ExpenditurePerCapita +  ...
normal = 20.382, p-value < 2.2e-16
alternative hypothesis: significant effects
```

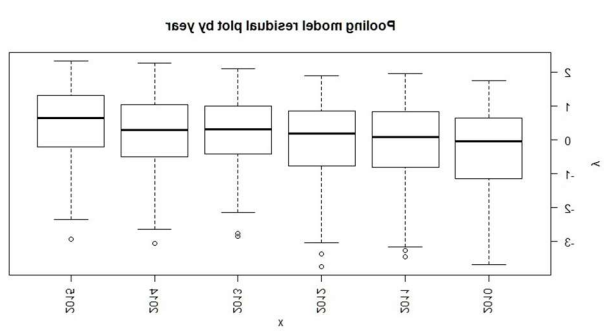
LMTEST is usually conducted for checking if the data is panel. Above test rejects the null hypothesis at p-value of 2.2e-16, suggesting the data is panel.

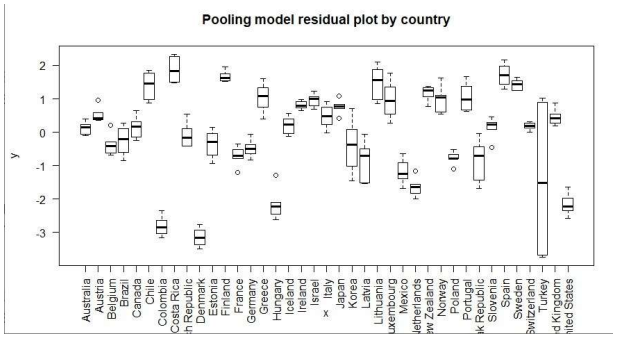
```
> phtest(pool_LEf,pool_LEr)

        Hausman Test

data:  Life_Expectancy_at_Birth ~ Hospitals + ExpenditurePerCapita +  ...
chisq = 27.091, df = 11,
p-value = 0.004452
alternative hypothesis: one model is inconsistent
```

We have conducted Hausman Test to compare both Fixed and Random effects models. Hausman test suggested that Fixed effect model is consistent than random effects models at p-value ~ 0.0044.

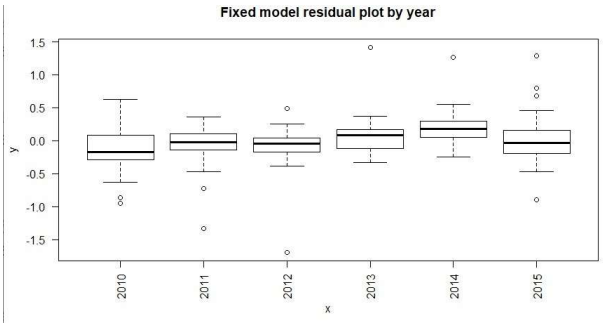
Residual Graphs

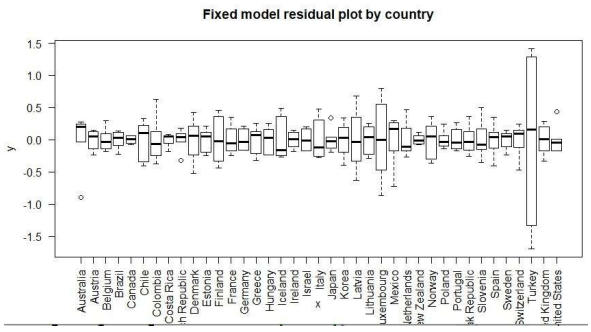
Pooling model



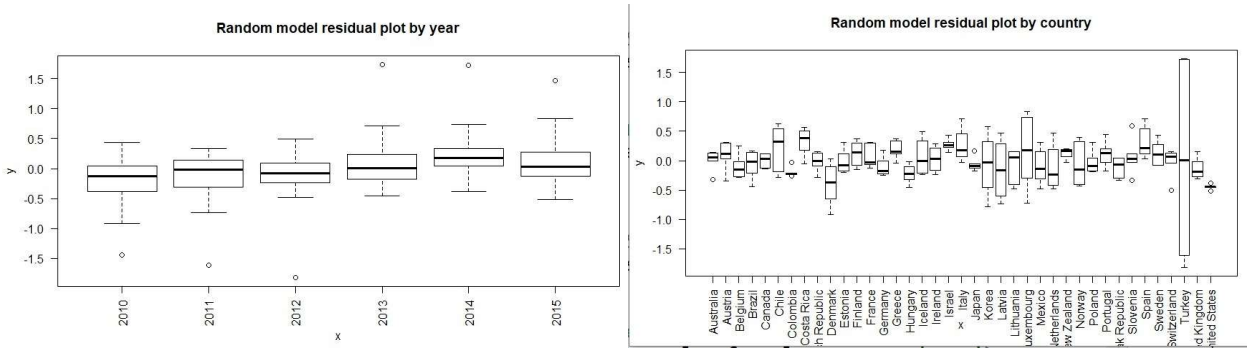


Fixed Effects Model





Random effects Model



The above residual graphs by time and by country are more stable and around the mean zero in case of Fixed effects model than random effects model and pooling model

Health Expenses model

```
> plmtest(pool_HE)

        Lagrange Multiplier Test - (Honda) for
        balanced panels

data:  log(ExpenditurePerCapita) ~ Diagnostic_Exams + Hospitals + PercPopulationabove65 +  ...
normal = 22.666, p-value < 2.2e-16
alternative hypothesis: significant effects
```

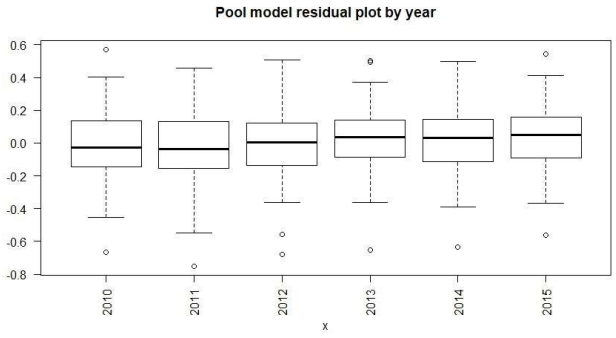
LMTEST is usually conducted for checking if the data is panel. Above test rejects the null hypothesis at p-value of 2.2e-16 that data is not panel, suggesting the data is panel.

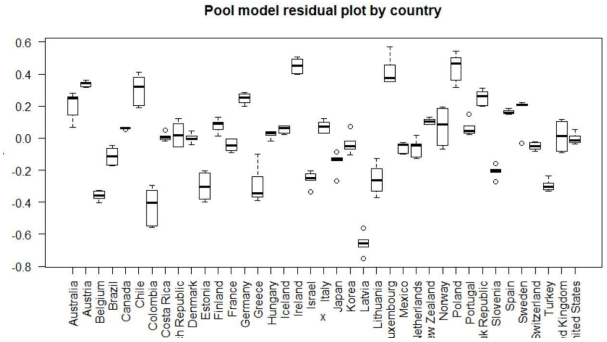
```
> phtest(pool_HEf,pool_HEr)

        Hausman Test

data:  log(ExpenditurePerCapita) ~ Diagnostic_Exams + Hospitals + PercPopulat
ionabove65 +  ...
chisq = 18.896, df = 7, p-value = 0.008521
alternative hypothesis: one model is inconsistent
```

We have conducted Hausman Test to compare both Fixed and Random effects models. Hausman test suggested that Fixed effect model is consistent than random effects models at p-value ~ 0.0085.

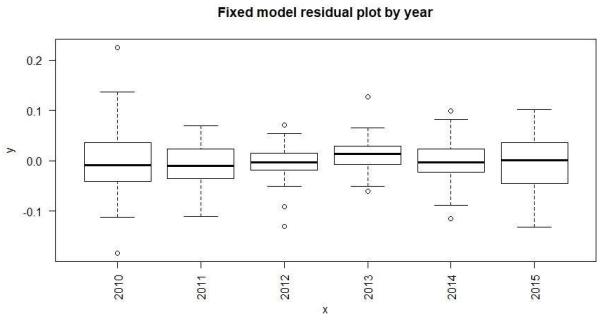
Residual Graphs

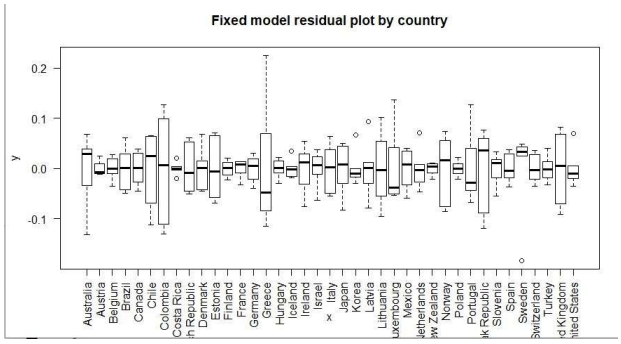
Pooling model



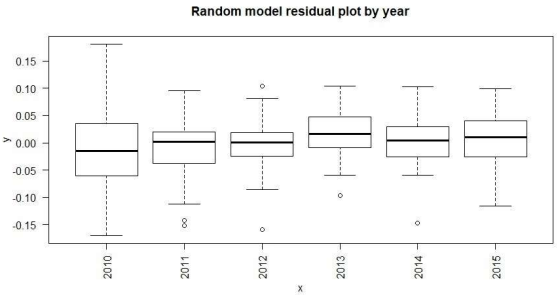


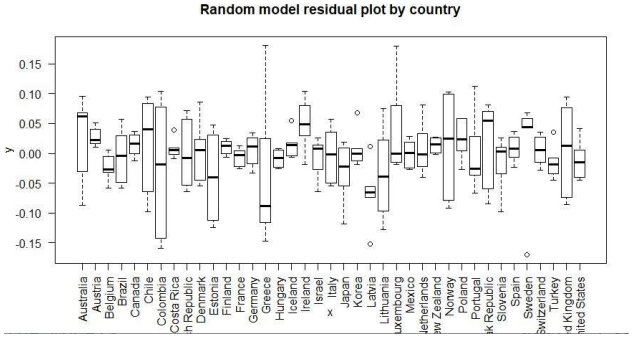
Fixed Effects Model





Random Effects Model





The above residual graphs by time and by country are more stable and around the mean zero in case of Fixed effects model than random effects model and pooling models.

- A quantitative measure for making this strategical decision is to build 10 hospitals per 1 million population, to increase Life expectancy by 6.7 years.
- To manage the health expenses effectively, the government should strive to increases the % of total population under public insurance.
- Quantitative measure for making this strategical decision would be to look at decrease in expenses by 1% per person with increasing 1% population into public insurance.
- Also, we observed that the peculiar case of USA with High health expenses is mostly due to less percentage of people under Public Insurance i.e., 30% if increased will decrease the health expenses and the recent presidential election 2020 campaign is all around "Medicare for All". Hope this model explains the argument of Campaigners.

References

1) Data Collection, OECD- https://stats.oecd.org/Index.aspx
2) Data Collection, WHO https://www.who.int/gho/database/en/
3) Panel Data Modelling - Panel Data Analysis Fixed & Random Effects - Oscar Torres-Reyna
4) Panel Data Modelling- Practical Guides To Panel Data Modeling: A Step by Step Analysis Using Stata , Hun Myoung Park, Ph.D
5) Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data- Andrew Bell, Kelvyn Jones-
6) Dormont, B., Grignon, M. & Huber, H., 2006. Health expenditure growth: reassessing the threat of ageing. Health Economics, 15(9), pp.947-963.
7) Dreger, C. & Reimers, H.E., 2005. Health care expenditures in OECD countries: a panel unit root and cointegration analysis.
8) Steingrímsdóttir, Ó.A., Næss, Ø., Moe, J.O. et al. Trends in life expectancy by education in Norway 1961–2009. Eur J Epidemiol 27, 163–171 (2012).
9) Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015
10) https://www.who.int/health_financing/documents/report_en_11_deter-he.pdf

Appendix

Link to GITHUB Repository (Code and Dataset)-

https://github.com/mangaln/Healthcare-Data-Analytics