UNIVERSITY OF
LIMERICK
OLLSCOIL LUIMNIGH

Econometric Analysis on relation between AIDS cases over time and vitality rate

Name	Aswathi Rajashekaran Nair
Student ID	21110557
Module Name	Applied Econometrics for Business
Module Code	EC6062
Semester/Year	Spring semester/2021-22

CONTENTS

1. INTRODUCTION

The current project intends to investigate the age disparity between aids cases reported between 1981 and 2002, as well as the progression and autocorrelation of the disease over the same time period. Acquired Immune Deficiency Syndrome (AIDS) is one of the most feared chronic diseases that human beings have ever encountered. Because it is generally considered embarrassing and difficult to discuss, there has been a long time of cases not recorded until awareness efforts encouraged patients to seek medical attention. As a result, reports have become increasingly accurate and reflect the true situation. Despite the fact that we can derive some insights about the evolution without turning to data, it is critical to evaluate those ideas in the context of data. Furthermore, this project demonstrates how different age groups differ in terms of the number and trajectory of cases. Finally, the project also intend to analyse whether or not the reports of dead or alive cases are in agreement with the number of cases.

2. LITERATURE REVIEW

At least two statistical tests were performed in the literature review of the publications that were included in this study. To my knowledge, the inclusion criteria and the nature of the studies included in this review all necessitated the use of a specific type of statistical analysis to evaluate the research topic. The main findings are consistent with those previously reported, which reveal that the vast majority of articles employ more than one statistical test. On the other hand, an assessment of study designs and statistical procedures in journals indicated that a small number of studies reported the use of multiple statistical tests, which is at odds with the findings of this review.

In order to estimate the prevalence and incidence of any disease, the majority of the published publications rely solely on the use of comprehensive case data analysis or readily available case data analysis (Boerma JT et al, 2003).
A couple of the papers explain ad hoc approaches for estimating disease number of new cases, such as the use of dummy variables and mean imputation, in order to obtain more accurate estimates (Schafer JL, 2002). In addition, It is observed that AIDS-related research on younger generations appears to be increasing, and that the time lag between publication and indexing remains a concern. In addition, the findings and recommendations contain information regarding various variables, such as publication kind, publishing year, entry date, language, source, publication country, geographic region, and gender. there are only a few papers that discuss more complex ways of correcting for missing data, such as inverse probability weighting, instrumental variables, and multiple imputation (Eekhout I et al, 2012).

3. DATA AND METHODOLOGY

The data used is from AIDS Public Information Data Set (APIDS) US Surveillance Data for 1981-2002, CDC WONDER On-line Database. The dataset contains 5858 observations and 9 variables. There are 4 variables which are both in their original form and their coded version.

- Year: Either a string telling the year in characters or in numbers. 1981 only contains the annual report per age category, so we will consider December as the month of report for it.

- Month: Also in characters and in coded format ("Year/Month"). We first convert the mention of 1981 reports being the sum over "All Months" to it being December ("1981/12"). Next, we add the first day of every month as the day of report and convert the coded variable to date type to make it easy to plot the temporal evolution for the number of cases.

- Age: 12 Age categories, coded from 0 to 10, and A, B, C in this order. We replace the A, B, C by 11, 12 and 13 in the coded version of the variable.

- Vital status: it has two categories, either reported dead before 2001 or not.

- Cases: The number of reported cases for a certain year, month, age category and vital status.

```
aids_data <- read.csv("./Aids_dataset.csv")
unique(aids_data$Age.at.Diagnosis.Code)
[1] "3" "4" "5" "6" "7" "8" "9" "A" "B" "C" "2" "1" "0"
```

```
aids_data <- aids_data %>%
  mutate(Age.at.Diagnosis.Code = recode(Age.at.Diagnosis.Code,
                                        "A"="10", "B"="11", "C"="12"),
         Age.at.Diagnosis.Code = as.numeric(Age.at.Diagnosis.Code),
         Month.Reported.Code = gsub("All Months", "12", Month.Reported.Code),
         Month.Reported.Code = paste0(Month.Reported.Code, "/01"),
         Month.Reported.Code = as.Date(Month.Reported.Code))
```

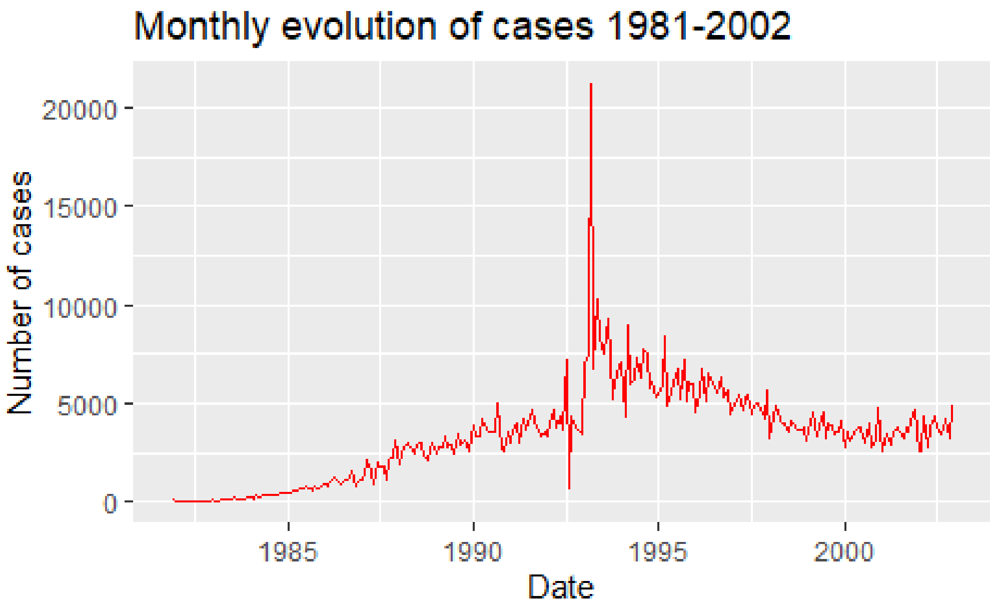
```
str(aids_data)
'data.frame':	5858 obs. of  9 variables:
 $ Year.Reported       : chr  "Before 1982" "Before 1982" "Before 1982" "Before 1982" ...
 $ Year.Reported.Code   : int  1981 1981 1981 1981 1981 1981 1981 1981 1981 1981 ...
 $ Month.Reported      : chr  "Before 1982" "Before 1982" "Before 1982" "Before 1982" ...
 $ Month.Reported.Code  : Date, format: "1981-12-01" ...
 $ Age.at.Diagnosis     : chr  "20 - 24 Years" "25 - 29 Years" "30 - 34 Years" "35 - 39 Years or age is missing" ...
 $ Age.at.Diagnosis.Code: num  3 4 5 6 7 7 8 9 9 10 ...
 $ Vital.Status         : chr  "Dead: Reported dead before 2001" "Dead: Reported dead before 2001" "Dead: Reported dead before 2001" "Dead: Reported dead before 2001" ...
 $ Vital.Status.Code    : int  1 1 1 1 0 1 1 0 1 1 ...
 $ Cases               : int  6 29 42 38 2 20 9 2 8 2 ...
```

i. TIME SERIES ANALYSIS: CASES

Plot of monthly evolution of cases:

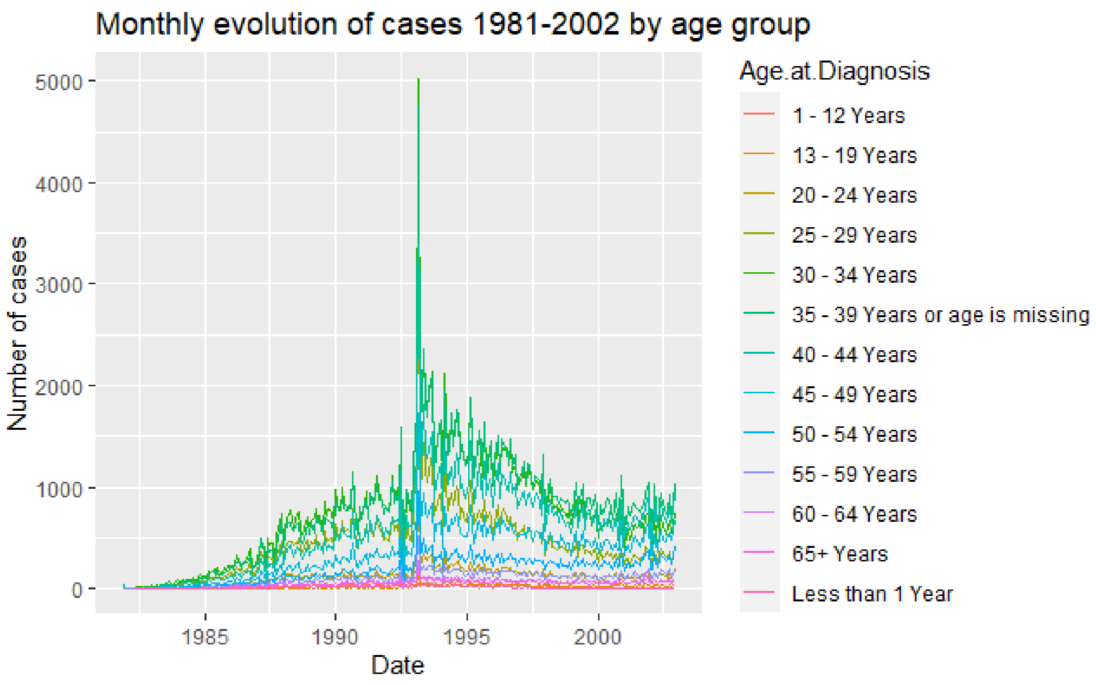
The monthly progression of the multilateral number of cases is presented first. Firstly, tallied the total number of reported cases across all categories for each month. Between 1981 and 1993, there was an upward trend in the number of cases. In 1993, there was a significant increase that did not reflect the number of cases in a single month, but rather a tally of unreported cases from previous months. After this maximum value, the number of cases decreases and stabilizes around 4000 after the year 2000.

```
monthly_total_evolution <- aids_data %>%
 group_by(Month.Reported.Code) %>%
 summarize(Cases = sum(Cases))

monthly_total_evolution %>%
 ggplot(aes(x = Month.Reported.Code, y = Cases)) +
 geom_line(color = "red") +
 ggtitle("Monthly evolution of cases 1981-2002") +
 xlab("Date") + ylab("Number of cases")
```



By categorizing the evolution by age group, it is observed that the groups that follow the general trend with the enormous peak in the 1990s are predominantly between the ages of 20 and 49. Very old and very young individuals have a stable evolution, with reported cases hovering around 0 and a small peak in the 1990s. 35 to 39 years old is the age group with the highest number of cases reported over time.

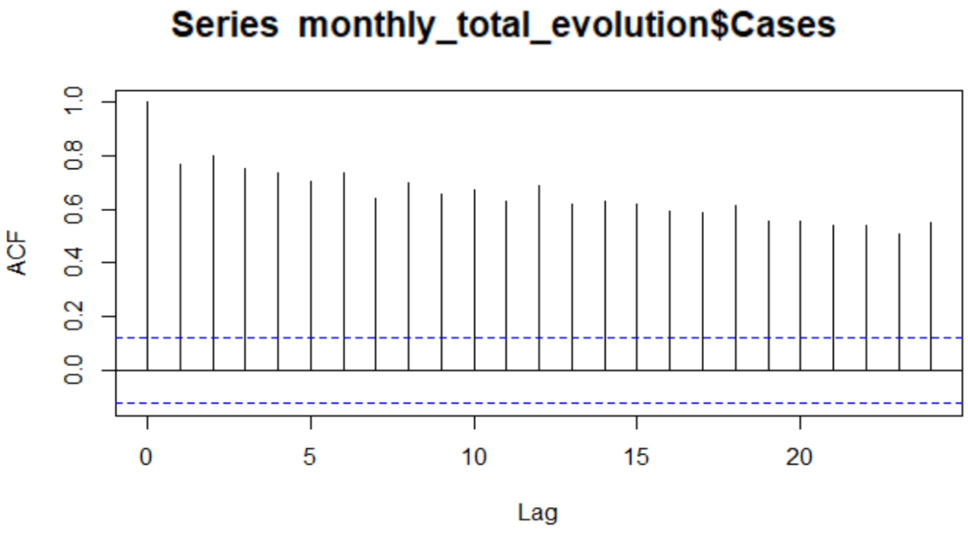
```
monthly_total_evolution_age <- aids_data %>%
 group_by(Month.Reported.Code, Age.at.Diagnosis) %>%
 summarize(Cases = sum(Cases))

monthly_total_evolution_age %>%
 ggplot(aes(x = Month.Reported.Code, y = Cases, color = Age.at.Diagnosis)) +
 geom_line() +
 ggtitle("Monthly evolution of cases 1981-2002 by age group") +
 xlab("Date") + ylab("Number of cases")
```



ii. AUTOCORRELATION

Plotting the autocorrelation plot for the number of cases over the years 1983 to 2002, combining all the categories. The series is evidently highly autocorrelated. There is no noticeable seasonality. However, the autocorrelation coefficient overall decreases for older lags. The highest autocorrelation is observed with the month preceding the 80 percent mark.

```
acf(monthly_total_evolution$Cases)
```



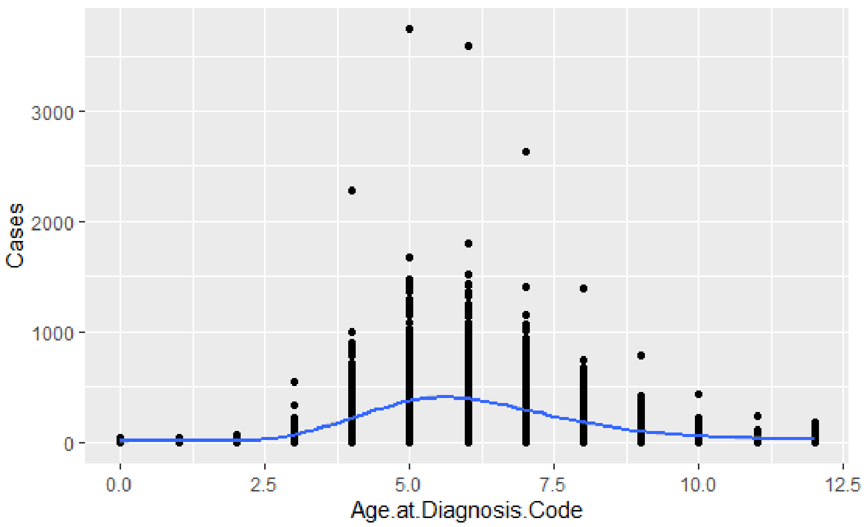
iii. AGE-CASES RELATIONSHIP

This section aims to examine the impact of age on the number of cases in greater depth. The age variable will be treated as a continuous variable, with the categories (0 to 12) serving as our vector, and the correlations with the number of cases and scatterplot will be examined. We will also consider age as a category and conduct statistical tests to determine whether there is a statistically significant difference between the various age groups.

#plot of cases and age

First, we show the scatterplot of the number of cases as a function of the age. We can clearly see that there is a bell curve shaped pattern, with the highest number of cases generally recorded for the middle aged people. Older and youger people tend to have less cases reported.

```
ggplot(aids_data, aes(x = Age.at.Diagnosis.Code, y = Cases)) +
 geom_point() +
 geom_smooth()
```



Considering the age as a category, we show the boxplot of the number of cases. We can clearly see that there is a difference between age groups but not between each pair. This means that there are bigger age groups among which we can see a significant difference. The bigger groups that we can define are as follow:

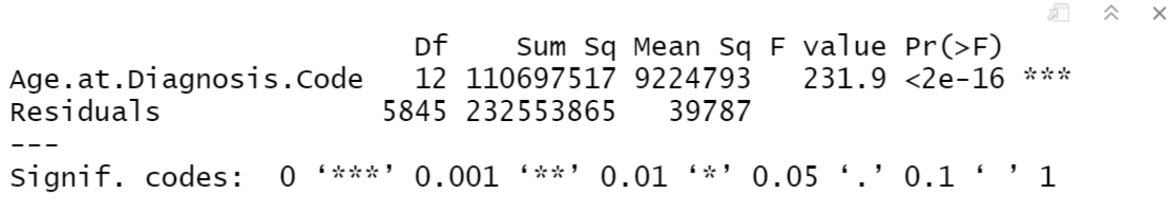
* 1 to 24.

* 25 to 49.

* 50 to +65.

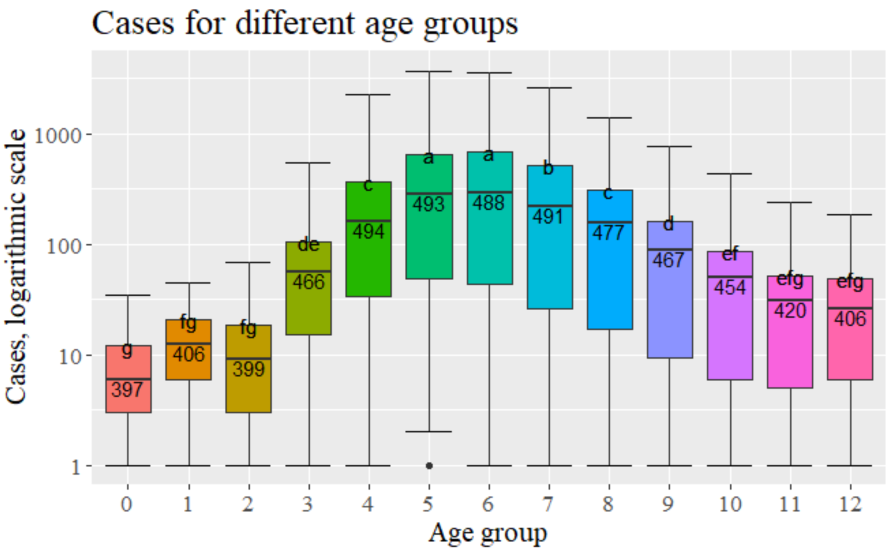
Next, we perform the anova test to see the difference in the number of cases between the age groups. The anova test shows a p-value is much lower than 5% which means that there is a statistically significant difference between age groups. We then plot the boxplot of the number of cases by age group, the number of cases is on a logarithmic scale to allow for a better visualization. The numbers shown in the boxplot reflect the median number of cases for each age group and the letters denote if there is a statistically pairwise significant difference. If two groups have no shared letters, then there is a statistically significant difference in number of cases between age groups. For instance, group 6 has the letter a and 8 has c, which means that they are significantly different. The pairwise comparison is done using the tukey test.

```
anova_test <- aov(Cases ~ Age.at.Diagnosis.Code, data = aids_data %>%
              mutate(Age.at.Diagnosis.Code = as.factor(Age.at.Diagnosis.Code)))
summary(anova_test)
```



```
                        Df    Sum Sq Mean Sq F value Pr(>F)    
Age.at.Diagnosis.Code   12 110697517 9224793   231.9 <2e-16 ***
Residuals             5845 232553865   39787                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
anova_test <- aov(Cases ~ Age.at.Diagnosis.Code, data = aids_data %>%
              mutate(Age.at.Diagnosis.Code = as.factor(Age.at.Diagnosis.Code)))
summary(anova_test)
tukey_result <- TukeyHSD(anova_test, conf.level=.95)

cld <- multcompLetters4(anova_test, tukey_result)
Tk <- aids_data %>%
  mutate(Age.at.Diagnosis.Code = as.factor(Age.at.Diagnosis.Code)) %>%
  group_by(Age.at.Diagnosis.Code) %>%
  summarise(mean=mean(Cases, na.rm = TRUE), quant = quantile(Cases, probs = 0.75,
                                     na.rm = TRUE)) %>%
  arrange(desc(mean))
# extracting the compact letter display and adding to the Tk table
cld <- as.data.frame.list(cld$Age.at.Diagnosis.Code)
Tk$cld <- cld$Letters

aids_data %>%
  mutate(Age.at.Diagnosis.Code = as.factor(Age.at.Diagnosis.Code)) %>%
  ggplot(aes(x = Age.at.Diagnosis.Code, y = Cases, fill = Age.at.Diagnosis.Code)) +
  stat_boxplot(geom ='errorbar', outlier.shape = 1) +
  geom_boxplot() +
  geom_text(data = Tk, aes(x = Age.at.Diagnosis.Code, y = quant, label = cld)) +
  # Add counts by group to boxplot
  annotate("text",
          x = 1:length(table(aids_data$Age.at.Diagnosis.Code)),
          y = aggregate(Cases ~ Age.at.Diagnosis.Code, aids_data, median)[ , 2],
          label = table(aids_data$Age.at.Diagnosis.Code),
          col = "black",
          vjust = 1) +
  ggtitle("Cases for different age groups") + xlab("Age group") +
  ylab("Cases, logarithmic scale") +
  scale_y_continuous(trans='log10') +
  geom_text(data = Tk, aes(x = Age.at.Diagnosis.Code, y = quant, label = cld))+
  theme(legend.position = "none", text=element_text(size=16,  family="serif"))
```

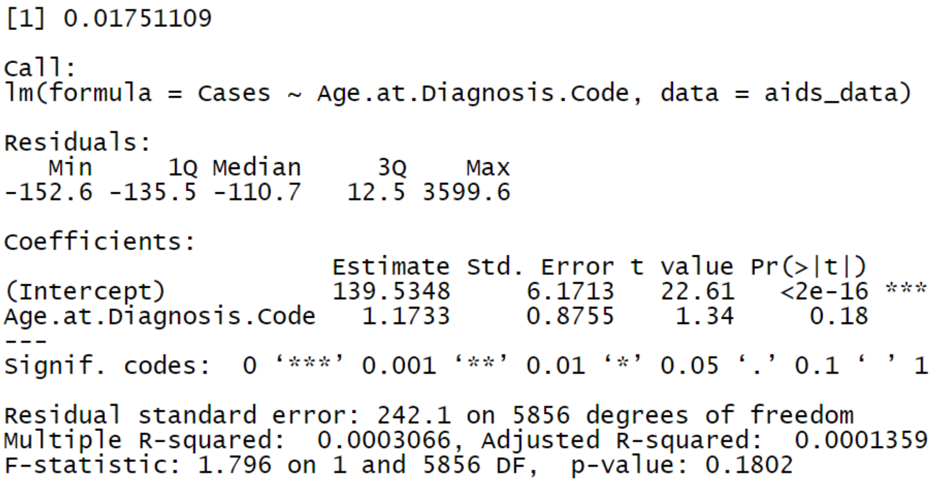


iv. **CORRELATION AND LINEAR REGRESSION MODEL**

The pearson correlation coefficient between age and the number of cases is estimated at 1.7%, which is very low. The linear regression model shows a p-value for the estimate of age of 18% which is significantly higher than 5%. Hence, there is no linear relationship between age and the number of cases. The estimate is equal to 1.17, which means that an increase in age, is estimated to increase the number of cases, in average, by 1.17.

```
cor(aids_data$Age.at.Diagnosis.Code, aids_data$Cases)
```

```
linear_model <- lm(Cases ~ Age.at.Diagnosis.Code, data = aids_data)
summary(linear_model)
```

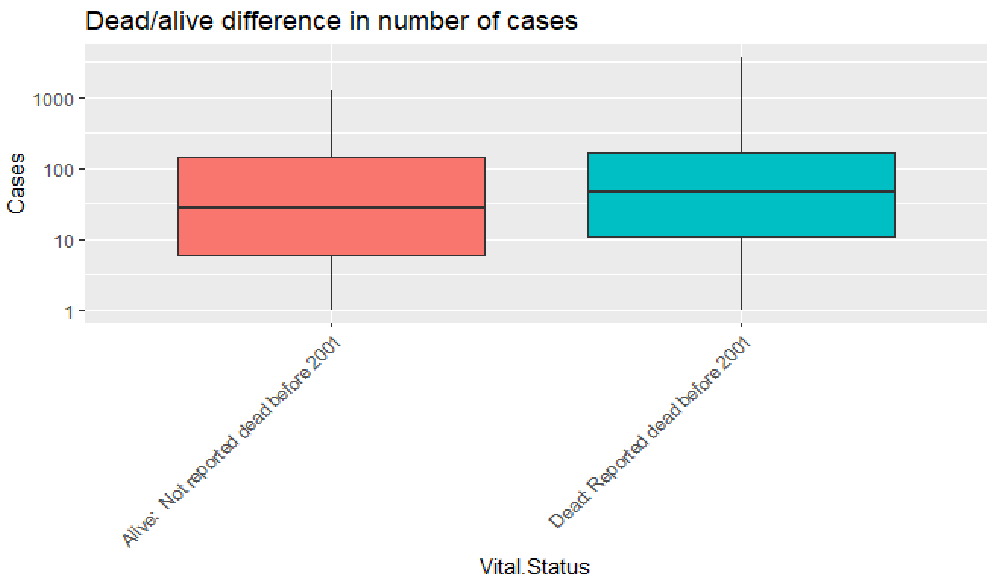


```
[1] 0.01751109

Call:
lm(formula = Cases ~ Age.at.Diagnosis.Code, data = aids_data)

Residuals:
   Min     1Q Median     3Q    Max 
-152.6 -135.5 -110.7   12.5 3599.6 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           139.5348     6.1713   22.61   <2e-16 ***
Age.at.Diagnosis.Code   1.1733     0.8755    1.34     0.18    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 242.1 on 5856 degrees of freedom
Multiple R-squared:  0.0003066,	Adjusted R-squared:  0.0001359 
F-statistic: 1.796 on 1 and 5856 DF,  p-value: 0.1802
```

v. IMPACT OF VITALITY STATUS

To see if there is a difference in the number of cases based on vital status (alive or dead), we plot the boxplot of the number of cases grouped by the vital status on a logarithmic scale. We can see that there is no statistically significant difference in the number of cases between the two categories.

```
aids_data %>%
 ggplot(aes(x = Vital.Status, y = Cases,
          fill = Vital.Status)) +
 geom_boxplot() +
 theme(legend.position = "none",
       axis.text.x = element_text(angle = 45, hjust = 1)) +
 ggtitle("Dead/alive difference in number of cases") +
 scale_y_continuous(trans='log10')
```



4. CONCLUSION

In conclusion, we have confirmed that during the 1990s, when public awareness of aids increased, there was a peak in the number of new cases reported (which were almost certainly cumulative from previous years), and that the number of cases gradually decreased until it reached an average of around 4000 cases per month across all categories. However, although the influence of age is not linearly significant, it is still present and we have seen a statistically significant difference across age categories when it comes to the total number of instances. The vital status, on the other hand, shows no sign of change.

5. BIBLIOGRAPHY

[1] US Department of Health and Human Services (US DHHS), Centers for Disease Control and Prevention (CDC), National Center for HIV, STD and TB Prevention (NCHSTP), AIDS Public Information Data Set (APIDS) US Surveillance Data for 1981-2002, CDC WONDER On-line Database, December 2005.

[2] Weir, S. S., Pailman, C., Mahlalela, X., Coetzee, N., Meidany, F., & Boerma, J. T. (2003). From people to places: focusing AIDS prevention efforts where it matters most. *AIDS, 17*(6), 895–903. https://doi.org/10.1097/00002030-200304110-00015

[3] Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

[4] Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C., & Heymans, M. W. (2012). Missing data: a systematic review of how they are reported and handled. Epidemiology (Cambridge, Mass.), 23(5), 729–732. https://doi.org/10.1097/EDE.0b013e3182576cdb

[5] http://www.sthda.com/english/wiki/ggplot2-axis-scales-and-transformations
(last accessed 08th May 2022)

[6] https://rpubs.com/cyobero/187387
(last accessed 06th May 2022)