Statistics OPRE 6359 Research Project:

Exploring the Factors which affects the Number of AIDS/HIV cases in different Countries in the World, Is Awareness an Important factor in high number of AIDS related deaths?

Saurabh Ranjan

Jindal School of Management

University of Texas at Dallas

Richardson, Texas, USA

sxr180117@utdallas.edu

Contents:

Abstract:

HIV/AIDS has been ranked 4th in Top causes of Death worldwide, it is large and growing causes of death and disease burden in the whole world, especially in sub-Saharan Africa. Some countries even have their 25% of population infected with HIV. We are just trying to understand what reason of such big differences could be, what all factors are important in having such high number of HIV cases and Finally, we are going to see if Awareness is big factor in affecting the high number

of AIDS related deaths. Dataset was prepared which has data for 108 different countries, columns like Number of HIV deaths, Awareness percentage, Country GDP, Life expectancy of countries and average number of schooling years people receive in that country were captured. Models were built using linear regression to predict the number of deaths due to AIDS taking expectancy, GDP, Education and Awareness as independent variables. Most of the countries in the Dataset are developing, hence, their Life expectancy, GDP and Education is going to be low comparatively but the interesting variable here is % Awareness. I would like to see how it behaves in different Linear models and whether it is significant factor or not in determining the Number of deaths per 100,000 people in different countries.

Introduction:

HIV stands for human immunodeficiency virus. It is the virus that can lead to acquired immunodeficiency syndrome or AIDS if not treated. Unlike some other viruses, the human body can't get rid of HIV completely, even with treatment. So, once we get HIV, we have it for life.

HIV attacks the body's immune system, specifically the CD4 cells (T cells), which help the immune system fight off infections. Untreated, HIV reduces the number of CD4 cells (T cells) in the body, making the person more likely to get other infections or infection-related cancers. Over time, HIV can destroy so many of these cells that the body can't fight off infections and disease. These opportunistic infections or cancers take advantage of a very weak immune system and signal that the person has AIDS, the last stage of HIV infection.

No effective cure currently exists, but with proper medical care, HIV can be controlled. The medicine used to treat HIV is called antiretroviral therapy or ART. If people with HIV take ART as prescribed, their viral load (amount of HIV in their blood) can become undetectable. If it stays undetectable, they can live long, healthy lives and have effectively no risk of transmitting HIV to an HIV-negative partner through sex. Before the introduction of ART in the mid-1990s, people with HIV could progress to AIDS in just a few years. Today, someone diagnosed with HIV and treated before the disease is far advanced can live nearly as long as someone who does not have HIV.

Scientists identified a type of chimpanzee in Central Africa as the source of HIV infection in humans. They believe that the chimpanzee version of the immunodeficiency virus (called simian immunodeficiency virus, or SIV) most likely was transmitted to humans and mutated into HIV when humans hunted these chimpanzees for meat and came into contact with their infected blood. Studies show that HIV may have jumped from apes to humans as far back as the late 1800s. Over decades, the virus slowly spread across Africa and later into other parts of the world. We know that the virus has existed in the United States since at least the mid to late 1970s.

When people get HIV and don't receive treatment, they will typically progress through three stages of disease. Medicine to treat HIV, known as antiretroviral therapy (ART), helps people at

all stages of the disease if taken as prescribed. Treatment can slow or prevent progression from one stage to the next. Also, people with HIV who take HIV medicine as prescribed and get and keep an undetectable viral load have effectively no risk of transmitting HIV to an HIV-negative partner through sex.

Stage 1: Acute HIV infection

Within 2 to 4 weeks after infection with HIV, people may experience a flu-like illness, which may last for a few weeks. This is the body's natural response to infection. When people have acute HIV infection, they have a large amount of virus in their blood and are very contagious. But people with acute infection are often unaware that they're infected because they may not feel sick right away or at all. To know whether someone has acute infection, either an antigen/antibody test or a nucleic acid (NAT) test is necessary. If you think you have been exposed to HIV through sex or drug use and you have flu-like symptoms, seek medical care and ask for a test to diagnose acute infection.

Stage 2: Clinical latency (HIV inactivity or dormancy)

This period is sometimes called asymptomatic HIV infection or chronic HIV infection. During this phase, HIV is still active but reproduces at very low levels. People may not have any symptoms or get sick during this time. For people who aren't taking medicine to treat HIV, this period can last a decade or longer, but some may progress through this phase faster. People who are taking medicine to treat HIV (ART) as prescribed may be in this stage for several decades. It's important to remember that people can still transmit HIV to others during this phase. However, people who take HIV medicine as prescribed and get and keep an undetectable viral load (or stay virally suppressed) have effectively no risk of transmitting HIV to their HIV-negative sexual partners. At the end of this phase, a person's viral load starts to go up and the CD4 cell count begins to go down. As this happens, the person may begin to have symptoms as the virus levels increase in the body, and the person moves into Stage 3.

Stage 3: Acquired immunodeficiency syndrome (AIDS)

AIDS is the most severe phase of HIV infection. People with AIDS have such badly damaged immune systems that they get an increasing number of severe illnesses, called opportunistic illnesses.

Without treatment, people with AIDS typically survive about 3 years. Common symptoms of AIDS include chills, fever, sweats, swollen lymph glands, weakness, and weight loss. People are diagnosed with AIDS when their CD4 cell count drops below 200 cells/mm or if they develop certain opportunistic illnesses. People with AIDS can have a high viral load and be very infectious.

How HIV spreads

To become infected with HIV, infected blood, semen or vaginal secretions must enter our body. This can happen in several ways:

- **By having sex.** You may become infected if you have vaginal, anal or oral sex with an infected partner whose blood, semen or vaginal secretions enter your body. The virus can enter your body through mouth sores or small tears that sometimes develop in the rectum or vagina during sexual activity.
- **From blood transfusions.** In some cases, the virus may be transmitted through blood transfusions. American hospitals and blood banks now screen the blood supply for HIV antibodies, so this risk is very small.
- **By sharing needles.** Sharing contaminated intravenous drug paraphernalia (needles and syringes) puts you at high risk of HIV and other infectious diseases, such as hepatitis.
- **During pregnancy or delivery or through breast-feeding.** Infected mothers can pass the virus on to their babies. HIV-positive mothers who get treatment for the infection during pregnancy can significantly lower the risk to their babies.

Problem Statement: Is Awareness (Correct knowledge of HIV) a significant factor in determining the number of AIDS related deaths?

Data-set and Variable Definitions:

Data had been collected from different sources such as UNICEF website, Kaggle etc. Data of Awareness percentage among youths about HIV for different country was collected for the year of 2015. Initially, there were 108 countries but due to missing records, the number of countries were reduced to 94.

TOP 20 Countries with highest number of HIV Death rate in HIV Data-set :

Country	Year	HIV Deaths (Total number per 100,000 individuals)	Awareness (correct knowledge of HIV)	Life Expectancy	GDP	Education
Lesotho	2015	555.68859	31.55	53.7	173.8289	10.7
Swaziland	2015	413.4952	0.04	58.9	3136.925	11.4
South Africa	2015	333.7744	37.725	62.9	5769.773	13
Mozambique	2015	297.31177	32.9425	57.6	528.3126	9.1
Botswana	2015	262.44002	0.08	65.7	6532.651	12.6
Namibia	2015	215.48398	50.516667	65.8	4737.67	11.7
Central African Republic	2015	188.1316	21.385	52.5	348.3814	7.1
Malawi	2015	183.78211	43.184091	58.3	362.6575	10.8
Zimbabwe	2015	173.9874	53.043333	67	118.6938	10.3
Zambia	2015	168.27112	38.58	61.8	1313.89	12.5

Kenya	2015	147.84414	53.226667	63.4	1349.971	11.1
Cameroon	2015	138.34643	30.173333	57.3	1244.429	10.4
Nigeria	2015	126.87363	26.696364	54.5	2655.158	10
Congo	2015	122.12494	24.244	64.7		11.1
Guinea-Bissau	2015	120.50259	18.756667	58.9	596.8717	9.2
Uganda	2015	95.643026	38.933333	62.3	693.8964	10
Togo	2015	93.971306	34.71	59.9	551.1383	12
Gabon	2015	77.541776	32.85	66	7388.984	12.6
Angola	2015	70.432555	32.9	52.4	3695.794	11.4
Gambia	2015	63.893592	30.5	61.1		8.9

Country: Country

Year: We are considering data for year 2015 only

HIV Deaths: Deaths due to HIV/AIDS - Sex: Both - Age: Age-standardized (Rate) (per 100,000 people)

Awareness: Per cent of young people with comprehensive, correct knowledge of HIV

Life Expectancy: Life Expectancy in age

GDP: Gross Domestic Product per capita (in USD)

Educ: Number of years of Schooling(years)

Hypothesis:

H0: Awareness is a significant factor in high number of HIV cases

H1: Awareness is not a significant factor in high number of HIV cases

Techniques used: Linear Regression Modelling and Anova in R

Descriptive Statistics:

Summary of Final HIV dataset looks like:

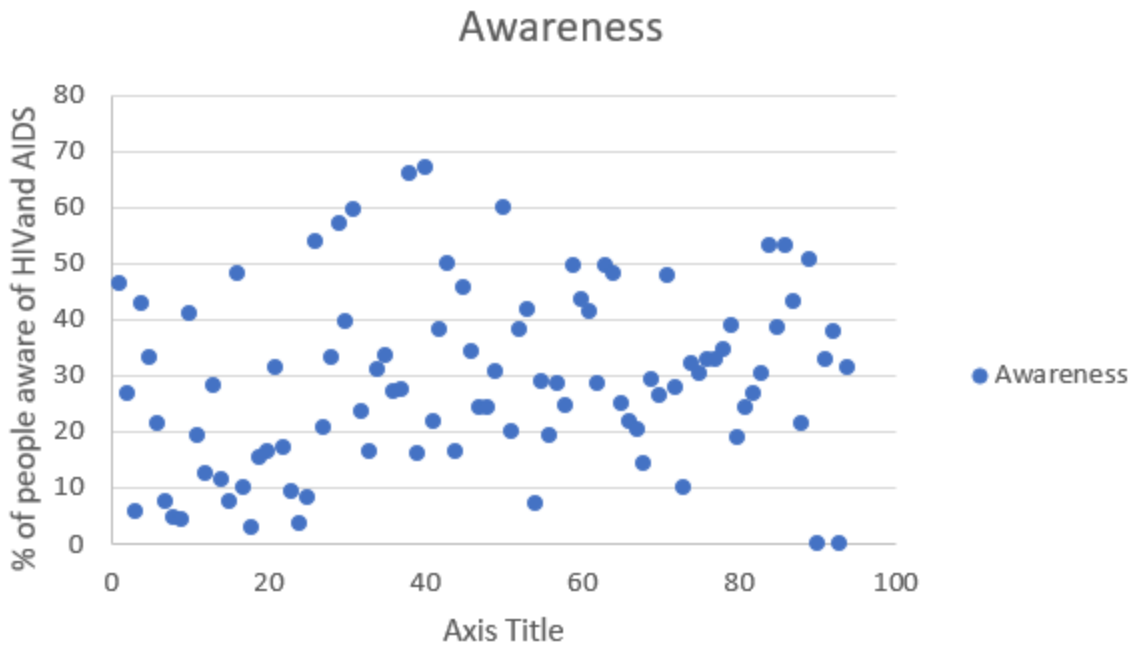
From R:

```
   HIV Deaths          Awareness      Life Expectancy       GDP              Education
Min.    :  0.0417   Min.    : 0.04   Min.    :51.00   Min.    :   33.68   Min.    : 4.90
1st Qu.:  2.3744   1st Qu.:18.88   1st Qu.:62.52   1st Qu.:  586.72   1st Qu.:10.12
Median : 12.4750   Median :28.67   Median :68.40   Median : 1355.54   Median :11.70
Mean   : 51.6293   Mean   :29.24   Mean   :67.77   Mean   : 3898.00   Mean   :11.58
3rd Qu.: 49.3523   3rd Qu.:39.51   3rd Qu.:74.25   3rd Qu.: 4765.75   3rd Qu.:13.18
Max.   :555.6886   Max.   :67.05   Max.   :79.60   Max.   :66346.52   Max.   :17.30
                                                   NA's   :6
```

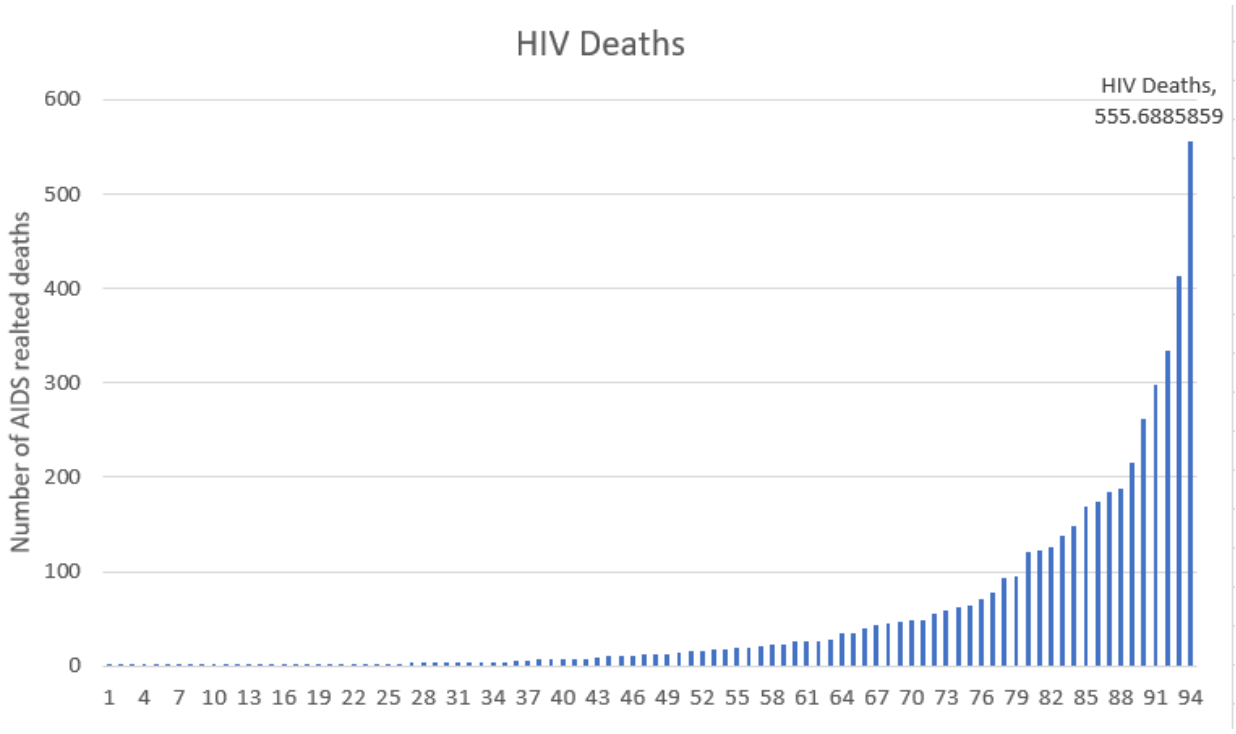
From Excel:

	HIV Deaths	*Awareness*	*Life Expectancy*	*GDP*	*Education*
Mean	51.62934839	29.24139576	67.76595745	3898.002981	11.58297872
Standard Error	9.715285996	1.619259064	0.751610112	825.4722717	0.254430601
Median	12.47503731	28.6707895	68.4	1355.542695	11.7
Mode	#N/A	#N/A	74.8	#N/A	12.7
Standard Deviation	94.19319246	15.6992991	7.287130403	7743.616305	2.466796203
Sample Variance	8872.357507	246.4679922	53.1022695	59963593.48	6.085083505
Kurtosis	10.88041078	-0.45151874	-0.766442319	49.30896831	0.197648988
Skewness	3.046247067	0.233865076	-0.34943891	6.327535034	-0.430829525
Range	555.6469253	67.01	28.6	66312.84145	12.4
Minimum	0.041660673	0.04	51	33.681223	4.9
Maximum	555.6885859	67.05	79.6	66346.52267	17.3
Sum	4853.158749	2748.691201	6370	343024.2624	1088.8
Count	94	94	94	88	94

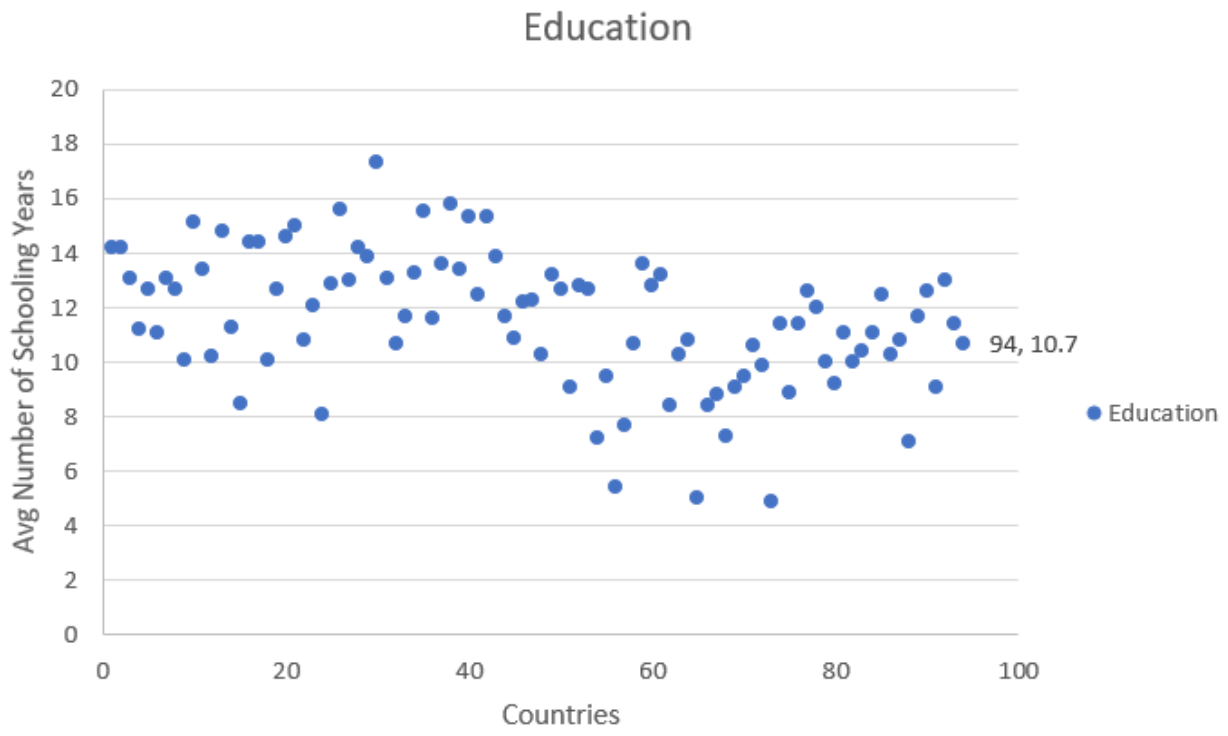
After removing the rows for which we didn't have enough data, there are 94 countries whose analysis we are going to perform here. My hypothesis focus more on HIV awareness variable hence I have not deleted the record for which we don't have GDP value.



The above graph shows the distribution of Awareness variable for all 94 countries.



The above graph shows the number of AIDS deaths for 94 countries in increasing order.



The above graph shows the relation between HIV Deaths in countries and avg schooling years .

Data Preparation and Model Building:

I needed to gather information about Countries, I gathered data from different sources. I downloaded awareness data from UNICEF website, Number of AIDS Deaths data from "ourworldindata.org" and Life Expectancy, GDP, Education data from Kaggle. After removing the unwanted and incomplete records, Data was cleaned and by using VLOOKUP function in excel, a final HIV dataset was prepared which I have used to build the model to perform data analysis. Steps involved in building the model were:

- Final Dataset was prepared after cleaning and consolidating from the original data source.
- Data was loaded in R and descriptive statistics was noted
- Train and test data were created through random sampling, 70 % of total data was collected in train sample and remaining 30 % were stored in test sample.
- Forward Step wise regression technique was used to build different model and model selection was done based on AIC value.
- Anova was conducted between two model to check if variable "Awareness" is a significant factor or not in determining the number of cases in different countries.

Creating Training and test (Validation) data through random sampling.

Code :

```
ind<-sample(0.70*nrow(dat))

tra<-dat[index,]

head(tra)

summary(tra)

tes<-dat[-index,]
```

Data Modelling:

```
model17<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy`)

model18<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy` + log(tra$Awareness))

anova(model17,model18)

model17<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy`+tra$GDP)

model18<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy`+tra$GDP + log(tra$Awareness))

anova(model17,model18)


model17<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy`+tra$GDP + tra$Education)

model18<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy` + tra$GDP + tra$Education +
log(tra$Awareness))

anova(model17,model18)


d<-lm(tra$`HIV Deaths` ~ tra$Education)

plot(tra$Education,  tra$`HIV Deaths`)

abline(d,lty=2)
```

Results, Graphs & Interpretation:

After building different models with and without Awareness variable, I tried to compare the models using ANOVA to find out whether the Awareness variable is significant or not.

I used Forward stepwise linear regression techniques to find the best fit of the models. Almost in every model, Log (Awareness) variable was significant. Logarithmic transformation of Awareness variable was used just to convert a highly skewed variable into a more normalized dataset.

In the Results below, we will see that whenever Log (Awareness) variable is introduced, it is significant and AIC value of model decreases and hence we were able to prove its significance by comparing the models through ANOVA.

When Life Expectancy is the only Independent Variable and then Log (Awareness) is introduced to compare the models.

Hypothesis for below Models:

H0: Models are same, no effect of Awareness variable

H1: Models are different, Awareness variable is highly significant

Below Explained Cases:

Case1: Comparison between model17 and model18

Case2: Comparison between model19 and model20

Case3: Comparison between model21 and model22

Case4: Effect of Log (Awareness) on HIV Deaths

FINAL Model: model22<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy` + tra$Education + log(tra$Awareness))

Case: 1

```
> model17<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy`)
> summary(model17)

Call:
lm(formula = tra$`HIV Deaths` ~ tra$`Life Expectancy`)

Residuals:
    Min      1Q  Median      3Q     Max
-132.91  -48.67  -12.37   13.34  393.44

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          558.520    106.240   5.257 1.85e-06 ***
tra$`Life Expectancy`  -7.379      1.609  -4.586 2.20e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 92.45 on 63 degrees of freedom
Multiple R-squared:  0.2503,    Adjusted R-squared:  0.2384
F-statistic: 21.03 on 1 and 63 DF,  p-value: 2.201e-05

> model18<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy` + log(tra$Awareness))
> summary(model18)

Call:
lm(formula = tra$`HIV Deaths` ~ tra$`Life Expectancy` + log(tra$Awareness))

Residuals:
    Min      1Q  Median      3Q     Max
-126.21  -44.88   -8.84   19.53  409.94

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          589.746    101.037   5.837 2.10e-07 ***
tra$`Life Expectancy`  -6.502      1.551  -4.192 8.94e-05 ***
log(tra$Awareness)    -27.472      9.450  -2.907  0.00505 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 87.42 on 62 degrees of freedom
Multiple R-squared:  0.3402,    Adjusted R-squared:  0.3189
F-statistic: 15.98 on 2 and 62 DF,  p-value: 2.521e-06

> anova(model17,model18)
Analysis of Variance Table

Model 1: tra$`HIV Deaths` ~ tra$`Life Expectancy`
Model 2: tra$`HIV Deaths` ~ tra$`Life Expectancy` + log(tra$Awareness)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     63 538411
2     62 473824  1     64587 8.4513 0.005054 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

model17 and model18 were compared using ANOVA, obtained p value was 0.0050 which signifies that Both the models are different and Log (Awareness) is a significant variable in the dataset.

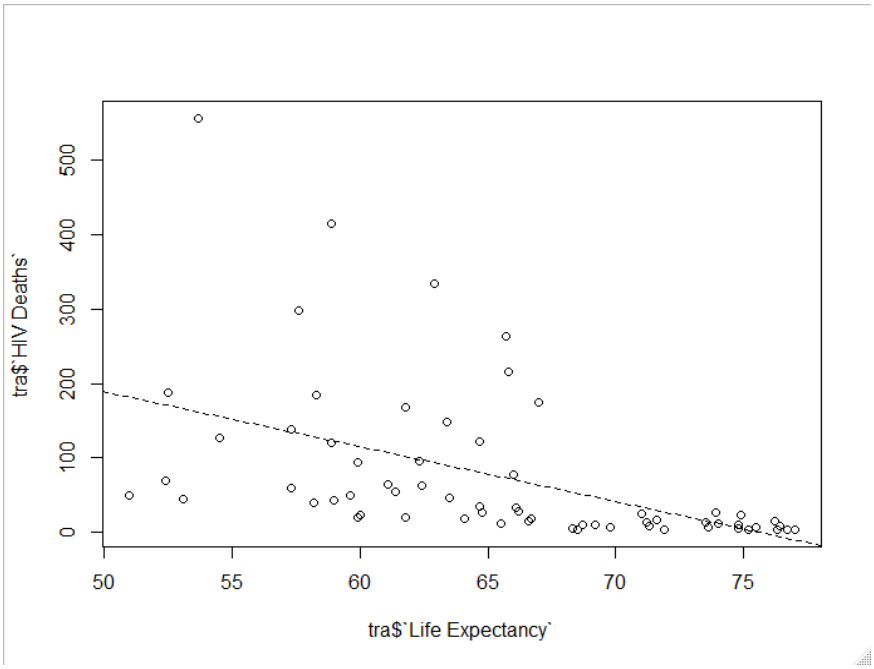


Fig:1 (Best Fit line for the tra$HIV Deaths and tra$Life Expectancy)

When GDP is also introduced, figure below:

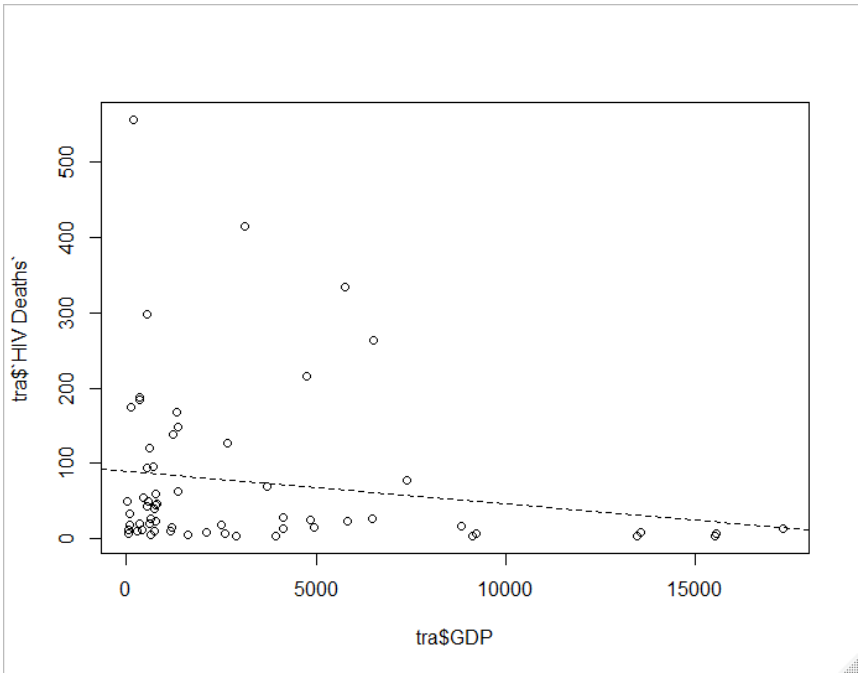


Fig:2 (Best Fit line for the tra$HIV Deaths and tra$GDP)

Best fit lines for models are included above.

Case 2:

```
> model19<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy`+tra$GDP)
> summary(model19)

Call:
lm(formula = tra$`HIV Deaths` ~ tra$`Life Expectancy` + tra$GDP)

Residuals:
    Min      1Q  Median      3Q     Max
-142.27  -44.75  -20.26   19.40  389.11

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          631.515041 126.947945   4.975 6.16e-06 ***
tra$`Life Expectancy`  -8.670176   2.016776  -4.299 6.65e-05 ***
tra$GDP                 0.003737   0.003399   1.099    0.276
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 94.85 on 58 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.2643,    Adjusted R-squared:  0.2389
F-statistic: 10.42 on 2 and 58 DF,  p-value: 0.0001363

> model20<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy`+tra$GDP + log(tra$Awareness))
> summary(model20)

Call:
lm(formula = tra$`HIV Deaths` ~ tra$`Life Expectancy` + tra$GDP +
    log(tra$Awareness))

Residuals:
    Min      1Q  Median      3Q     Max
-133.58  -40.56   -8.70   14.79  405.60

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          664.549933 120.507916   5.515 8.84e-07 ***
tra$`Life Expectancy`  -7.829269   1.928548  -4.060 0.000152 ***
tra$GDP                 0.003743   0.003211   1.165 0.248712
log(tra$Awareness)    -27.429574   9.711857  -2.824 0.006517 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 89.62 on 57 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.3546,    Adjusted R-squared:  0.3206
F-statistic: 10.44 on 3 and 57 DF,  p-value: 1.423e-05

> anova(model19,model20)
Analysis of Variance Table

Model 1: tra$`HIV Deaths` ~ tra$`Life Expectancy` + tra$GDP
Model 2: tra$`HIV Deaths` ~ tra$`Life Expectancy` + tra$GDP + log(tra$Awareness)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     58 521831
2     57 457768  1     64063 7.9769 0.006517 **
```

model19 and model20 were compared using ANOVA, obtained p value was 0.0065 which signifies that Both the models are different and Log (Awareness) is a significant variable in the dataset.

Case 3:

When Education is also considered, figure below:

```
> model21<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy`+tra$Education)
> model22<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy` + tra$Education + log(tra$Awareness))
> anova(model21,model22)
Analysis of Variance Table

Model 1: tra$`HIV Deaths` ~ tra$`Life Expectancy` + tra$Education
Model 2: tra$`HIV Deaths` ~ tra$`Life Expectancy` + tra$Education + log(tra$Awareness)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     62 466182
2     61 398253  1     67929 10.405 0.002022 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Now, Education is introduced, and GDP is removed. Then, Model21 and model22 were compared using ANOVA, obtained p value was 0.0020 which signifies that Both the models are different and Log (Awareness) is a significant variable in the dataset.

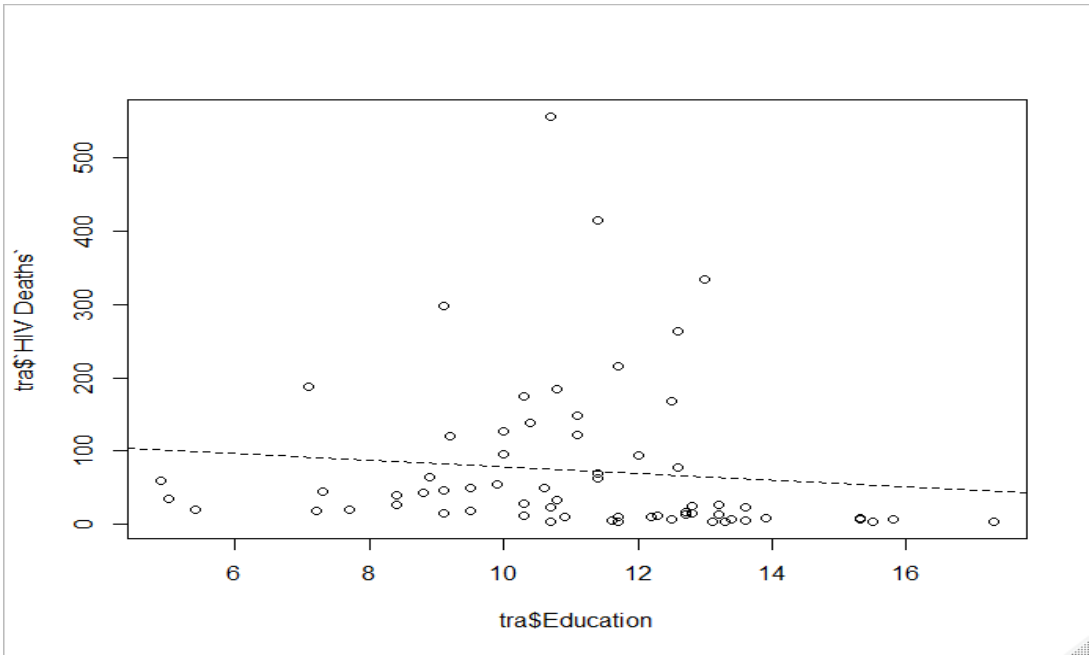


Fig: 3

Best Fit line for (tra$HIV Death ~ tra$Education)

```
> model21<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy`+tra$GDP+tra$Education)
> summary(model21)

Call:
lm(formula = tra$`HIV Deaths` ~ tra$`Life Expectancy` + tra$GDP +
    tra$Education)

Residuals:
    Min      1Q  Median      3Q     Max
-169.50  -46.10  -11.77   16.11  344.92

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          643.446688 120.248229   5.351 1.62e-06 ***
tra$`Life Expectancy` -12.143530   2.281741  -5.322 1.80e-06 ***
tra$GDP                -0.000436   0.003551  -0.123  0.90270
tra$Education          20.514282   7.380596   2.779  0.00736 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 89.79 on 57 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.3521,    Adjusted R-squared:  0.318
F-statistic: 10.33 on 3 and 57 DF,  p-value: 1.584e-05

> model22<-lm(tra$`HIV Deaths`~ tra$`Life Expectancy`+tra$GDP + tra$Education + log(tra$Awareness))
> summary(model22)

Call:
lm(formula = tra$`HIV Deaths` ~ tra$`Life Expectancy` + tra$GDP +
    tra$Education + log(tra$Awareness))

Residuals:
    Min      1Q  Median      3Q     Max
-155.98  -46.13   -6.42   19.54  360.36

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          678.115931 112.428955   6.032 1.35e-07 ***
tra$`Life Expectancy` -11.401060   2.136189  -5.337 1.77e-06 ***
tra$GDP                -0.000585   0.003304  -0.177  0.86010
tra$Education          21.274877   6.871522   3.096  0.00306 **
log(tra$Awareness)    -28.419293   9.059523  -3.137  0.00272 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 83.54 on 56 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.4489,    Adjusted R-squared:  0.4096
F-statistic: 11.41 on 4 and 56 DF,  p-value: 7.697e-07

> anova(model21,model22)
Analysis of Variance Table

Model 1: tra$`HIV Deaths` ~ tra$`Life Expectancy` + tra$GDP + tra$Education
Model 2: tra$`HIV Deaths` ~ tra$`Life Expectancy` + tra$GDP + tra$Education +
    log(tra$Awareness)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     57 459546
2     56 390863  1     68683 9.8405 0.002721 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, All the Independent variables are included. Then, Model21 and model22 were compared using ANOVA, obtained p value was 0.0027 which signifies that Both the models are different and Log (Awareness) is a significant variable in the dataset.

Case 4:

Considering Only "Awareness" variable

```
> model18<-lm(tra$`HIV Deaths`~ log(tra$Awareness))
> summary(model18)

Call:
lm(formula = tra$`HIV Deaths` ~ log(tra$Awareness))

Residuals:
    Min      1Q  Median      3Q     Max
-101.12  -52.48  -33.36   12.53  489.26

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          187.84      35.81   5.246 1.93e-06 ***
log(tra$Awareness)   -35.17      10.42  -3.376  0.00126 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 98.25 on 63 degrees of freedom
Multiple R-squared:  0.1532,    Adjusted R-squared:  0.1398
F-statistic:  11.4 on 1 and 63 DF,  p-value: 0.001262
```

And it is highly significant.

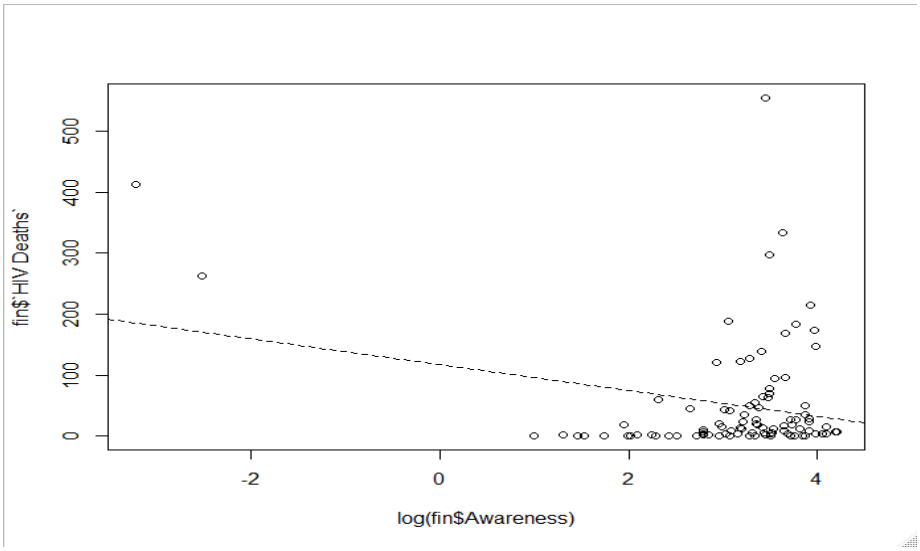


Fig: 4

Best Fit line for (tra$HIV Death ~ tra$Education)

The above best line suggests that HIV Deaths and % Awareness are inversely proportional, hence if Awareness increases, Number of HIV deaths for that year would decrease.

The graph shows the relationship perfectly, higher the awareness lower would be number of deaths.

In the above model, when we have (HIV Deaths ~ Log (Awareness)), Coefficient of log(Awareness) is -35 which is huge. Hence, 1 % increase in Awareness would result in 35 less number of AIDS related deaths every year.

In all the three ANOVA cases where eight different models were compared, Log (Awareness) was highly significant in all the cases. Hence, we can conclude that the null hypothesis is true, Awareness is an important significant factor in influencing the number of AIDS related death in different countries in the world.

Conclusion:

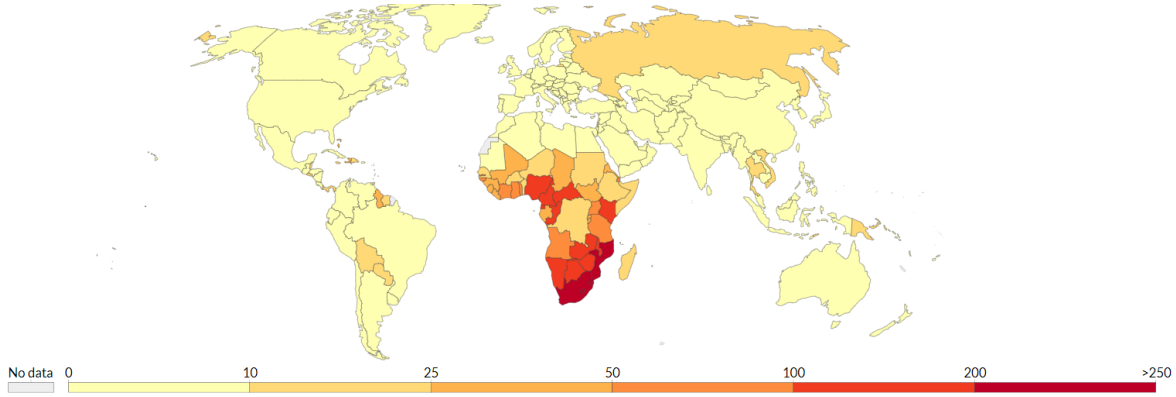
Awareness was one of the favorite independent variables whose values were collected through surveys and yet it came as the highly significant factor in AIDS related deaths. life Expectancy and GDP were also significant with negative sign that means they are inversely proportion.

43 countries out of 94 are African, they belong to African Continent. They are developing countries with less GDP. According to the model22, we can say that Life expectancy, HIV Awareness, GDP and Education of a country affects the total number of HIV deaths significantly. Hence, Government should must work towards creating a HIV Awareness program for their citizens, making them aware of the causes, cure and prevention of AIDS. This would surely help in convincing people to use proper precaution while having unsafe sex.

Finally, we can conclude that “Percentage of people having correct knowledge of HIV” has an important role in determining the number of AIDS deaths.

Appendix:

World Map: Death rates from HIV/AIDS, measured as the number of deaths per 100,000 individuals.



ANOVA: Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample.

Linear Regression: In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.

All the data analysis and graph plotting were performed in R and Excel.

References:

https://data.unicef.org/resources/dataset/hiv-aids-statistical-tables/

https://ourworldindata.org/grapher/hiv-death-rates?tab=chart&time=1990..2017

http://ghdx.healthdata.org/gbd-results-tool

https://www.kaggle.com/kumarajarshi/life-expectancy-who

https://www.cdc.gov/hiv/basics/whatishiv.html

https://en.wikipedia.org/wiki/Linear_regression

https://en.wikipedia.org/wiki/Analysis_of_variance