Strathmore
UNIVERSITY

A health economic model: Prediction of the prevention of mother-to-child transmission intervention cost in Central Kenya.

Natacia Rutendo Magombo - 101520

Submitted in partial fulfillment of the requirements for the Degree of Bachelor of Business Science in Actuarial Science at Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

February 2021

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the Research Project contains no material previously published or written by another person except where due reference is made in the Research Project itself.

Natacia Rutendo Magombo

February 9, 2021

This Research Project has been submitted for examination with my approval as the Supervisor.

Dr Collins Odhiambo

February 9, 2021

Strathmore Institute of Mathematical Sciences

Strathmore University

Contents

Many initiatives have been put in place to reduce the incidences and deaths due to HIV. As per the cabinet secretary of the ministry of health, Hon. James Wainaina Macharia's words, in the (National AIDS Control Council, 2014), increasing domestic and sustainable financing for HIV is priority. One of the most pursued intervention amongst the HIV programs is the prevention of mother to child transmission (PMTCT) and the target is to eliminate mother to child

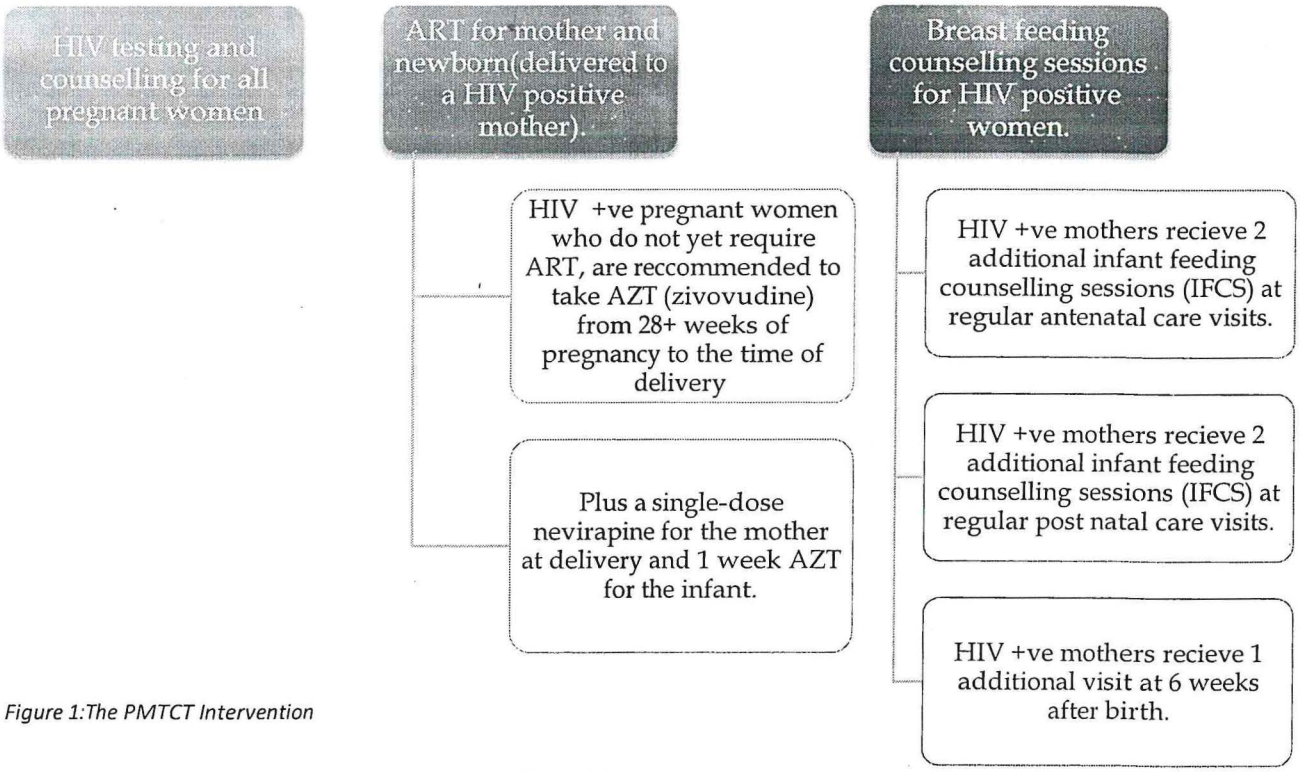


Figure 1:The PMTCT Intervention

transmission (eMTCT) by 2021. The Intervention is summarized in Fig 1. To show its determination, Kenya was quick to adopt the new PMTCT guidelines in 2012. The new improvement was the introduction of the Option B+ into the treatment regimen for pregnant women. The country has shown an increase in PMTCT coverage and results show that women under these new guidelines were less likely to transmit the HIV virus (Pricilla et al., 2018).

PMTCT intervention, among many other interventions, requires constant monitoring on the budget and policies put in force. One of the methods applied in adequate financial planning and policy making in health economics field is a cost analysis (Vassall et al, 2017). In most HIV intervention cost studies, the perspective is likely to be that of the provider. The provider perspective gives information on the expenditure required for the health providers or funders to deliver services (Vassall et al, 2017). This costing perspective proves to be more relevant since

HIV interventions are funded for in Kenya. The international HIV funding community has been very active in helping to tackle the epidemic over the years, especially in low and middle-income countries. As of late, there has been emphasis on HIV domestic funds to encourage more sustainable methods of spending the funds as a result of improved country ownership. This is aimed at helping each government to manage the funds more effectively and take more initiative towards the epidemic. As a result of the new concerns, the America's President's Emergency Plan for AIDS Relief (PEPFAR) has been cutting its funds globally and this action does not spare Kenya. According to the Health gap organization (Milanga, 2019), the Kenya PEPFAR fund was cut from the $505 million that was given in 2018 to $375 million in 2019. Kenya will have to start making crucial investment decisions to make sustain the intervention.

The significance of calculating the costs borne by the patients can reveal the actual affordability of an intervention as shown by Pillai et al. (2019). The study on HIV patient cost in Gauteng revealed that even though HIV services are free in South Africa, healthcare can still be expensive for low income patients. Patient costs can be estimated using statistical models such as linear, non-linear and generalized linear regression models (Jones, 2010).

Problem Statement

The PMTCT intervention is the most successful HIV prevention measure (Plessis et al., 2014). The program is heavily funded in Kenya by donors such as PEPFAR and the Global fund, leaving the patients with only the direct non-medical costs to bear. The ongoing progressive fund cuts have left questions on sustainability of the intervention and cost studies can help the government make crucial financial decisions (Creese et al,.1994). A provider's perspective has been given in the government reports, what is missing is a cost analysis taking on a patient cost perspective.

Research Objective

Objective

To use a health economic model to predict the expected cost of the prevention of mother to child transmission intervention to the patient.

Research question

I. What are the patient costs pertaining to PMTCT patients?
II. Are the costs borne by the patients bearable?
III. Which variables significantly affect the PMTCT patient costs?

Significance of research

I. This study will help the, national boards concerned with HIV in Kenya to be aware of the discrepancies (if any) between the available resources and what is required.
II. This research will then be useful for re-strategizing the PMTCT policy so that the available funds are effectively used and that
III. Pregnant women with HIV will be the prime end beneficiaries to this research. Also, initiators of the PMTCT programs, as their effectiveness is dependent on the funds available to them.

Chapter 2:Literature Review

Depending on the objective of the study, cost can be calculated and predicted using different methods. The method of calculation and forecasting is quite simple when the perspective is that of the provider. The objective of such researches or projects are usually aimed at calculating the intervention expenditure on the supply side and in most cases, the outcomes of the intervention. The supply commonly being the government, NGO's or Healthcare providers. Most cost estimations with statistical approaches take a societal perspective with a focus on the patient cost data, as we will review later. Cost data can be used to make predictions of disease-specific healthcare cost burdens (Gregori et al., 2011). The effect or outcome of the intervention is not being assessed in this paper, hence, only the cost of PMTCT is being analyzed.

Provider Cost Perspective

Costing begins with specifying the purpose, the population of interest and the period in which the data collection is conducted. Vassall et al. (2017) and Creese et al. (1994) are both standardized frameworks for costing healthcare programs. These frameworks contain empirical methods of analyzing the cost data and predicting future costs. There are two costing approaches used in both frameworks is the bottom-up approach, the bottom-up and top-down approach. The bottom-up approach involves getting the costs and quantities of the activities or inputs used for the program and using them to form the aggregate cost of the program. Top-down approach does the opposite and is more concerned with getting the aggregate cost and applying it to the inputs or activities. This method is best when there is no extensive data on the program being analyzed and it does not require much time and expenses of getting the data. The trade off to all those benefits is that the top-down method is less accurate compared to the bottom-up costing approach. (Vassall et al., 2017). It is encouraged that data be collected over a recent year period to avoid distortions such as seasonal effects from shorter periods and irrelevance from older data.

Cost data must be grouped. The following factors should be considered when grouping the costs; the relevancy of the cost to the problem at hand, the mutual exclusivity of the classes (to avoid double counting of expenditures) and lastly, the exhaustion of all possibilities in the health program. The following are examples of cost classes: Resource inputs; source of funding; currency; activity; and level of use (Creese et al., 1994). Splitting costs into homogenous groups

is necessary when costing a health program. Some costs are more sensitive or essential than the other and so, grouping them enables one to treat them with extra care when modelling.

The costs of the intervention can be viewed in two ways, financially or economically. Financial cost is the money paid for (or the value of) the resources used. It is therefore price oriented. Financial costs often give answers to budget related issues. They are useful for assessing the amount spent between periods of time and how expenditure varies as the program matures. Financial costs do not account for the opportunity cost that is found in the health programmes such as the opportunity cost of the volunteers in the program. That cost is not met by the budget allocations but by another separate entity such as, the government or citizens. Using financial costs in themselves to calculate future cost demands on the program is not sufficient for decision making. The financial cost projection passively assume that donors will continue to donate in the future, hence costs paid for by the donors are not included in the projection. Financial costs are therefore not adequate to analyse sustainability (Creese et al., 1994). The economic view allows the researchers to look at the sustainability of the program or to compare the benefit between programs (U.S. Centers for Diseases Control and Kenya Ministry of Health, 2013). It is commonly used to assess the effectiveness of the program and is not a replacement of the financial costs but rather an additional component.

In empirical methods concerning costing, how to calculate capital inputs should be well established and clarified. Inputs can be divided into two sub-groups, that is, recurrent inputs and capital inputs. Recurrent inputs tend to be short term and are repurchased in every period, typically a year. A perfect example in a PMTCT health program can be the ARV drugs taken by a HIV woman planning on getting pregnant. Capital inputs, on the other hand, are long-term and their use can be over several period. Assumptions need to be made when deciding on when to account for their expenditure and the method that will be applied. When calculating the expected cost per chosen period of the intervention, capital costs can cause a bias if not spread throughout the periods. Both Vassall et al. (2017) and Creese et al. (1994) provide a method on how to spread the financial cost for capital inputs. This is simply the current cost of purchasing the capital item divided by the useful life from the date of purchase and this gives the average annual cost. This smoothens out the pattern of spending funds in the cost data since most capital is purchased in the first period. Through this method, we see that financial costs do not account for the time value of money.

Vassall et al. (2017) brings in the aspect of annualization, which is further exhausted by Walker and Kumaranayake in their paper on discounting and annualization. According to Walker and Kumaranayake (2002), annualization permits the combination of recurrent cost incurred in a specific year and the equivalent capital cost for that same year. Annualization can branch into two perspectives: economical and financial cost. In the financial cost view, we use the straight-line depreciation method. This method assumes that the value of the capital input is used up equally in its lifetime. This was illustrated in the previous paragraph. The simple depreciation method is not appropriate when handling economic costs due to their nature. We use the amortization method when calculating the economic cost per period for the capital input. All three , Vassall et al. (2017); Creese et al. (1994); and Walker and Kumaranayake (2002) outline the same method on amortization (also known as annualization method). To calculate the economic cost attributed to a certain year for a capital item, we need the replacement cost of capital, its useful life of the capital and the discount rate. We form an annualization factor using the equation: $\frac{[(1+r)^n-1]}{[r(1+r)^n]}$, where (n) is the useful life of the capital item and r is the discount rate (Walker an Kumaranayake, 2002). Instead, we can use annualization tables which give the factor for a given useful life and discount rate found in (Creese, 1994). The minimum discount rate that is been widely acceptable is 3% although this can be different depending on the user and the purpose of the analysis. . Another alternative is using the borrowing rate for the national government to borrow international funds (Vassall et al., 2017).

Patient Cost Perspective

Literature concerning economic evaluations has been based mostly on the providers view such as the government or the health institutions, ignoring the patient's perspective in making policy decisions. This maybe because of the fully funded interventions that offer free services and medical products to the patients such as the PMTCT intervention. The patient cost consists of direct medical cost, direct non-medical cost and indirect costs and to wholly estimate these costs is a challenge because the data available might not necessarily have those details. There is also inconsistency of the patient's cost definition among studies (Tai, 2016). Cost from the patient's perspective is defined as the amount paid out of pocket for healthcare services (AMA Journal of Ethics, November 2015). More developing countries are accommodating the use of economic evaluations to inform their policy making decisions and a choice of perspective has a big role to

play in the evaluations. The patient perspective of cost analysis in Kenya would be good to give a picture of the costs still borne by the patients after all that funding. The costs borne by the patients can pose as barriers to the elimination of mother-to-child transmission as drawn from the studies on HIV patient costs in South Africa (Pillai et al., 2019).

The costing models applied on patient costs data are rather more complex compared to the previous approach. Econometric regression models can be applied in the analysis of the healthcare cost data. The commonly used examples of these models are Simple linear regression models; Nonlinear regression models and Generalized linear models. The basic concept of these models is that we are analyzing the relationship between the costs data and the characteristics of the patients. In other words, how are the attributes of the individual patients affecting the cost? (Gregori et al., 2011). The "garbage-in-garbage" concept applies even in the healthcare cost analysis. The data available will determine the model to be fitted and if any corrective measures need to be applied to the cost data. Individual patient cost data has the following main issues: zero-cost – an indication that no actual cost have been recorded for the patient , skewness - a measure of asymmetry of the distribution variable and censoring – the value of an observation that is only partially known (Gregori et al., 2011). The extent to which the issues occur can be reduced by collecting data prospectively. Although this option might not be feasible.

Simple Linear Regression

In simple linear regression models, the individual cost data is a function of an explanatory variable(s) and residuals(also known as error terms). The residuals of the individuals are assumed to follow a normal distribution with a zero mean. This can be shown in an equation as:

$$y_i = x_i'\beta + \varepsilon_i \ , \ \varepsilon_i \sim N(0, \sigma^2)$$

This assumption allows the modelling of the expected mean cost to depend on the characteristic(s) of the patients.

$$E(y_i) = x_i'\beta$$

The above equation and OLS assumptions can be used to estimate the parameters of the regression model. This model has an advantage on its simplicity. Estimating the parameters does not require much computation and are easy to interpret (Jones, 2010). The issue with this model is that the cost data and error terms might not necessarily follow a normal distribution

because of the nature of the cost data. To 'fix' this, the data can be transformed by applying logs, square roots, other power functions or the Box-Cox. The transformations aim to somewhat normalize the skewed data (Gregori et al., 2011; Jones, 2010). This also poses other problems such as interpretation of the parameters and in the Box-Cox case, the model might become over specified.

Non-Linear Regression

In these models, the cost data and the variables have a non-linear relationship. The basic exponential conditional mean model that illustrates this relationship is:

$$E[y_i|x_i] = \phi\, exp(x_i'\beta)$$

The equation above can then be extended further to form a poisson, negative binomial and hazard models (Jones, 2010). Some models in this group address the latent variable issue.

Generalized Linear Models, GLMs

GLMs are the most used models to estimate and predict individual cost data. The cost data is said to follow a probability distribution,F. The conditional mean of the cost data is then specified directly to a probability distribution function and this is shown by: $E[y_i|x_i] = f(x_i'\beta)$.

If the chosen probability distribution is an exponential function, then $E[y_i|x_i] = exp(x_i'\beta)$. The model features a link function, $g(.)$. This is the link between the conditional mean (which can be written as, μ_i) and the linear function of the explanatory variables, $x_i'\beta$. It tells us how the conditional mean relates to the explanatory variables (Jones, 2010). This can be illustrated in the equation below:

$$g\big(E(y_i|x_i)\big) = g(\mu_i) = x_i'\beta$$

Hence:

$$\mu_i = g^{-1}(x_i'\beta) = f(x'_i\beta)$$

With GLMs, prediction and estimation of cost is focused on the means, (μ_i) and that makes them attractive to researchers (Barber and Thompson, 2004). Other advantages are that no transformations are required on the data.(Jones, 2010). GLM comes with a problem of choosing the appropriate link and a distribution that extensively describes the cost data. The cost distribution should also describe the relationship between the variance and a function of the mean, such that: $Var(y_i) \propto \mu_i^{\lambda}$, where λ is an integer whose value is based on the distribution

family (Barber and Thompson, 2004). There can also be other alternatives to the functions of the mean such as the quadratic form for other base distribution families.

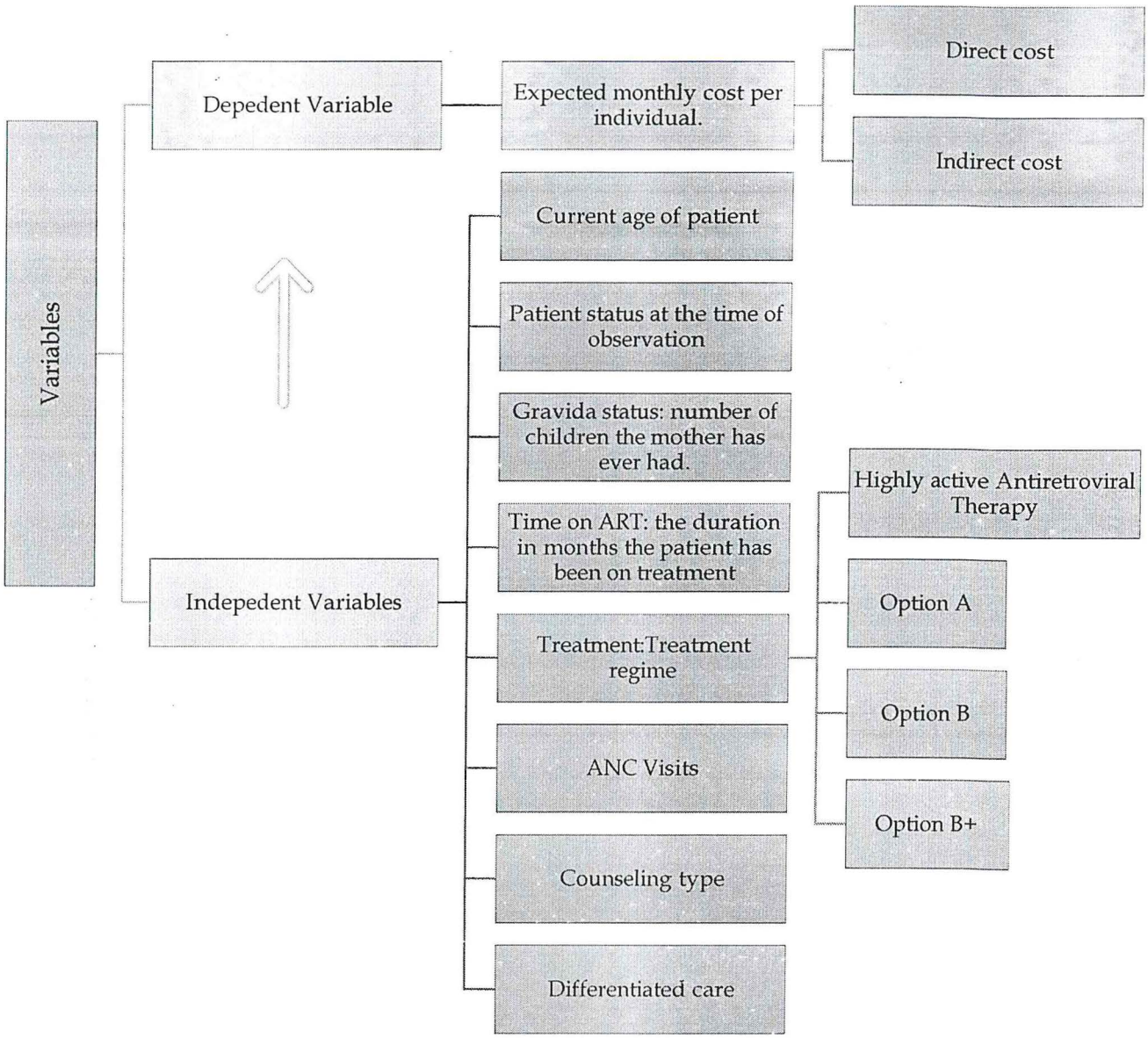
Other Issues

Among the three main data issues apparent in cost data, skewness has already been dealt with. The issue of zero-cost is dominant in studies that take on a broader societal point of view. This is because, not every individual will spend on the service of interest. For example, if we are studying the cost of HIV PMTCT services, not all HIV patients will go for any treatment related with PMTCT. In such a case, the patient costs are given the zero variable because they have made no payment toward the service of interest in the study. This issue can be solved from a model point of view or from a conceptual point of view (Gregori et al., 2011). The idea of latent variables and the use of models such as Tobit models, Threshold legit models and mixed models may satisfy the zero-cost condition. Another commonly used method to deal with many zero costs in data is to use a two-part model as implemented by Sande et al., (2018).

Censoring, as described before, is when the value of an observation is only partially known. In most studies, censoring is a problem because the total cost of an individual in a given period is not known. This may be due to death before the period of study was complete in a trial study. In such cases, Lin (2000) shows that survival models might not be appropriate for this kind of censoring. The survival models can also be inserted in a two-part model to account for both zero cost and censoring (Gregori et al., 2011).

Conceptual framework

Variables of interest



I expect a causal relationship between the following variables:

Chapter 3: Research Methodology

Research design

In this quantitative research, we want to analyse the cross-sectional patient cost data and predict the expected cost of the intervention for the patient costs. Observed data will be collected from a secondary source and analysed using GLM techniques on R.

Setting and Population

The costs for receiving PMTCT services are catered by the donors in Kenya rather than the patients. Although the patients still must cater for other non-medical costs such as transportation to access the medical facilities. The observed data describes a group of patients that get their PMTCT services from 13 level III and IV facilities. The 13 hospitals are in Central Kenya. The HIV prevalence is 6% in that region. The PMTCT services in Kenya are uniformly implemented as per WHO guidelines.

Data

The data was collected from a secondary source. It is retrospective data on the individual costs of 212 patients that were observed over a period between January 2013 and April 2018. To predict the patient cost of the intervention, other characteristics that may be correlated to both the direct and indirect costs have been provided. The information of the following is given: age of patient at observation, number of children the patient has ever had, current HIV status, time on ARV, direct cost per month, treatment regimen, Antenatal visits, counselling type (Antenatal and Postnatal care), differentiated care, and the indirect cost per month.

In this study, direct costs are aggregated amounts of all the costs required to deliver the intervention to an individual per month given the above-mentioned variables. Indirect costs are the aggregated costs incurred by the patient to access the services, also known as 'out-of-pocket' costs.

As expected of individual healthcare cost data (Barber and Thompson, 2004), the cost data has a skewed distribution. The data does not exhibit censoring nor zero-costs. In a case where the data only has a skewness problem, the highly preferable model to be specified to the data is the GLM model (Gregori et al.,2011). The GLM model will be used for this study.

Model Specification

The GLM model is flexible and does not require transformation of the data. Instead, it allows for the skewness of the data. (Jones, 2010; Gregory et al., 2011; Barber and Thompson, 2004; Deb and Norton, 2018). The model has the following components which will be calculated:

1. A linear predictor:

$$y_i = x_i'\beta$$

$$x_i'\beta = \sum_{k=1}^{K} x_{ik}\beta_k$$

where $x_i'\beta$ is a linear index of covariates and coefficients (Deb and Norton, 2018) So in this study, $y_i = (age\beta_1 + \cdots + differentiatedCare\beta_k)$. The cost data is known to follow a distribution $F, y_i \sim F$.

2. A link function:

$$g\big(E(y_i|x_i)\big) = g(\mu_i) = x_i'\beta$$

Hence:

$$\mu_i = g^{-1}(x_i'\beta) = f(x'_i\beta)$$

This is a relationship between the expected cost and the characteristics of the patients (covariates). It is expected that the appropriate link for healthcare expenditure is the log-link (Deb and Norton, 2018) and the most commonly used is the identity link when the covariates have an additive effect to the expected cost rather than a multiplicative one (Jones, 2011).

3. The variance function, which is a function $v(.)$ of the mean of the distribution.

$$var(y_i|x_i) = v(\mu_i)$$

The distribution family of the dependent variable, the cost, is used to relate the conditional mean of the cost to the covariates (Jones, 2010). To estimate the parameters, the MLE (maximum likelihood estimator) is used.

Chapter 4: Data Analysis

The research project was aimed at estimating a GLM model that will predict the expected cost, including both the direct and indirect cost, pertaining to the 'prevention of mother to child transmission intervention patient. Using the data described in Chapter 3, a new variable was created by combining the indirect and direct cost per month.

Exploratory data analysis

The objective of an Exploratory data analysis is to unearth patterns, verify any assumptions and to identify anomalies.

Descriptive and Summary statistics

The data contains five discrete variables which are well described below. Each variable has 212

```
descrete.data

 5  Variables      212  Observations
---------------------------------------------------------------------
PatientID
       n  missing distinct
     212        0      212

lowest : ST10  ST100 ST101 ST102 ST103, highest: ST95  ST96  ST97  ST98  ST99
---------------------------------------------------------------------
CurrentStatus
       n  missing distinct
     212        0        3

Value            Active         LTFU Transfer-out
Frequency           185            7           20
Proportion        0.873        0.033        0.094
---------------------------------------------------------------------
Treatment
       n  missing distinct
     212        0        4

Value          HAART  Option A  Option B Option B+
Frequency        183        12        10         7
Proportion     0.863     0.057     0.047     0.033
---------------------------------------------------------------------
DifferentaitedCare
       n  missing distinct
     212        0        2

Value         No   Yes
Frequency     15   197
Proportion 0.071 0.929
---------------------------------------------------------------------
Counseling_Type
       n  missing distinct
     212        0        2

Value          ANC ANC/PNC
Frequency      117      95
Proportion   0.552   0.448
---------------------------------------------------------------------
```

observations and non-missing values. Each patient has a distinct ID. According to the summary above, the common patient found in this region has a current active status, is on HAART treatment, has differentiated care and attends the ANC counselling sessions.

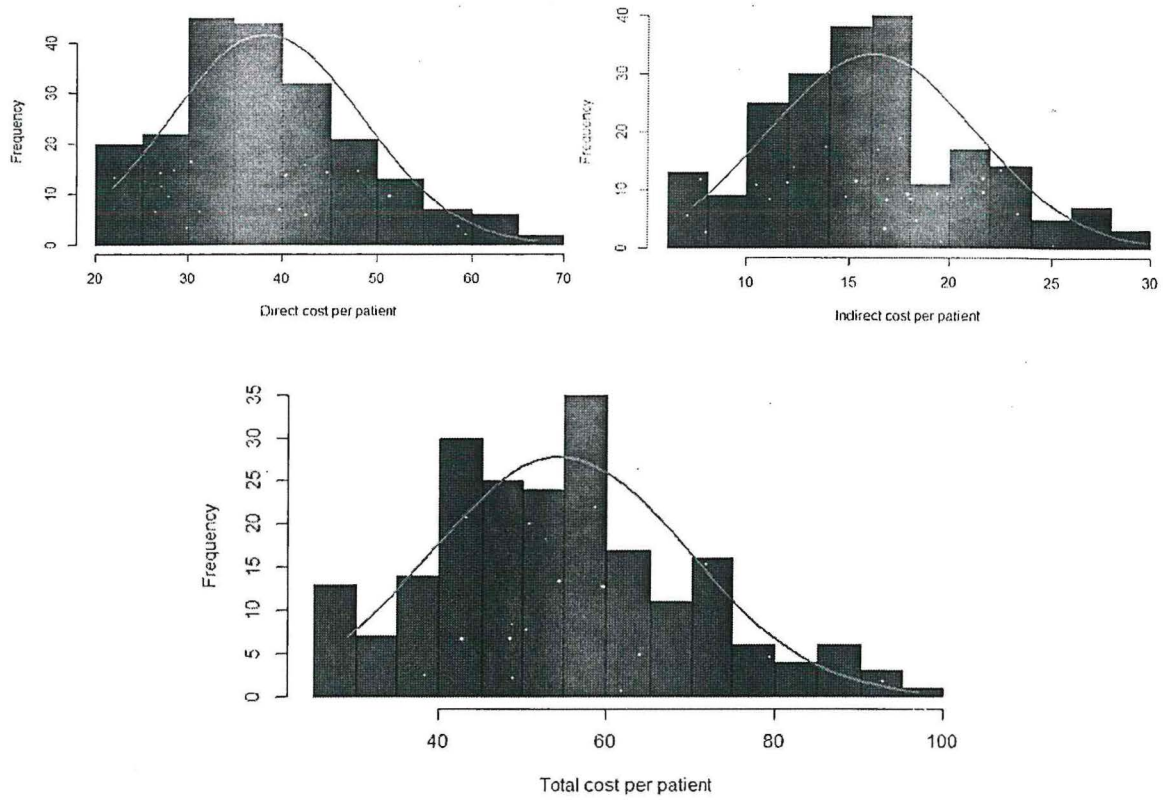
The table below represents the summarizes the continuous data variables. The `pastecs` package was used to produce a table with both the descriptive statistics (such as the mean, standard deviation) and basic statistics (such as the number of data points and the number of null data).

	CurrentAge	Gravida	ANCvisits	TimeonART	Directcost	Indirectcost	Totalcost
nbr.val	212.00	212.000	212.00	212.00	212.00	212.00	212.00
nbr.null	0.00	7.000	7.00	0.00	0.00	0.00	0.00
nbr.na	0.00	0.000	0.00	0.00	0.00	0.00	0.00
min	20.00	0.000	0.00	5.00	21.82	7.15	28.97
max	40.00	5.000	20.00	43.00	67.03	30.00	97.03
range	20.00	5.000	20.00	38.00	45.21	22.85	68.06
sum	6083.00	381.000	1524.00	6284.00	8094.83	3417.74	11512.57
median	28.50	1.500	6.00	31.00	37.25	16.00	53.19
mean	28.69	1.797	7.19	29.64	38.18	16.12	54.30
SE.mean	0.35	0.074	0.30	0.53	0.70	0.35	1.04
CI.mean.0.95	0.69	0.146	0.58	1.04	1.37	0.69	2.06
var	26.13	1.158	18.52	58.81	102.88	25.82	230.81
std.dev	5.11	1.076	4.30	7.67	10.14	5.08	15.19
coef.var	0.18	0.599	0.60	0.26	0.27	0.32	0.28

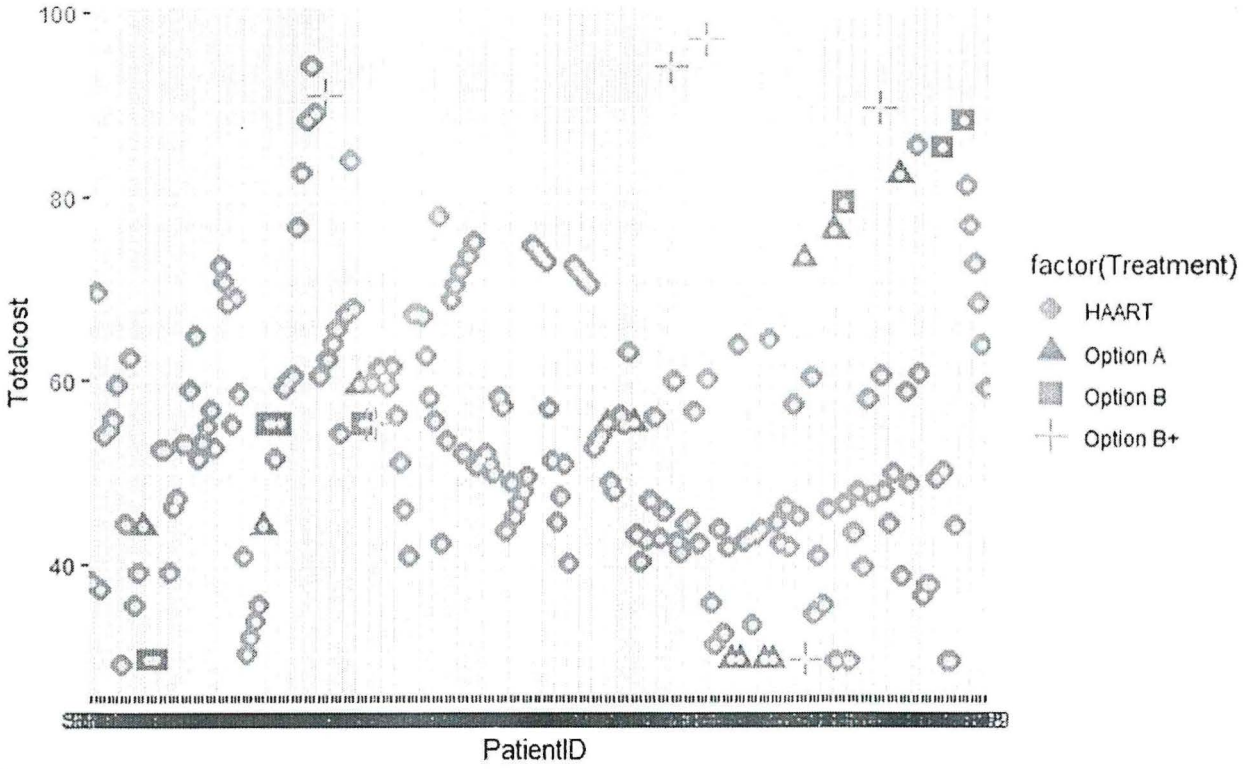
Visual Analysis

The visual analysis on R was done using the `tiyverse` and `ggplot2` libraries. In the visual analysis section, the trends of the cost data were more apparent. We made use of tools such as the scatter plot and the bar graph with a normal curve. By plotting the graphs, we also got insights of the customers as described in the table under Descriptive and Summary statistics.

BAR GRAPHS



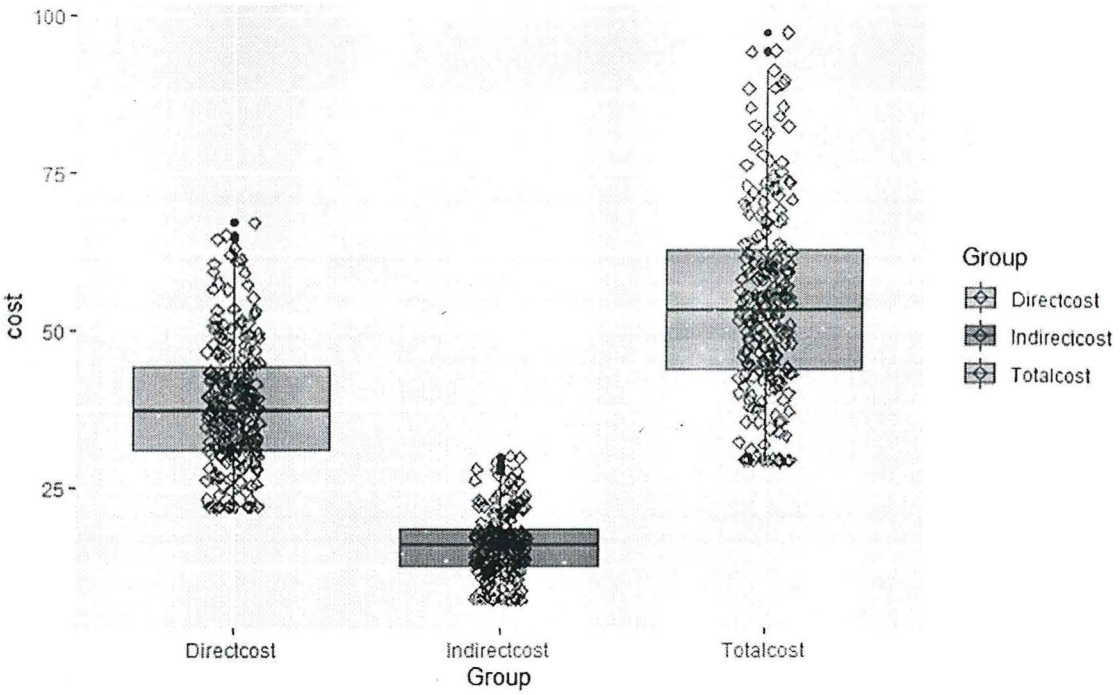
The bar graphs present the frequencies of the cost ranges. The direct cost data appeared rightly skewed and the highest cost frequency among the patients ranged between $30-40. It is unimodal because the data has a single peak as shown above. Due to the similar values of the mode (55.2), mean (54.3) and the median (53.19) and by simple judgement of the smoothed curve, we approximated that the indirect cost graph is approximately normal. The third graph is a combination of the direct and indirect cost.



SCATTER PLOT

The scatter plot gives insight on the total cost data per patient. This scatter plot differentiates the color and shape of the data points according to the treatment the patient received. From the graph above, most patients in Central Kenya received the HAART treatment. Option B+ treatment was the most expensive, ranging from approximately, $89-98. This excludes the 3 outliers, two of them with a cost of approximately $55 and one with a cost of $30.

BOX PLOT



The box plots are used to visualize how the cost data is distributed. Indirect cost is generally lower than the direct cost.

The 50th percentile of the total cost ranged from $44-63. The cost data boxplots are negatively skewed, with a few outliers indicated by the black solid points.

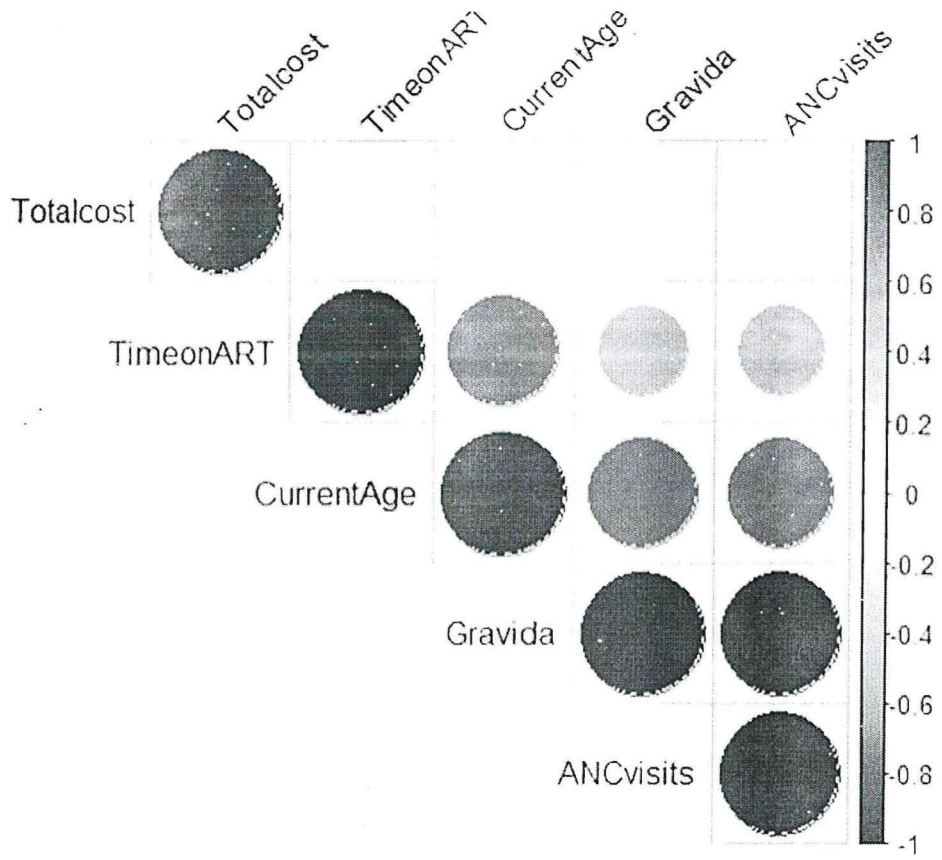
CORRELATION

	CurrentAge	Gravida	ANCvisits	TimeonART	Totalcost
CurrentAge	1.00	0.78	0.78	0.78	-0.10
Gravida	0.78	1.00	1.00	0.54	-0.04
ANCvisits	0.78	1.00	1.00	0.54	-0.04
TimeonART	0.78	0.54	0.54	1.00	0.02
Totalcost	-0.10	-0.04	-0.04	0.02	1.00

n= 212

P

	CurrentAge	Gravida	ANCvisits	TimeonART	Totalcost
CurrentAge		0.0000	0.0000	0.0000	0.1312
Gravida	0.0000		0.0000	0.0000	0.5729
ANCvisits	0.0000	0.0000		0.0000	0.5729
TimeonART	0.0000	0.0000	0.0000		0.7315
Totalcost	0.1312	0.5729	0.5729	0.7315	



We also visualized the correlation matrix as shown below.

These are the correlation coefficients found between the continuous variables in our data. The 'Gravida', 'ANCvisits' and 'TimeonART' variables were highly correlated with the patient's age and they were statistically significant. 'ANCvisits' and the 'Gravida' variable had a perfect positive correlation, which indicates perfect multicollinearity. Multicollinearity is an interdependent disorder, according to Farrar and Glauber (1967). It is defined in terms of lack of independence, which means high inter-correlation within a set of variables.

This means that each variable out of the two can explain the other using a perfect linear function and that can be a problem when coming up with the linear function. The 'Totalcost' variable had low correlation coefficients with the explanatory variables. This is not a good thing since the explanatory variables should be highly related to indicate that they can explain the response variable. Nevertheless, correlation does not always indicate a causal relationship and especially when the correlation coefficients are insignificant.

Stepwise regression Analysis

Stepwise regression consists of recursively removing unnecessary determinants in the predictive model, to find a subset of variables in the data set leading to the most efficient model,

```
Call:
lm(formula = Totalcost ~ CurrentAge + Gravida + TimeonART, data = regdata1)

Residuals:
    Min      1Q  Median      3Q     Max
-27.841  -9.619   0.155   7.325  43.809

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.3718     7.5748   9.554  < 2e-16 ***
CurrentAge   -1.3891     0.4430  -3.136  0.00196 **
Gravida       2.3366     1.5599   1.498  0.13567
TimeonART     0.5935     0.2182   2.720  0.00708 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.92 on 208 degrees of freedom
Multiple R-squared:  0.04927,   Adjusted R-squared:  0.03556
F-statistic: 3.593 on 3 and 208 DF,  p-value: 0.01453

1  linear dependencies foundSubset selection object
Call: regsubsets.formula(Totalcost ~ ., data = regdata1, nvmax = 3,
    method = "seqrep")
4 Variables  (and intercept)
           Forced in Forced out
CurrentAge     FALSE      FALSE
Gravida        FALSE      FALSE
TimeonART      FALSE      FALSE
ANCvisits      FALSE      FALSE
1 subsets of each size up to 3
Selection Algorithm: 'sequential replacement'
         CurrentAge Gravida TimeonART ANCvisits
1  ( 1 ) "*"        " "     " "       " "
2  ( 1 ) "*"        " "     "*"       " "
3  ( 1 ) "*"        "*"     "*"       " "
```

i.e., a model that reduces the error of estimation.

There are three strategies for step-by-step regression (Stepwise Regression Essentials in R, 2018): Backward selection, forward selections and stepwise selection. Stepwise selection, which is a mixture of forward and back selections was used in this study. We started with no predictors, then sequentially added the most contributory predictors (like forward selection). After adding the new variables, we removed the variables that no longer enhanced the model fit (like backward selection).This was all done in R.

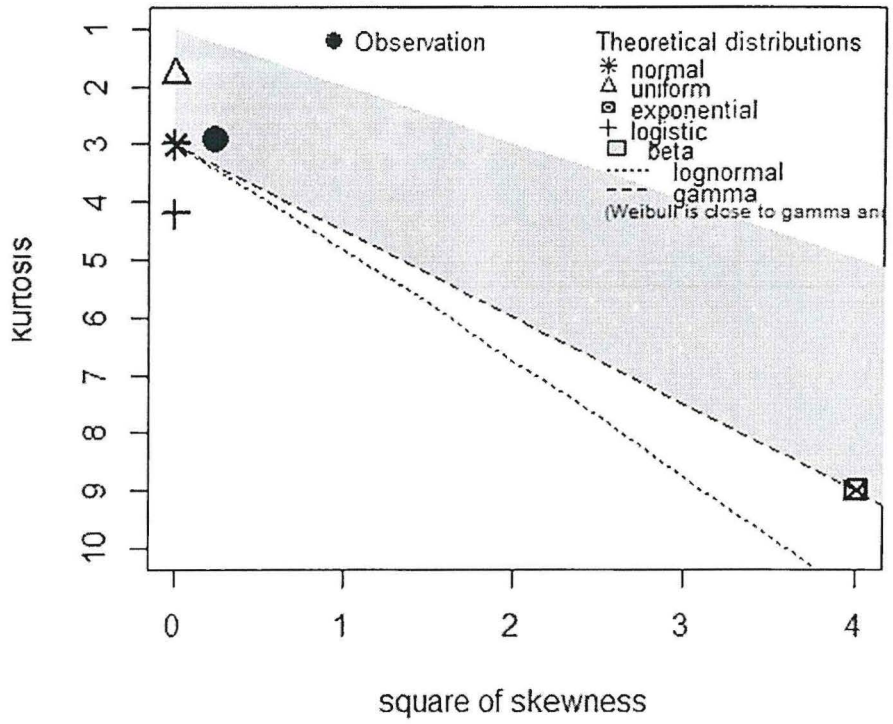
```
  nvmax     RMSE   Rsquared      MAE   RMSESD RsquaredSD    MAESD
1     1 15.08317 0.06537964 12.18167 1.704188 0.07753230 1.345685
2     2 14.88728 0.05438763 11.93533 1.777514 0.06565981 1.425631
3     3 14.89534 0.06297146 11.94723 1.820797 0.10100163 1.494907
  nvmax
2     2
[1] "the best model is the one with 2 variables"
Subset selection object
4 Variables  (and intercept)
           Forced in Forced out
CurrentAge     FALSE      FALSE
Gravida        FALSE      FALSE
TimeonART      FALSE      FALSE
ANCvisits      FALSE      FALSE
1 subsets of each size up to 2
Selection Algorithm: backward
         CurrentAge Gravida TimeonART ANCvisits
1  ( 1 ) "*"        " "     " "       " "
2  ( 1 ) "*"        " "     "*"       " "
(Intercept)  CurrentAge   TimeonART
 65.2885829  -0.9328320   0.5324327
```

Three possible regressions were generated with different number of variables, ranging from 1-3, because we had set the maximum to be two using the function, `nvmax`. According to (Stepwise Regression Essentials in R, 2018), the best model is the one with the least Root square mean error (RSME). In our case, the best model was one with two variables. The best multilinear regression was:

$Totalcost_i = 65.289 - 0.933 CurrentAge_i + 0.532 TimeonART_i$.

Family link

To find the GLM that best fits the data, we first determined the distribution that best fits the Total cost data. According to (Delignette-Muller & Dutang, 2015, p. 17), we can use the Collen and Frey graph to help make the choice of the distributions that fit the data.



Cullen and Frey Graph

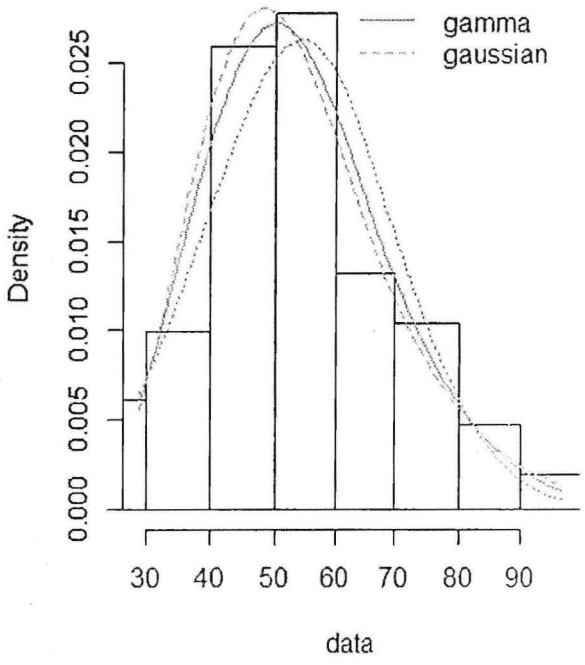
The closest distributions were the Gamma, Lognormal and the Normal distributions.

Results

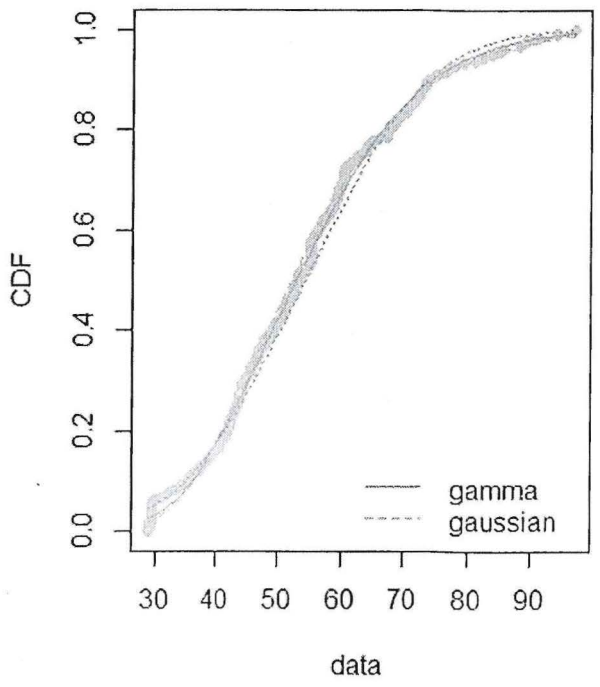
According to the results shown below, the Gamma distribution was the best fit. This conclusion was drawn from the visual results and the gamma line having the lowest AIC(Akaike Information Criterion) and BIC (Schwarz's Bayesian information criteria).

```
Fitting of the distribution ' norm ' by maximum likelihood
Parameters :
     estimate Std. Error
mean 54.30458  1.0409504
sd   15.15647  0.7360631
Loglikelihood:  -877.1215   AIC:  1758.243   BIC:  1764.956
Correlation matrix:
     mean sd
mean    1  0
sd      0  1

Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters :
         estimate Std. Error
meanlog 3.9552987 0.01944049
sdlog   0.2830578 0.01374573
Loglikelihood:  -871.7722   AIC:  1747.544   BIC:  1754.258
Correlation matrix:
              meanlog         sdlog
meanlog  1.000000e+00 -7.594955e-12
sdlog   -7.594955e-12  1.000000e+00

Fitting of the distribution ' gamma ' by maximum likelihood
Parameters :
        estimate Std. Error
shape 12.8842285 1.23519696
rate   0.2372627 0.02319405
Loglikelihood:  -871.1458   AIC:  1746.292   BIC:  1753.005
Correlation matrix:
          shape      rate
shape 1.0000000 0.9806587
rate  0.9806587 1.0000000
```

Histogram and theoretical densities



Empirical and theoretical CDFs



Chapter 5:Discussion

The aim of this project was to fit the total cost data to a GLM, and this model is to be used to predict the patient's total cost, given the patient details. In this study, the patient details that had a significant effect on the total cost were, the time in months spent on ART and their age.

GLM Results

```
Call:
glm(formula = Totalcost ~ CurrentAge + TimeonART, family = Gamma(link = "identi
ty"),
    data = regdata1)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.57015  -0.20154  -0.00316   0.15704   0.68357

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  65.0781     5.9380  10.960  < 2e-16 ***
CurrentAge   -0.9449     0.3181  -2.970  0.00332 **
TimeonART     0.5512     0.2022   2.727  0.00694 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.07590855)

    Null deviance: 16.667  on 211  degrees of freedom
Residual deviance: 16.003  on 209  degrees of freedom
AIC: 1741.6

Number of Fisher Scoring iterations: 5
```

The following is the summary output of the GLM.

The results above show a summary of the deviance residuals. As per the results, the median value was fairly close to zero and this shows that our model is not prejudiced in one way (R on datascienceblog.net: R for Data Science, 2018).

Our model had a low Null deviance value. This not good since it indicates that the model is a good enough predictor with no other explanatory variables apart from the intercept. On the other hand, the residual deviance was low, and this is a good thing because it shows that the model was a good fit for the data. Also, the coefficients of the explanatory variables were both significant. All in all, the GLM model was a good fit for the data.

Predictions

We used the model to predict the total cost required by a patient to receive treatment per month

```
      fit             se.fit        residual.scale
 Min.   :47.13   Min.   :1.102   Min.   :0.2755
 1st Qu.:51.99   1st Qu.:1.196   1st Qu.:0.2755
 Median :54.63   Median :1.540   Median :0.2755
 Mean   :54.31   Mean   :1.658   Mean   :0.2755
 3rd Qu.:57.30   3rd Qu.:1.891   3rd Qu.:0.2755
 Max.   :58.90   Max.   :3.501   Max.   :0.2755
```

in the Central region. Below is a summary of the predictions for the total cost.

The variable, 'fit' are the fitted values of the total cost from the model. The average total cost of a PMTCT patient in that region was $54.31 per month, with a standard deviation of $1.66. The total cost ranged from $47.13 to an approximate maximum of $58.90.

To be able to answer the second research question indicated in Chapter 1, we also predicted the cost borne by the patients. The cost required for a PMTCT patient to receive treatment is not necessarily the cost borne by the patients because the direct costs are subsidized. Therefore, we

```
      fit             se.fit          residual.scale
 Min.   :32.95   Min.   :0.7318   Min.   :9.99
 1st Qu.:37.08   1st Qu.:0.8207   1st Qu.:9.99
 Median :38.55   Median :0.9915   Median :9.99
 Mean   :38.18   Mean   :1.1236   Mean   :9.99
 3rd Qu.:40.01   3rd Qu.:1.3170   3rd Qu.:9.99
 Max.   :40.89   Max.   :2.5641   Max.   :9.99
```

also predicted the direct cost, and the summary results are shown below.

The average predicted direct cost is $38.18. Hence, the cost borne by the patient ranged from $8.95 to $20.72 on an average.

Limitations

Unlike most healthcare intervention cost estimation research, this research investigates sustainability from a wholesome point of view. That is, the indirect cost born by the patients are included to get a full picture. The limitation we faced is that the data received is second hand and the indirect cost data might have been an estimation rather than the real experience of the patients.

The total cost data had a low correlation with its explanatory variables. This might have limited the results we got.

Conclusion

This study was carried out to create a mathematically economic tool, so as to estimate the expected cost of a PMTCT intervention patient. The study focused on the Patients in the Central region of Kenya which has an HIV prevalence of 6%. Unlike other studies, we took into consideration the cost still borne by the patient even when the direct cost is subsidized.

According to the research, a GLM model can be specified onto intervention cost data and can be used to predict cost. In the current study, we found that it costs an average total of $54.31 for a patient in the Central Region of Kenya to access and receive PMTCT treatment. Although, through the international funding, the direct cost is subsidized in Kenya, leaving the patient with an average range of $8.95 to $20.72 to bear.

Understanding the individual total cost can help the government in making more efficient decisions concerning the intervention and it will help end the HIV pandemic. If the cost to access the treatment is too high, this might become a hinderance to the patient from an effective treatment experience.

References

National AIDS and STI Control Programme (NASCOP). (2019). Preliminary KENPHIA 2018 Report. NASCOP.

National AIDS Control Council. (2014, June). *Kenya AIDS Strategic Framework.* https://nacc.or.ke/kenya-aids-strategic-framework-kasf/

nMilanga, M. (2019, November). *PEPFAR funding cuts threaten the future of people living with HIV in Kenya.* Health GAP (Global Access Project). https://healthgap.org/pepfar-funding-cuts-threaten-the-future-of-people-living-with-hiv-in-kenya/

du Plessis, E., Shaw, S. Y., Gichuhi, M., Gelmon, L., Estambale, B. B., Lester, R., Kimani, J., & Avery, L. S. (2014). Prevention of mother-to-child transmission of HIV in Kenya: challenges to implementation. *BMC health services research, 14 Suppl 1*(Suppl 1), S10. https://doi.org/10.1186/1472-6963-14-S1-S10

Vassall, A., Sweeney, S., Kahn, J., Gomez, G., Bollinger, L., & Marseille, E., & Herzel, B., Plosky, W., Cunnama, L., Sinanovic, E., & Group, GHCC & Group, GHCC & Harris, K., Levin, C. (2017). Reference case for estimating the costs of global health services and Interventions.

Gregori, D., Petrinco, M., Bo, S., Desideri, A., Merletti, F., & Pagano, E. (2011). Regression models for analyzing costs and their determinants in health care: an introductory review. *International journal for quality in health care : journal of the International Society for Quality in Health Care, 23 3,* 331-4.

Creese, A.L., Parker, D., & World Health Organization. (1994). Cost analysis in primary health care : a training manual for programme managers / edited by Andrew Creese and David Parker. World Health Organization. https://apps.who.int/iris/handle/10665/40030

Walker, D.G., & Kumaranayake, L. (2002). Allowing for differential timing in cost analyses: discounting and annualization. *Health policy and planning, 17 1,* 112-8 .

Tai, B. B., Bae, Y. H., & Le, Q. A. (2016). A Systematic Review of Health Economic Evaluation Studies Using the Patient's Perspective. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research, 19*(6), 903-908. https://doi.org/10.1016/j.jval.2016.05.010

The challenge of understanding health care costs and charges. (n.d.). Journal of Ethics | American Medical Association. https://journalofethics.ama-assn.org/article/challenge-understanding-health-care-costs-and-charges/2015-11

Pillai, N., Foster, N., Hanifa, Y., Ndlovu, N., Fielding, K., Churchyard, G., Chihota, V., Grant, A. D., & Vassall, A. (2019). Patient costs incurred by people living with HIV/AIDS prior to ART initiation in primary healthcare facilities in Gauteng, South Africa. *PLOS ONE, 14*(2), e0210622. https://doi.org/10.1371/journal.pone.0210622

Jones, A.M. (2010). HEDG Working Paper 10 / 01 Models For Health Care.

Barber, J., & Thompson, S. (2004). Multiple regression of cost data: Use of generalised linear models. *Journal of Health Services Research & Policy, 9*(4), 197-204. https://doi.org/10.1258/1355819042250249

Sande, L., Maheswaran, H., Mangenah, C., Mwenge, L., Indravudh, P., Mkandawire, P., Ahmed, N., D'Elbee, M., Johnson, C., Hatzold, K., Corbett, E. L., Neuman, M., & Terris-Prestholt, F. (2018). Costs of accessing HIV testing services among rural Malawi communities. *AIDS Care, 30*(sup3), 27-36. https://doi.org/10.1080/09540121.2018.1479032

Lin, D. (2000). Linear regression analysis of censored medical costs. *Biostatistics, 1*(1), 35-47. https://doi.org/10.1093/biostatistics/1.1.35

Deb, P., & Norton, E. C. (2018). Modeling health care expenditures and use. *Annual Review of Public Health, 39*(1), 489-505. https://doi.org/10.1146/annurev-publhealth-040617-013517

Pricilla, R. A., Brown, M., Wexler, C., Maloba, M., Gautney, B. J., & Finocchario-Kessler, S.(2018). Progress toward eliminating mother to child transmission of HIV in Kenya: Review of treatment guidelines uptake and pediatric transmission between 2013 and 2016—A follow up. Maternal and Child Health Journal, 22(12), 1685-1692. https://doi.org/10.1007/s10995-018-2612-0

U.S. Centers for Diseases Control and Kenya Ministry of Health, (2013), The Cost of Comprehensive HIV Treatment in Kenya. Report of a Cost Study of HIV Treatment Programs in Kenya. Atlanta, GA (USA) and Nairobi, Kenya.

Stepwise Regression Essentials in R. (2018, March 11). Articles - STHDA. http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/#computing-stepwise-regression

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. The Review of Economics and Statistics, 49(1), 92. https://doi.org/10.2307/1937887

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: AnRPackage for Fitting Distributions. *Journal of Statistical Software*, *64*(4), 1–23. https://doi.org/10.18637/jss.v064.i04

R on datascienceblog.net: R for Data Science. (2018, November 10). *Interpreting Generalized Linear Models*. R-Bloggers. https://www.r-bloggers.com/2018/11/interpreting-generalized-linear-models/

APPENDIX

Codes used for the Exploratory Data Analysis.

```
#Descrete data
install.packages("Hmisc")
library(Hmisc)
descrete.data= data[,c(1,6,7,8,9)]
describe(descrete.data)

#Continuous data
install.packages("pastecs")
library(pastecs)
continuous.data= data[,c(2,3,4,5,12)]
options(scipen=100)
options(digits=2)
stat.desc(continuous.data)

#Bar graph
x <- data$Totalcost
h<-hist(x, breaks=10, col="blue", xlab="Total cost per patient",
        main="Bar graph with Normal Curve")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="red", lwd=2)

#Scatterplot
scatter_plot <- ggplot(data, aes(x = PatientID, y = Totalcost)) +
    geom_point(aes(shape = factor(Treatment), color = factor(Treatment)), siz
e = 4) + geom_point(color = "grey90", size = 1.5)
scatter_plot


#Box plot
```

```
a = data.frame(Group = "Totalcost", cost = c(data$Totalcost))
b = data.frame(Group = "Directcost", cost = c(data$Directcost))
c = data.frame(Group = "Indirectcost", cost = c(data$Indirectcost))
Costs = rbind(a,b,c)
ggplot(Costs, aes(x=Group, y=cost, fill=Group)) +
  geom_boxplot()+ geom_jitter(shape=5, position=position_jitter(0.1))

#Correlation
df = data[,c(2,5,6,8,11)] #This is to remain with the continuous variables fr
om the dataset.
library(Hmisc)
Corr_matrix<- rcorr(as.matrix(df))
Corr_matrix

#Visual Correlation
install.packages("corrplot")
library(corrplot)
M<-cor(df)
corrplot(M, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45, is.corr = TRUE)
```

Stepwise regression Analysis codes

```
#Load packages, MASS, caret, leaps and tidyverse
library(tidyverse)
library(caret)
library(leaps)
library(MASS)
#Create new data frame with only the "Totalcost" dependent variable and the c
ontinuous explanatory variables
data<-read.csv("costdata.csv")
regdata1<- data[, c("CurrentAge", "Gravida", "TimeonART", "ANCvisits", "Total
cost")]
# Fit the full model
full.model <- lm(Totalcost ~., data = regdata1)
# Stepwise regression model
step.model <- stepAIC(full.model, direction = "both",
                      trace = FALSE)
summary(step.model)
models <- regsubsets(Totalcost~., data = regdata1, nvmax = 3,
                     method = "seqrep")
summary(models)
# Set seed for reproducibility
set.seed(123)
# Set up repeated k-fold cross-validation
train.control <- trainControl(method = "cv", number = 10)
# Train the model
step.model <- train(Totalcost ~., data = regdata1,
                    method = "leapBackward",
                    tuneGrid = data.frame(nvmax = 1:3),
                    trControl = train.control
)
#Results
step.model$results
step.model$bestTune
```

```
"the best model is the one with 2 variables"
summary(step.model$finalModel)
coef(step.model$finalModel, 2)
```