Predictors of Antenatal Care Count Data in Bangladesh: relative importance analysis

M.S. Thesis

Submitted by

Kaniz Fatema

Registration no: 2015-017-257

Exam roll: 36504

Session: 2019-20

Under the supervision of

Md. Zillur Rahman Shabuz, PhD

Associate Professor

Department of Statistics

University of Dhaka, Dhaka-1000

August, 2022

Acknowledgment

Firstly, I would like to express my admiration to Almighty who give me the opportunity and the patience to accomplish my thesis under my respected supervisor, Md. Zillur Rahman Shabuz, PhD, Associated Professor, Department of Statistics, Bio-statistics & Informatics, University of Dhaka, who has inspired, guided and encouraged me to accomplish my thesis. His simple presentation to a complex topics, valuable advice and continuous support have helped me to overcome the problems and complete my work. My heartiest gratitude goes to my respected supervisor.

I want to give my sincere thanks to my respected teacher Professor Sayema Sharmin, Chairman, Department of Statistics, Bio-statistics & Informatics, University of Dhaka, for permitting me to do this research work and give me the opportunity to use the seminar as well as the computer and other facilities of this department for this purpose.

I want to express my gratitude to Md. Ershadul Haque, Assistant Professor, Department of Statistics, Bio-statistics & Informatics, University of Dhaka, Ummay Nayeema Islam, lecturer, Department of Statistics, Bio-statistics, & Informatics, University of Dhaka and Nasrin Sultana, Assistant Professor, Department of Statistics, Bio-statistics, & Informatics, University of Dhaka, for helping me in computation and construction of data.

Among my friends, I want to thank specially Md. Lutful Kader who helped me in coding and helping me in different parts of my thesis.

My deeply thanks goes to my friends Halima Akter prova, Progna paul, Fatema jasmin and all of my friends for their continuous help, cooperation and encouragement at different stages of this research.

Finally, I am really grateful to my beloved parents, my elder brother and sister for their continuous love, support and encouragement throughout my whole study period in this department.

Kaniz Fatema

August, 2020

Abstract

Given that maternal mortality is a major global health concern, multiple measures including antenatal care visits have been promoted by the global community. However, most pregnant women in low and medium income countries, primarily in Sub-Saharan Africa do not attain the recommended timelines, in addition to a slower progress towards meeting the required minimum of eight visits stipulated by the World Health Organization. In regression analysis researcher's commonly face the situation of relative importance of the variable. Similar to the multiple linear regressions, researchers may be interested to find out the most important predictor among the several predictors in a Poisson regression model. There exists several method for measuring relative importance among the independent variable. Dominance analysis technique has been widely used by researchers to examine the predictor importance more accurately in the linear regression. Budescu (1993) introduced this most popular methods, dominance analysis. Azen and Budescu refined and expanded this (2003). Therefore, this study explored the trends in antenatal care visits and the associated factors in Bangladesh from BDHS, 2017-18 data using the multiple covariates. Women who are less likely to achieve optimal antenatal care visits should be targeted by policies for reducing moralities and other complications. Poverty-reduction policies, promoting maternal and girl education, improving general livelihood in rural settings, expanding health facility coverage and infrastructural access, are required to increase antenatal care utilization.

Contents

Acknowledgment **I**

Abstract **IV**

1 Introduction **1**

1.1 Preface 1

1.2 Literature review 3

1.3 Objective of this study 5

1.4 Organization of the Study 6

2 Methodology **7**

2.1 Poisson Regression 7

2.2 Inference Procedure of Regression Parameters 10

2.2.1 Estimation of regression parameters 10

2.3 Global test 13

2.4 Local test 14

2.4.1 Subset of parameter 14

2.4.2 Individual parameter 16

2.5 Overdispersion 17

2.5.1 Detection of Overdispersion 17

2.5.2 Test of overdispersion 18
2.6 Negative Binomial Regression Model 19
2.6.1 Inference Procedure of Regression Parameter 19
2.6.2 Negative Binomial (NB2) regression as GLM 19
2.6.3 Estimation of parameter vector, $\boldsymbol{\theta}$ 23
2.7 Method of measuring variable importance in regression analysis 33

3 Data and variable **38**
3.1 Source of Data 38
3.2 Variables 40

4 Results **43**
4.1 Univariate analysis 43
4.2 Bivariate analysis 46
4.3 Application of Poisson and Negative binomial (NB) regression to the number of ANC visits 48
4.3.1 Poisson regression model 48
4.3.2 Overdispersion tests 51
4.3.3 Negative binomial regression model 52
4.3.4 Result obtained from Dominance analysis 55

5 Conclusion and discussion **58**
5.1 Discussion 58
5.2 Recommendation 60
5.3 Further scope of the study 60

References **62**

Appendix **65**

List of Figures

4.1 Bar diagram of number of ANC visits 43

4.2 Relative importance among the predictors 57

List of Tables

2.1 Dominanc analysis for several variable 36

3.1 Variables and their categories 42

4.1 Frequency distribution and corresponding percentage distribution of the categories of the covariates 45

4.2 Bivariate associations between the number of antenatal care visits and different predictors in Bangladesh. 47

4.3 Estimates, p-values and incidence rate ratio (IRR) of Poisson regression model for the determinants of number of ANC visits. 49

4.4 Score tets detection overdispersion to the number of ANC visits in Bangladesh 51

4.5 Estimates, p-values and incidence rate ratio (IRR) of Negative Binomial regression model for the determinants of number of ANC visits. 53

4.6 Variable importance obtained using Poisson regression model 55

4.7 Variable importance obtained using negative binomial regression model 56

Chapter 1

Introduction

1.1 Preface

Maternal health care throughout pregnancy is essential to a country's development. Globally, around 3,00,000 pregnant women die each year because of pregnancy and complications during delivery (Sarker et al., 2020); (Duodu et al., 2022). Only two out of every three pregnant women (65%) receive at least four antenatal appointments, even though 86 percent of pregnant women receive expert health care at least once. The vast majority of maternal deaths (94%) happened in low and medium income countries, primarily in Sub-Saharan Africa. Perhaps fewer women got at least four antenatal visits, such like Sub-Saharan Africa and South Asia (52 percent and 49 percent, respectively) (UNICEF).Women who miss ANC appointments are more likely to have pregnancy complications which including preeclampsia, eclampsia, and anemia, and also a higher probability of severe birth outcomes such as preterm birth, weight problem and stillbirth (Abbas, Rabeea, Hafiz, & Ahmed, 2017). Worldwide, the early ANC visits increased from 40.9% to 58.6% from the year 1990 to 2013. However, in developed and developing countries this result differs. Only four ANC visits decrease

newborn mortality, the new recommendation raises the number of visits a pregnant woman has with health practitioners from four to eight during her pregnancy (WHO). Recent research suggests that having more antenatal contact with a health practitioner by mothers and adolescent girls is linked to a lower risk of stillbirth. When compared to four ANC visits , eight or more contacts can lower prenatal mortality by up to eight per 1000 newborns. The proportion of mothers getting appropriate antenatal care (4 to 7) remains nearly constant between 2006 to 2011 and increased from 49.98% in 2011 to 58.61% in 2017-2018. These have prompted calls for further action to address the problem, as highlighted in SDG 3, which aims to minimize worldwide maternal mortality about less than 70 per 100,000 live births by 2030 (Duodu et al., 2022).
According to the 2007 BDHS, only 22% of pregnant women got four or more ANC visits. The percentage of ANC visits increased from 22% in 2007 to 47% in 2017-2018 (NIPORT, 2020). The structure of the Health, Population, and Nutrition Sector Development Program (HPNSDP) calls for 50% of pregnant women to join at least four prenatal care meetings by 2016 (N. Sultana & Bari, 2017);(M. Sultana et al., 2017). Hence, Bangladesh is considerably behind in meeting this goal. Because of a scarcity of health-care providers and establishments, nearly three-quarters of Bangladeshi mothers (73%) do not receive four or more ANC visits from qualified health experts, let alone the eight 'contacts' recommendation of the World Health Organization (WHO) (Jo et al., 2019).

None of the studies rank the order of the determinants of the number of ANC. The number of ANC visits is count response and Poisson regression is a popular approach to analyze count data. Similar to multiple linear regressions, researchers may be interested to find out the most important predictor among the several predictors in a

Poisson regression model. Dominance analysis technique has been widely used by researchers to examine the predictors importance in the linear regression. Dominance analysis was suggested by Budescu (Budescu, 1993) and was extended and modified by Azen and Budescu (Azen & Budescu, 2003). Recently Azen and Traxel used dominance analysis to logistic regression analysis to rank order the predictors (Azen & Traxel, 2009). We will also use dominance analysis to count regression models to rank order the predictors of the number of ANC visits.

1.2 Literature review

One of the primary significant features of antenatal care is to offer women and their infants with health-related advice and services that can significantly boost their health. The 2016 Ethiopian Demographic and Health Survey was employed to investigate the variables that affect the amount of antenatal care helps in providing among Ethiopian pregnant women. Excess zeros had been discovered in the data (35 percent). As a result, several regression model have been fitted. Result obtained from exploratory analysis it was discovered that 2240 (34.7%) of the pregnant women did not attend any antenatal care service throughout her pregnancy months. The simulation study was implemented to compare models based on their Akaike information criterion value in order to determine the model that best suited for the data. The simulation experimental results demonstrated that zero-inflated data models fit the data effectively than classical models. All of these zero-inflated models was evaluated by comparing

using the Voung test, and the Hurdle model fit the data effectively than any other zero-inflated model, which was described by too much zeros and wide variation in the non-zero findings. According to the Voung test, the Hurdle model fit zero-inflated data effectively compared to any other zero-inflated model (Bekalo & Kebede, 2021).

Another study conducted where find negative binomial regression is best fitted. The results demonstrate that respondents' educational qualification, socio-economic status, place of residence, accessibility, and birth order each have a significant impact on antenatal care service utilization (Islam, Sen, & Bari, 2022).

Using the hurdle negative binomial regression model one study was conducted in Ethiopia. The intention of this article was to figure out the variables that influence child-bearing mothers' use of ANC visits in the Kaffa, and Bench-Maji zones of Ethiopia's Southern Nation Nationalities and Peoples Region. This analysis shows that Hurdle regression fits well (Terefe & Gelaw, 2019).

Another study's goal was to evaluate the condition of ANC utilization and to develop a Bayesian Count Regression model for the factors of ANC visits among pregnant women in the Amhara regional state. It consisted of a society based analytical cross-sectional research of reproductive-age (15–49) women in the Amhara region. According to the findings of this study, in order to increase the number for ANC visits in the Amhara region, women with minimal educational status and rural women should be prioritized (Workie & Lakew, 2018).

Another study in Bangladesh used the hurdle negative binomial regression model for

handling the count data. This model is appropriate for handling over-dispersion and excess zeroes (Bhowmik, Das, & Islam, 2020).

For determining the main factors of ANC visits research has been carried out in Bangladesh using zero-truncated negative binomial regression (0-NBR model). The study's findings show that place of residence, birth order, exposure to mainstream media, socio-economic status, and mother's education have a considerable effect on women's ANC status during pregnancy (Hossain, Akter, Sultana, & Kabir, 2020).

1.3 Objective of this study

The primary goal of this research is to figure out the factors that are associated with the number of ANC visits.

Specific objectives are given below:

1. To estimate the prevalence rate of ANC,
2. To select an appropriate model for identifying the factors associated with the number of ANC visits,
3. To assess the contributions of individual variables to the number of ANC visits in Bangladesh.

1.4 Organization of the Study

This thesis paper is organized in eight chapters. These chapters are

Chapter 1 refers overall short review if relative importance, related literature review where using Poisson and Negative binomial (NB) model and additionally the main purpose of the study.

Chapter 2 refers theoretical description of Poisson regression model and its etimation procedure. Also several kind of test for hypothesis, theoretical part of over-dispersion and several kind of test for checking over-dispersion, theoretical description of Negative binomial regression model and its etimation procedure. This part give the theoretical review of relative importance analysis.

Chapter 3 refers Data and variable used in our study and complete description about the procedure of modifying variable.

Chapter 4 indicates result obtained by Univariate analysis, Bivariate analysis. Application of GLM model (Poisson and NB model) and also provide over-dispersion test. Finally, this part give the results of relative importance using general dominance analysis.

Chapter 5 indicates overall summary of this study and some recommendation after completing this research and some scope for future.

Chapter 2

Methodology

This chapter will discuss theoretical description of Poisson regression and Negative binomial regression model and its etimation procedure. Also several kind of test for overdispersion.

2.1 Poisson Regression

One of the most fundamental distributions for the count data is the Poisson distribution named after the French mathematician and physicist Simeon Denis Poisson who proposed it in 1837 (DeJardine, 2013). Let Let $\boldsymbol{Y} = (Y_1, \cdot\cdot, Y_j, \cdot\cdot, Y_n)$ be the a discrete random variable, It is said to have a Poisson distribution with parameter

$$\lambda > 0.$$

where λ is the shape parameter indicating the expected number of events occurring in the specified interval. because it represents the number of occurrences in a given

interval., Then the probaility mass function (pmf) is given by

$$f(y_i) = \frac{e^{-\lambda_j}\lambda_j^{y_j}}{y_i!}, \quad \lambda_j > 0, \quad y_j = 0, 1, 2, \cdots.$$

Also, suppose $\boldsymbol{X} = (X_1, ..., X_p)^\top$ is the $q \times 1$ vector of covariates with response Y_j and $\boldsymbol{\beta} = (\beta_1, ..., \beta_k, ..., \beta_q)^\top$ is the $q \times 1$ vector of regression coefficient. Poisson count model which is a member of generalized linear model(GLM) because :

(a) The probability distribution function of $Y_j, f(y_j)$, is a member of exponential family with canonical form, i.e.

$$\begin{aligned} f(y_j) &= \frac{e^{-\lambda_j}\lambda_j^{y_j}}{y_j!} \\ &= exp(y_j ln\lambda_j - \lambda_j - lny_j!). \end{aligned}$$

That is, Poisson distribution belongs to exponential family of distribution with

$$\begin{aligned} a(y_j) &= y_j \\ b(\lambda_j) &= ln\lambda_j \\ c(\lambda_j) &= -\lambda_j \\ d(y_j) &= -lny_j!. \end{aligned}$$

where $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ and $d(\cdot)$ are known functions. If $a(y) = y$, the distribution is said to be in canonical (i.e. standard) form. The function $b(\theta)$ is called the natural parameter of the distribution. If there are other parameters in addition to the parameter of interest θ, these are termed as nuisance parameters forming parts of the functions $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ and $d(\cdot)$ and they are treated as if they were known.

Thus, Poisson distribution has a canonical form as $a(y_j) = y_j$ with natural parameter $b(\lambda_j) = ln\lambda_j$. (b) One can use the above natural parameter $b(\lambda_j)$ as a link function in GLM, if

1. $b(\lambda_j)$ is a function of mean $\mu_j = E(y_j)$ i.e. $b(\lambda_j) = g(\mu_j)$;

2. $g(\mu_j)$ is differentiable function of μ_j;

3. The range of $g(\mu_j)$ is between $-\infty$ and $+\infty$.

From exponential family of distributions, one may write

$$E[a(y_j)] = E(y_j) = -\frac{c'(\lambda_j)}{b'(\lambda_j)}$$

$$\Rightarrow \mu_j = -\frac{-1}{\frac{1}{\lambda_j}} = \lambda_j.$$

That is, one can write $b(\lambda_j) = ln\lambda_j = ln\mu_j = g(\mu_j)$.

Now,

$$g'(\mu_j) = \frac{d}{d\mu_j} g(\mu_j) = \frac{d}{d\mu_j} ln\mu_j = \frac{1}{\mu_j}.$$

Note that $g'(\mu_j)$ exists if $\mu_j > 0$. Here, $g'(\mu_j) > 0$ for all $\mu_j > 0$ which implies that $g(\mu)$ is differentiable monotonic increasing function for all values of μ_j, when $\mu_j > 0$.

(c) The range of $g(\mu_j) = ln\mu_j$ is $(-\infty, +\infty)$, since $\mu_j > 0$. Therefore, one can use natural parameter $b(\lambda_j) = g(\mu_j) = ln\lambda_j$ as a link function in GLM.

The GLM for count response is given by

$$ln\lambda_j = \boldsymbol{x}_j^\top \boldsymbol{\beta}$$

That is,

$$\lambda_j = e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}}$$

Thus,

$$\mu_j = \lambda_j = e^{\eta_j} \quad \text{with} \quad \eta_j = \boldsymbol{x}_j^\top \boldsymbol{\beta}.$$

2.2 Inference Procedure of Regression Parameters

2.2.1 Estimation of regression parameters

Under GLM, regression coefficient which is actually regression parameters can be estimated which are given below:

Likelihood function

Let $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ be a random sample of size n from Poisson distribution function $f(y; \lambda)$, where λ is the of parameters to be estimated. So, the likelihood function is

$$\begin{aligned} L(\boldsymbol{\lambda}; \boldsymbol{y}) &= \prod_{j=1}^{n} f(y_j; \lambda_j) \\ &= \prod_{j=1}^{n} exp\left[y_j ln\lambda_j - \lambda_j - lny_j!\right], \quad \text{where } \lambda_j = e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}} \end{aligned} \tag{2.1}$$

Usually, we take the natural logarithm of the likelihood function due to mathematical tractability. Then the log-likelihood function, denoted by $l(\lambda; y)$, is given by

$$\begin{aligned} l(\boldsymbol{\lambda}; \boldsymbol{y}) &= lnL(\boldsymbol{\lambda}; \boldsymbol{y}) \\ &= \sum_{j=1}^{n} (y_j ln\lambda_j - \lambda_j - lny_j!) \\ &= \sum_{j=1}^{n} (y_j ln\mu_j - \mu_j - lny_j!), \text{ as } \lambda_j = \mu_j. \end{aligned} \tag{2.2}$$

Note that the mean response, μ_j is a function of regression parameter, β.

Score function

The score function, denoted by $U(\beta)$, being the first derivative of log-likelihood function with respect to the parameter of interest, can be defined as

$$U(\boldsymbol{\beta}) = \frac{\delta}{\delta\beta}\ l(\boldsymbol{\lambda}; \boldsymbol{y}).$$

The k^{th} element of score function can be obtained as

$$\begin{aligned} U_k(\boldsymbol{\beta}) &= \frac{\delta}{\delta\beta_k}\ l(\boldsymbol{\lambda}; \boldsymbol{y}) \\ &= \sum_{j=1}^{n} \frac{(y_j - \mu_j)}{Var(Y_j)}\ x_{jk}\ \frac{\delta}{\delta\eta_j}\mu_j\ ; \quad k = 1, 2, \cdots, q, \end{aligned}$$

where $\frac{\delta}{\delta\eta_j}\mu_j = e^{\eta_j} = \mu_j$. Therefore,

$$\begin{aligned} U_k(\boldsymbol{\beta}) &= \sum_{j=1}^{n} \frac{(y_j - \mu_j)}{\mu_j}\ x_{jk}\ \mu_j \\ &= \sum_{j=1}^{n} (y_j - \mu_j) x_{jk}\ ; \quad k = 1, 2, \cdots, q. \end{aligned}$$

Information matrix

The information matrix, denoted by $I(\boldsymbol{\beta})$, is the variance covariance matrix of the score function $U(\boldsymbol{\beta})$ which is given by

$$I(\boldsymbol{\beta}) = Var\left[U(\boldsymbol{\beta})\right] = \left[I_{kg}(\boldsymbol{\beta})\right]; \quad k = 1, 2, \cdots, q \; ; \quad g = 1, 2, \cdots, q.$$

The $(k, g)^{th}$ element of the information matrix can be defined as

$$\begin{aligned} I_{kg}(\boldsymbol{\beta}) &= -E\left[\frac{\delta}{\delta\beta_g} U_k(\boldsymbol{\beta})\right] \\ &= \sum_{j=1}^{n} \frac{x_{jk}\; x_{jg}}{Var(y_j)} \left(\frac{\delta}{\delta\eta_j}\mu_j\right)^2 \\ &= \sum_{j=1}^{n} \frac{x_{jk}\; x_{jg}}{\mu_j}\; \mu_j^2 \\ &= \sum_{j=1}^{n} x_{jk}\; x_{jg}\; \mu_j, \quad k = 1, 2, \cdots, q \; ; \quad g = 1, 2, \cdots, q. \end{aligned}$$

Estimating equation

The maximum likelihood estimating equation for β is given by

$$U(\boldsymbol{\beta}) = \mathbf{0}.$$

Using Newton-Raphson iterative procedure that equation can easily be solved for $\boldsymbol{\beta}$. The estimate found at the $r^{th}(r = 1, 2, 3, ...)$ iteration is given by

$$\widehat{\boldsymbol{\beta}}^{r} = \widehat{\boldsymbol{\beta}}^{r-1} + \left[\boldsymbol{I}(\boldsymbol{\beta})\right]^{-1}_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{r-1}} U(\boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{r-1}}.$$

Here, the asymptotic distribution of the estimator $\widehat{\boldsymbol{\beta}}$ for the regression parameter is normal with mean $\boldsymbol{\beta}$ and variance covariance matrix $I^{-1}\boldsymbol{\beta}$ i.e.

$$\widehat{\boldsymbol{\beta}} \backsim N_p[\boldsymbol{\beta}, I^{-1}(\boldsymbol{\beta})] \text{ as } n \to \infty.$$

Therefore, the distribution of $\widehat{\boldsymbol{\beta_k}}$ is $\widehat{\boldsymbol{\beta_k}} \backsim N[\boldsymbol{\beta_k}, I^{kk}(\boldsymbol{\beta})]$ as $n \to \infty$

2.3 Global test

There are three main test for hypotheses about regression parameters $\boldsymbol{\beta}$**.** the first test is the usual test based on asymptotic normality of the maximum likelihood estimates, called Wald's test.

Wald test

For large sample, $\widehat{\boldsymbol{\beta}}$ has a p-variate normal distribution with mean $\boldsymbol{\beta}$ and variance covariance estimated by $I(\widehat{\boldsymbol{\beta}})^{-1}$ A test of global hypothesis of $H_0 \ : \ \boldsymbol{\beta} = \ \boldsymbol{\beta}_0$ is

$\chi^2_w = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^\top \mathbf{I}\left(\widehat{\boldsymbol{\beta}}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$

Which has a large sample chi-square distribution with p degrees of freedom if H_0 is true for large sample.

likelihood ratio test

The second test is the likelihood ratio test of the hypothesis of $H_0 \ : \ \boldsymbol{\beta} = \ \boldsymbol{\beta}_0$ and

$\chi^2_{\text{LR}} = 2[LL\left(\widehat{\boldsymbol{\beta}}\right) - \text{LL}\left[\boldsymbol{\beta}_0\right]$

Which has a large sample chi-square distribution with p degrees of freedom under H_0.

Score test

The third test is the score test. It is based on the different scores, $U(\boldsymbol{\beta}) = (U_1(\boldsymbol{\beta}), U_2(\boldsymbol{\beta}), ..., U_p(\boldsymbol{\beta}))^\top$ For large sample , $U(\boldsymbol{\beta})$ is asymptotically p-variate normal with mean 0 and covariance $I(\boldsymbol{\beta})$ the test of $H_0 \ : \ \boldsymbol{\beta} = \ \boldsymbol{\beta}_0$ is

$\chi^2_{\mathrm{SC}} = \mathbf{U}(\boldsymbol{\beta}_0)^\top \left[\mathbf{I}\left(\boldsymbol{\beta}_0\right)\right]^{-1} \mathbf{U}(\boldsymbol{\beta}_0)$

Which has a large sample chi-square distribution with p degrees of freedom under H_0 for large n.

2.4 Local test

2.4.1 Subset of parameter

If one is interested in testing a hypothesis about a subset of the $\boldsymbol{\beta}$'s. The hypothesis is then $H_0 \ : \ \boldsymbol{\beta}_1 = \ \boldsymbol{\beta}_{10}$, where $\boldsymbol{\beta}$=$(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$. Here $\boldsymbol{\beta}_1$ is a $q \times 1$ vector of the $\boldsymbol{\beta}$'s of interest and $\boldsymbol{\beta}_2$ is the vector of the remaining $p - q$ $\boldsymbol{\beta}$'s.

The Wald test statistic

The Wald test of $H_0 \ : \ \boldsymbol{\beta}_1 = \ \boldsymbol{\beta}_{10}$ is based on the maximum partial likelihood estimators of $\boldsymbol{\beta}$. Let $\widehat{\boldsymbol{\beta}}$=$(\widehat{\boldsymbol{\beta}}_1^\top, \widehat{\boldsymbol{\beta}}_2^\top)^\top$ be the maximum partial likelihood estimator of $\boldsymbol{\beta}$ .suppose we partition the information matrix $\mathbf{I}$ as $\mathbf{I} = \begin{pmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{pmatrix}$

$\chi^2_w = \left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\right)^\top \left[\mathbf{I}^{11}\left(\widehat{\boldsymbol{\beta}}\right)\right]^{-1} (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})$

Where $\mathbf{I}^{11}$ is the upper $q \times q$ sub matrix of $\mathbf{I}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)$. Let $\widehat{\boldsymbol{\beta}}_2$ be the maximum likelihood estimate of $\boldsymbol{\beta}_2$ based on the log likelihood with the first $\boldsymbol{\beta}$'s fixed at a value $\boldsymbol{\beta}_{10}$.

The likelihood ratio test

the likelihood ratio test of the hypothesis of $H_0 \ : \ \boldsymbol{\beta} = \ \boldsymbol{\beta}_{10}$ $\mathbf{H_0} : \boldsymbol{\beta_1} = \boldsymbol{\beta_{10}}$ is expressed as

$\chi^2_{\mathrm{LR}} = 2\{LL\left(\widehat{\boldsymbol{\beta}}\right) - LL\left[\boldsymbol{\beta}_{10},\ \widehat{\boldsymbol{\beta}}_2\left(\boldsymbol{\beta}_{10}\right)\right]\}$

Which has a large sample chi-square distribution with q degrees of freedom under H_0.

The Score test

To test $H_0 \ : \ \boldsymbol{\beta}_1 = \ \boldsymbol{\beta}_{10}$ using the score statistic, let $\mathbf{U}_1\left[\boldsymbol{\beta}_{10},\ \widehat{\boldsymbol{\beta}}_2\left(\boldsymbol{\beta}_{10}\right)\right]$ be the $q \times 1$ vector of scores for $\boldsymbol{\beta}_1$ and at the restricted partial maximum likelihood estimator for $\boldsymbol{\beta}_2$. Then

$\chi^2_{\mathrm{SC}} = \mathbf{U}_1\left[\boldsymbol{\beta}_{10},\ \widehat{\boldsymbol{\beta}}_2\left(\boldsymbol{\beta}_{10}\right)\right]^{\top}[\mathbf{I}^{11}\left(\boldsymbol{\beta}_{10},\ \widehat{\boldsymbol{\beta}}_2\left(\boldsymbol{\beta}_{10}\right)\right)]\mathbf{U}_1\left[\boldsymbol{\beta}_{10},\ \widehat{\boldsymbol{\beta}}_2\left(\boldsymbol{\beta}_{10}\right)\right]$

Which has a large sample chi-square distribution with q degrees of freedom under H_0.

(Klein & Moeschberger, 2003)

2.4.2 Individual parameter

Wald Statistic

Wald Statistic is for $H_0 = \beta = \beta_0$ given below

$$z^2 = \left(\frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} \right)^2$$

z^2 has (asymptotically) a chi-squared distribution with $df = 1$

Wald statistics can be used to test 2-sided alternatives, while z can be used to test 1-sided as well as 2-sided alternatives.

Likelihood Ratio Test Statistic

The Likelihood ratio test statistic for $H_0 = \beta = \beta_0$ equals

$\lambda_{LR} = -2(l(\beta_0) - l(\hat{\beta}))$

Where $l(\hat{\beta}) = log(L(\hat{\beta}))$ and $l(\beta_0) = log(L(\beta_0))$ are the "maximized log-likelihood functions".

Score Test Statistic

The score test statistic equals for $H_0 = \beta = \beta_0$

$\chi^2_{\mathrm{SC}} = \mathbf{U}(\beta_0)\ [\mathbf{I}\,(\beta_0)]^{-1}\ \mathbf{U}(\beta_0)$

Which has a large sample chi-square distribution with p degrees of freedom under $H_0\ :\ \beta =\ \beta_0$

2.5 Overdispersion

Overdispersion in statistics refers to the existence of more variability in a data set than might be expected based on the specified statistical model. When the observed variance is more than the the variance of a theoretical model, this is referred to as overdispersion. Underdispersion, on the other hand, indicates that there would be less variation in the data than expected. When fitting very simple parametric models, like those depending on the Poisson distribution, Overdispersion is frequently encountered. The Poisson distribution only has parameter. This kind of phenomenon occurs for the Poisson model when variance of the observation becomes higher than the mean. Overdispersion should be considered when analyzing count data to avoid making incorrect inferences.

2.5.1 Detection of Overdispersion

The sum of all Pearson residual is defined as the Pearson chi-square statistic which is denoted by χ^2 (McCullagh & Nelder, 2019) , (Wilson, 1989). The Pearson residuals may be defined as follows (Hilbe, 2011)

$$\chi^2 = \sum_{i}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \tag{2.3}$$

The above expression is called the dispersion statistic where μ_i is the expected counts. If value for the dispersion statistics greater than 1, than the model may considered as Overdispersed. In such a situation it is more reliable to switch negative binomial distribution.

2.5.2 Test of overdispersion

For detecting overdispersion, there are two common statistical test such as Z-score, Lagrange multiplier tests (Hossain et al., 2021). We applied these test in our data using Poisson regression model.

Score test

- Dean and Lawless (1989): The formula of Z-score is given by

$$Z_i = \frac{(\ y_i - \mu_i)^2 - \ y_i}{\mu_i \sqrt{(2)}} \tag{2.4}$$

  where μ_i be the expected counts.

- Winkelmann (2008): The formula of Z-score is given by

$$Z_i = \frac{(\ y_i - \mu_i)^2 - \ y_i}{2\mu_i} \tag{2.5}$$

- Cameron and Trivedi (2013): The formula of Z-score is given by

$$Z_i = \frac{(\ y_i - \mu_i)^2 - \ y_i}{\mu_i} \tag{2.6}$$

Lagrange multiplier test

This test has been continued by using χ^2 and the test-statistics is defined (Hilbe,2011) as

$$Ł = \frac{\sum_i^n (\mu_i - y_i)^2}{2\sum_i^n \mu_i^2} \tag{2.7}$$

which follows χ^2 distribution with 1 degrees of freedom.

2.6 Negative Binomial Regression Model

The Poisson regression model is generally too limited for count data., necessitating the use of alternative models. When the equidispersion property of the Poisson distribution is breached, going to result in excess dispersion, the mean and variance are no longer the same.The most common type of dispersion is overdispersion. That is, the nominal variance exceeds the response's mean. The counts are then interpreted as being generated by a Poisson process, however the researcher may be unable to correctly describe the rate parameter of this procedure. Instead, the rate parameter is a random variable.

2.6.1 Inference Procedure of Regression Parameter

2.6.2 Negative Binomial (NB2) regression as GLM

Under a Negative Binomial (NB2) count model, the response variable Y has a Negative Binomial distribution. Suppose that there are n independent response variables $Y_1, \cdot\cdot, Y_j, \cdot\cdot, Y_n$ with the following probability mass function

$$f(y_j) = \frac{\Gamma(y_j + k^{-1})}{y_j!\Gamma(k^{-1})} \left(\frac{1}{1+k\mu_j}\right)^{k^{-1}} \left(\frac{k\mu_j}{1+k\mu_j}\right)^{y_j}; \quad \mu_j > 0,\ k > 0,\ y_j = 0, 1, 2, \cdot\cdot\cdot\cdot$$

$$i.e \;\; Y_j \sim NeBin\left(k^{-1}, k\mu_j\right), \text{ where } k \text{ is the overdispersion parameter.}$$

Now, the mean and variance of the Negative Binomial response, (Y_j) are

$$E(Y_j) = \mu_j \;\; \text{and}$$

$$Var(Y_j) = \mu_j + k\mu_j^2 \;,$$

which implies that the nominal variance exceeds the mean of response since $k > 0$. Also, suppose that $\boldsymbol{x_j} = (x_{j1}, \cdots, x_{jk}, \cdots, x_{jp})^\top$ is the p×1 vector of covariates associated with response Y_j and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_k, \cdots, \boldsymbol{\beta}_p)^\top$ is the p×1 vector of regression coefficients. Let us define the parameter vector, $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, k)'$ to be estimated.

NB model is a member of generalized linear model(GLM) because of the following reasons:

(a) The probability distribution function of Y_j, $f(y_j)$, is a member of exponential family with canonical form, i.e.

$$f(y_j) = \frac{\Gamma(y_j + k^{-1})}{y_j!\Gamma(k^{-1})} \left(\frac{1}{1+k\mu_j}\right)^{k^{-1}} \left(\frac{k\mu_j}{1+k\mu_j}\right)^{y_j}$$

$$= exp\left[y_j ln\left(\frac{k\mu_j}{1+k\mu_j}\right) - \frac{1}{k} ln\left(1+k\mu_j\right) + ln\left(\frac{\Gamma(y_j + k^{-1})}{y_j!\Gamma(k^{-1})}\right)\right]$$

$$= exp\left[a(y_j)b(\mu_j) + c(\mu_j) + d(y_j)\right].$$

That is, Negative Binomial distribution belongs to an exponential family of distributions with

$$a(y_j) = y_j,$$

$$b(\mu_j) = ln\left(\frac{k\mu_j}{1+k\mu_j}\right),$$

$$g(\mu_j) = -\frac{1}{k} ln\left(1+k\mu_j\right),$$

$$\text{and } d(y_j) = ln\left(\frac{\Gamma(y_j + k^{-1})}{y_j!\Gamma(k^{-1})}\right).$$

Since $a(y_j) = y_j$, the Negative Binomial distribution is in canonical form with natural parameter

$$b(\mu_j) = ln\left(\frac{k\mu_j}{1+k\mu_j}\right).$$

Note that this natural parameter ranges between $-\infty and 0$. It can not be used as a link function to construct the GLM. Because the range of link function in a GLM should be between $-\infty and +\infty$ as it is equated to the linear predictor, $\eta_j = \boldsymbol{x_j}^\top \boldsymbol{\beta}$ taking any value on the real line.

One may use the **log-link** function to construct the GLM for Negative Binomial response. Utilizing the log-link allows a direct comparison with the Poisson model, which is NB2 with $k = 0$ (Hilbe, 2011). That is, the natural log of the mean response, $ln\mu_j$ can be used as a link function, $g(\mu_j)$ due to the following reasons:

- $ln\mu_j$ is a function of mean response, μ_j;
- $ln\mu_j$ is a monotonic differentiable function of μ_j ; and
- The range of $ln\mu_j$ is between $-\infty and +\infty$ as $\mu_j > 0$.

(b) The GLM for Negative Binomial response can be written as

$$g(\mu_j) = \eta_j = \boldsymbol{x}_j^\top \boldsymbol{\beta}$$
$$\Rightarrow ln\mu_j = \boldsymbol{x}_j^\top \boldsymbol{\beta},$$

which is called Negative Binomial count model with log link function. For this model, it can be shown that

$$\mu_j = e^{\eta_j},$$

where $\eta_j = \boldsymbol{x}_j^\top \boldsymbol{\beta}$.

2.6.3 Estimation of parameter vector, $\boldsymbol{\theta}$

Under a Negative Binomial(NB2) count GLM, our main interest is to estimate the regression parameters along with dispersion parameter using the maximum likelihood approach which is briefly described below:

Likelihood function

Let $\boldsymbol{Y} = (Y_1, \cdot\cdot, Y_j, \cdot\cdot, Y_n)^t op$ be a random sample of size n from a Negative Binomial (NB2) distribution function, $f(y; \mu, k)$, where k is the dispersion parameter. Also, suppose that $\boldsymbol{x}_j = (x_{j1}, \cdot\cdot\cdot, x_{jk}, \cdot\cdot\cdot, x_{jp})^\top$ is the p×1 vector of covariates associated with j^{th} response, Y_j and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdot\cdot\cdot, \boldsymbol{\beta}_k, \cdot\cdot\cdot, \boldsymbol{\beta}_q)^\top$ is the q×1 vector of regression coefficients. Note that the mean response, μ_i is a function of regression parameter, $\boldsymbol{\beta}$. Let us define the parameter vector, $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, k)^\top$ which is to be estimated using the maximum likelihood approach.

The likelihood function is a function of statistical model parameters that is important in statistical inference, particularly in parameter estimation of interest. Thus, the likelihood function of NB2 distribution is given by

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{j=1}^{n} f(y_j; \mu_j, k) \tag{2.8}$$

$$= \prod_{j=1}^{n} exp\left[y_j ln\left(\frac{k\mu_j}{1+k\mu_j}\right) - \frac{1}{k} ln\left(1+k\mu_j\right) + ln\left(\frac{\Gamma(y_j + k^{-1})}{y_j! \Gamma(k^{-1})}\right)\right], \tag{2.9}$$

$$\tag{2.10}$$

where $\mu_j = exp\left(\boldsymbol{x_j}^\top \boldsymbol{\beta}\right)$.

Thus, the log-likelihood function, denoted by $l(\boldsymbol{\theta};\boldsymbol{y})$, is given by

$$l(\boldsymbol{\theta};\boldsymbol{y}) = lnL(\boldsymbol{\theta};\boldsymbol{y}) \tag{2.11}$$

$$= \sum_{j=1}^{n}\left[y_j ln\left(\frac{k\mu_j}{1+k\mu_j}\right) - \frac{1}{k}ln\left(1+k\mu_j\right) + ln\left(\frac{\Gamma(y_i+k^{-1})}{y_j!\Gamma(k^{-1})}\right)\right]. \tag{2.12}$$

Score function

The score function, denoted by $U(\boldsymbol{\theta})$, is a $(p+1)\times 1$ vector of score components which can be defined as

$$U(\boldsymbol{\theta})_{(p+1)\times 1} = \frac{\delta}{\delta\theta}\, l(\boldsymbol{\theta};\boldsymbol{y}) \;;\; \text{where } \boldsymbol{\theta} = (\boldsymbol{\beta}^\top, k)^\top$$

$$= \left[U_1(\boldsymbol{\theta}), U_2(\boldsymbol{\theta}), \cdots, U_g(\boldsymbol{\theta}), \cdots, U_p(\boldsymbol{\theta}), U_{p+1}(\boldsymbol{\theta})\right]'.$$

The g^{th} $(g = 1, 2, \cdots\cdots, p)$ element of score function can be obtained as

$$
\begin{aligned}
U_g(\boldsymbol{\theta}) &= \frac{\delta}{\delta\beta_g} l(\boldsymbol{\theta}; \boldsymbol{y}) \\
&= \frac{\delta}{\delta\beta_g} \sum_{j=1}^{n} \left[y_j ln \left(\frac{k\mu_j}{1+k\mu_j} \right) - \frac{1}{k} ln\,(1+k\mu_j) + ln \left(\frac{\Gamma(y_j + k^{-1})}{y_j! \Gamma(k^{-1})} \right) \right] \\
&= \sum_{j=1}^{n} \left[y_j \left(\frac{1+k\mu_j}{k\mu_j} \right) \ \frac{(1+k\mu_j)\ k\mu_j\ x_{jg} - k\mu_j\ k\mu_j\ x_{jg}}{(1+k\mu_j)^2} - \frac{k\ \mu_j\ x_{jg}}{k(1+k\mu_j)} \right] \\
&= \sum_{j=1}^{n} \left[y_j\ \frac{x_{jg}}{1+k\mu_j} - \frac{\mu_j\ x_{jg}}{1+k\mu_j} \right] \\
&= \sum_{j=1}^{n} \frac{x_{jg}(y_j - \mu_j)}{(1+k\mu_j)}\ ; \quad g = 1, 2, \cdots\cdots, p. \qquad (2.13)
\end{aligned}
$$

$$\text{and } U_{p+1} = \frac{\delta}{\delta g}\, l(\boldsymbol{\theta}; \boldsymbol{y})$$

$$= \frac{\delta}{\delta k} \sum_{j=1}^{n} \left[y_j ln \left(\frac{k\mu_j}{1+k\mu_j} \right) - \frac{1}{k} ln\left(1+k\mu_j\right) + ln\Gamma(y_j + k^{-1}) - ln\Gamma(y_j+1) - ln\Gamma(k^{-1}) \right.$$

$$= \sum_{j=1}^{n} \left[\frac{y_j\ k\mu_j}{(1+k\mu_j)}\ \frac{\mu_j}{(k\mu_j)^2} + \frac{1}{k^2} ln\left(1+k\mu_j\right) - \frac{\mu_j}{k(1+k\mu_j)} + \Psi\left(y_j + k^{-1}\right) - \Psi\left(\frac{1}{k}\right) \right]$$

$$= \sum_{j=1}^{n} \left[\frac{y_j}{k(1+k\mu_j)} - \frac{\mu_j}{c(1+k\mu_j)} + \frac{1}{k^2} ln\left(1+k\mu_j\right) + \Psi\left(y_j + k^{-1}\right) - \Psi\left(\frac{1}{k}\right) \right]$$

$$= \sum_{j=1}^{n} \left[\frac{1}{k^2} \left\{ ln\left(1+k\mu_j\right) + \frac{k(y_j-\mu_j)}{(1+k\mu_j)} \right\} + \Psi\left(y_j + k^{-1}\right) - \Psi\left(\frac{1}{k}\right) \right], \quad (2.14)$$

where $\Psi(\cdot)$ is called the digamma function which is the first derivative of the log-gamma function, $ln\Gamma(\cdot)$.

Information matrix

The information matrix, denoted by $I(\boldsymbol{\theta})$, is the variance covariance matrix of the score function $U(\boldsymbol{\theta})$ which is a $(p+1) \times (p+1)$ matrix given by

$$I(\boldsymbol{\theta}) = Var\left[U(\boldsymbol{\theta})\right] = \left[I_{ll'}(\boldsymbol{\theta})\right]; \quad l = 1, 2, \cdots, (p+1)\ ; \quad l' = 1, 2, \cdots, (p+1).$$

The observed information matrix, $I^*(\boldsymbol{\theta})$ is defined as the negative of the second partial derivative of the log-likelihood with respect to the parameters, while the expected information matrix, $I(\boldsymbol{\theta})$, is the expected value of the observed information matrix.

That is,

$$I^*(\boldsymbol{\theta}) = -\frac{\delta^2}{\delta\boldsymbol{\theta}\delta\boldsymbol{\theta}'}\, l(\boldsymbol{\theta}, y)$$

$$\text{and } I(\boldsymbol{\theta}) = E\left[I^*(\boldsymbol{\theta})\right].$$

It can be shown that

$$I^*(\boldsymbol{\theta})_{(p+1)\times(p+1)} = \begin{bmatrix} I^*_{11}(\boldsymbol{\theta}) & I^*_{12}(\boldsymbol{\theta}) \\ \\ I^*_{21}(\boldsymbol{\theta}) & I^*_{22}(\boldsymbol{\theta}) \\ \\ \end{bmatrix},$$

where $I^*_{11}(\boldsymbol{\theta}) = -\dfrac{\delta^2 l}{\delta\boldsymbol{\beta}\delta\boldsymbol{\beta}'}$ is a $p \times p$ matrix, $I^*_{12}(\boldsymbol{\theta}) = I^*_{21}(\boldsymbol{\theta}) = -\dfrac{\delta^2 l}{\delta\boldsymbol{\beta}\delta k}$ is a $p \times 1$ vector and $I^*_{22}(\boldsymbol{\theta}) = -\dfrac{\delta^2 l}{(\delta k)^2}$ is a scalar component. Now, using equation (3.7), the $(g, g')^{th}$ element of $I^*_{11}(\boldsymbol{\theta})$ can be defined as

$$I^*_{11(gg')}(\boldsymbol{\theta}) = -\frac{\delta^2 l}{\delta\boldsymbol{\beta}_g\, \delta\boldsymbol{\beta}_{g'}} = -\frac{\delta}{\delta\boldsymbol{\beta}_{g'}}\, U_g(\boldsymbol{\theta})\ ; \quad g = 1, 2, \cdots, p \text{ and } g' = 1, 2, \cdots, p \tag{2.15}$$

$$= -\frac{\delta}{\delta\boldsymbol{\beta}_{g'}}\left[\sum_{j=1}^{n} \frac{x_{jg}(y_j - \mu_j)}{(1 + k\mu_j)}\right] \tag{2.16}$$

$$= - \sum_{j=1}^{n} x_{jg} \left[\frac{(1+k\mu_j)(-\mu_j \; x_{jg}) - (y_j - \mu_j) \; c\mu_j \; x_{jg'}}{(1+k\mu_j)^2} \right]$$

$$= \sum_{j=1}^{n} x_{jg} \left[\frac{(1+ky_j) \; \mu_j \; x_{jg'}}{(1+k\mu_j)^2} \right] \tag{2.17}$$

$$\therefore \; I_{11}^{*}(\boldsymbol{\theta}) = \sum_{j=1}^{n} x_j \left[\frac{\mu_j \; (1+ky_j)}{(1+k\mu_j)^2} \right] x_j'. \tag{2.18}$$

$$\tag{2.19}$$

Again, using equation (3.8), the g^{th} element of $I^*_{12}(\boldsymbol{\theta}) = I^*_{21}(\boldsymbol{\theta})$ can be defined as

$$I^*_{12(g)}(\boldsymbol{\theta}) = -\frac{\delta^2 l}{\delta \boldsymbol{\beta}_j \, \delta k} = -\frac{\delta}{\delta \boldsymbol{\beta}_g} U_{p+1}(\boldsymbol{\theta}) \; ; \quad g = 1, 2, \cdots, p$$

(2.20)

$$= -\frac{\delta}{\delta \boldsymbol{\beta}_g} \sum_{j=1}^{n} \left[\frac{1}{k^2} \left\{ ln\left(1 + k\mu_j\right) + \frac{k(y_j - \mu_j)}{(1 + k\mu_j)} \right\} + \Psi\left(y_j + k^{-1}\right) - \Psi\left(\frac{1}{k}\right) \right]$$

(2.21)

$$= -\sum_{j=1}^{n} \frac{1}{k^2} \left[\frac{k\mu_j \; x_{jg}}{(1 + c\mu_j)} + \frac{k(1 + k\mu_j)(-\mu_j x_{jg}) - k(y_j - \mu_j) k\mu_j \; x_{jg}}{(1 + k\mu_j)^2} \right]$$

(2.22)

$$= -\sum_{j=1}^{n} \frac{\mu_j \; x_{jg}}{k} \left[\frac{1}{1 + k\mu_j} - \frac{1 + ky_j}{(1 + k\mu_j)^2} \right]$$

(2.23)

$$= \sum_{j=1}^{n} \frac{\mu_j \; (y_j - \mu_j) \; x_{jg}}{(1 + k\mu_j)^2}.$$

(2.24)

Therefore, $I^*_{12}(\boldsymbol{\theta}) = I^*_{21}(\boldsymbol{\theta}) = \sum_{j=1}^{n} \frac{\mu_j \; (y_j - \mu_j) \; x_j}{(1 + k\mu_j)^2}.$ (2.25)

Finally, the scalar component of the observed information matrix, $I^*(\boldsymbol{\theta})$ denoted by $I^*_{22}(\boldsymbol{\theta})$ can be defined, using equation (3.8), as

$$I^*_{22}(\boldsymbol{\theta}) = -\frac{\delta^2 l}{(\delta k)^2} = -\frac{\delta}{\delta k}\, U_{(p+1)}(\boldsymbol{\theta})$$

(2.26)

$$= -\frac{\delta}{\delta k}\sum_{j=1}^{n}\left[\frac{1}{k^2}\left\{ln\left(1+k\mu_j\right)+\frac{k(y_j-\mu_j)}{(1+k\mu_j)}\right\}+\Psi\left(y_j+k^{-1}\right)-\Psi\left(\frac{1}{k}\right)\right]$$

(2.27)

$$= -\sum_{j=1}^{n}\left[\frac{1}{k^2}\left\{\frac{\mu_j}{1+k\mu_j}+\frac{(1+k\mu_j)(y_j-\mu_j)-k(y_j-\mu_j)\mu_j}{(1+k\mu_j)^2}\right\}\right.$$
$$\left.-\frac{2}{k^3}\left\{ln\left(1+k\mu_j\right)+\frac{k(y_j-\mu_j)}{(1+k\mu_j)}\right\}+\Psi'\left(y_j+k^{-1}\right)-\Psi'\left(\frac{1}{k}\right)\right]$$

(2.28)

$$= -\sum_{j=1}^{n}\left[\frac{1}{k^2}\left\{\frac{\mu_j(1+k\mu_j)+(y_j-\mu_j)}{(1+k\mu_j)^2}\right\}-\frac{2}{k^3}\left\{ln\left(1+k\mu_j\right)+\frac{k(y_j-\mu_j)}{(1+k\mu_j)}\right\}\right.$$
$$\left.+\Psi'\left(y_j+k^{-1}\right)-\Psi'\left(\frac{1}{k}\right)\right]$$

(2.29)

$$= -\sum_{j=1}^{n}\left[-\frac{1}{k^3}\left\{2ln\left(1+k\mu_j\right)+\frac{2k(y_j-\mu_j)}{1+k\mu_j}-\frac{k\mu_j(1+k\mu_j)+k(y_j-\mu_j)}{(1+k\mu_j)^2}\right\}\right.$$
$$\left.+\Psi'\left(y_j+k^{-1}\right)-\Psi'\left(\frac{1}{k}\right)\right]$$

(2.30)

$$= -\sum_{j=1}^{n}\left[-\frac{1}{k^3}\left\{2ln\left(1+k\mu_j\right)+\frac{2k(y_j-\mu_j)(1+k\mu_j)-k\mu_j(1+k\mu_j)-k(y_j-\mu_j)}{(1+k\mu_j)^2}\right\}\right.$$
$$\left.+\ \Psi'\left(y_j+k^{-1}\right)-\Psi'\left(\frac{1}{k}\right)\right]$$

(2.31)

$$= -\sum_{j=1}^{n}\left[-\frac{1}{k^3}\left\{2ln\left(1+k\mu_j\right)+\frac{(y_j-\mu_j)(2k+2k^2\mu_j-k)-k\mu_j(1+k\mu_j)}{(1+k\mu_j)^2}\right\}\right.$$
$$\left.+\ \Psi'\left(y_j+k^{-1}\right)-\Psi'\left(\frac{1}{k}\right)\right]$$

(2.32)

$$= -\sum_{j=1}^{n}\left[-\frac{1}{k^3}\left\{2ln\left(1+k\mu_j\right)+\frac{(y_j-\mu_j)k(1+2k\mu_j)-k\mu_j(1+k\mu_j)}{(1+k\mu_j)^2}\right\}\right.$$
$$\left.+\ \Psi'\left(y_j+k^{-1}\right)-\Psi'\left(\frac{1}{k}\right)\right]$$

(2.33)

$$= \sum_{j=1}^{n}\left[\frac{1}{k^3}\left\{\frac{k(1+2k\mu_j)(y_j-\mu_j)-k\mu_j(1+k\mu_j)}{(1+k\mu_j)^2}+2ln(1+k\mu_j)\right\}\right.$$
$$\left.-\ \Psi'\left(y_j+k^{-1}\right)+\Psi'\left(\frac{1}{k}\right)\right]. \tag{2.34}$$

Maximum likelihood estimating equation

The maximum likelihood estimating equation for the parameter vector, $\boldsymbol{\theta}$ can be defined by equating the score function, $U(\boldsymbol{\theta})$ to zero as follows:

$$U\left(\boldsymbol{\theta}\right) = \mathbf{0}. \tag{2.35}$$

This equation can easily be solved for $\boldsymbol{\theta}$ by Newton Raphson iterative procedure.The estimate obtained at the $r^{th}(r = 1, 2, \cdots)$ iteration is given by

$$\widehat{\boldsymbol{\theta}}^{(r)} = \widehat{\boldsymbol{\theta}}^{(r-1)} + \left[I(\boldsymbol{\theta})\right]^{-1}_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(r-1)}} U\left(\boldsymbol{\theta}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(r-1)}}, \tag{2.36}$$

where $I(\boldsymbol{\theta})$ is the expected information matrix defined as

$$\begin{aligned} I(\boldsymbol{\theta}) &= -E\left[\frac{\delta}{\delta\boldsymbol{\theta}}\, U'(\boldsymbol{\theta})\right] \\ &= \left\{I_{ll'}\right\}_{l=1,2,\cdots,(p+1);\ l'=1,2,\cdots,(p+1)}. \end{aligned}$$

Note that the maximum likelihood estimator, $\widehat{\boldsymbol{\theta}}$ is asymptotically distributed as Normal with mean vector $\boldsymbol{\theta}$ and variance-covariance matrix $\left[I(\boldsymbol{\theta})\right]^{-1}$, i.e.

$$\widehat{\boldsymbol{\theta}}_{(p+1)\times 1} \sim N_{p+1}\left[\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1}\right] \quad as \quad n \to \infty.$$

Therefore, the distribution of $\widehat{\boldsymbol{\beta}}_g$ $(g = 1, 2, \cdots, p)$ and $\hat{k}$ are

$$\widehat{\boldsymbol{\beta}}_g \sim N\left[\boldsymbol{\beta}_g, I^{gg}(\theta)\right] \quad as \quad n \to \infty; \quad j = 1, 2, \cdots, p$$

$$\text{and } \widehat{k} \sim N\left[k, I^{(p+1),(p+1)}(\boldsymbol{\theta})\right] \quad as \quad n \to \infty,$$

where $I^{gg}(\boldsymbol{\beta})$ and $I^{(p+1),(p+1)}(\boldsymbol{\theta})$ are the $(g,g)^{th}$ and $((p+1),(p+1))^{th}$ element of $[I(\boldsymbol{\theta})]^{-1}$ respectively.

2.7 Method of measuring variable importance in regression analysis

In regression analysis researcher's commonly face the situation of relative importance of the variable. Similar to the multiple linear regressions, researchers may be interested to find out the most important predictor among the several predictors in a Poisson regression model. There exists several method for measuring relative importance among the independent variable such that 1. Standardized regression coefficients. 2. Orthogonal counterparts measure of Gibson (1962) and R.M. Johnson (1966). 3. Relative Weights (RW) measure of J. W. Johnson (2000). 4. Average sequential R^2 overall possible ordering suggested by Lindman, Merenda and Gold (1980) . 5. Dominance analysis of Azen and Brudescu.

Dominance analysis technique has been widely used by researchers to examine the predictor importance more accurately in the linear regression. R^2 for Poisson Model based on Pearson Residuals :

$$R^2 = 1 - \sum_{i}^{N} \frac{\left(y_i - \hat{\mu}_i\right)^2 / \hat{\mu}_i^2}{\left(y_i - \bar{y}\right)^2 / \bar{y}^2} \tag{2.37}$$

Dominance analysis is the method that involves examining the R^2 values for all subset, is refined by incorporating numerous quantitative measures of dominance, i.e,

the enhance in R^2 from introducing a different variable to the model which contains other predictors. For example, the involvement of X_a to the model of subsets X_c is $R^2_{Y.X_aX_c} - R^2_{Y.X_c}$ where, $R^2_{Y.X_aX_c}$ is the value of R^2 when Y is regressed on X_a and X_c, while $Y.{X_c}^2$ is the R^2 regressed on X_c.

The standard summary statistics for a regression analysis include an estimate of the , the standard error and regression coefficients coefficients, and the t- statistics, as well as the p- values corresponding to the t- statistics.

Dominance analysis technique has been widely used by researchers to examine the predictors importance in the linear regression. Dominance analysis was suggested by Budescu (Budescu, 1993) and was extended and modified by Azen and Budescu (Azen & Budescu, 2003). Recently Azen and Traxel used dominance analysis to logistic regression analysis to rank order the predictors (Azen & Traxel, 2009).
Complete dominance analysis is difficult to carried out and rarely used. Conditional dominance is comparatively weaker than complete dominance but the same thing happened for conditional dominance analysis which is also difficult to carried out.

The strongest dominance is complete dominance where if additional contribution X_a is greater than X_b for all subsets model, then X_a dominates X_b then X_b completely dominates X_c then in accordance with Brudescu X_a will dominates X_c. Complete dominates has some problem, i.e, X_a dominates X_c but sometimes X_c dominates X_b. That means reverse can be happened. If average additional contribution of X_b is greater than X_c among each model , then X_b conditionally dominates X_c. Just Like complete dominance, conditional dominance often does not exist for all pairs of predictors. The same

thing happen for conditional dominance. Averaging all conditional statistics yields the weakest kinds of dominance, known as general dominance.The model R^2 is equal to the sum of general dominance over all predictor variables.

Table 2.1: Dominanc analysis for several variable

Subset model (X)	**Submodel size (l)**	**Method**		
		X_a	X_b	X_c
Null	0	$R^2_{Y.X_a}$	$R^2_{Y.X_b}$	$R^2_{Y.X_c}$
Average (CD_1)		$R^2_{Y.X_a}$	$R^2_{Y.X_b}$	$R^2_{Y.X_c}$
X_a	1	-	$R^2_{Y.X_aX_b}$- $R^2_{Y.X_a}$	$R^2_{Y.X_aX_c}$ - $R^2_{Y.X_a}$
X_b	1	$R^2_{Y.X_aX_b} - R^2_{Y.X_b}$	-	$R^2_{Y.X_bX_c} - R^2_{Y.X_b}$
X_c	1	$R^2_{Y.X_aX_c} - R^2_{Y.X_c}$	$R^2_{Y.X_bX_c} - R^2_{Y.X_c}$	-
Average (CD_2)		$\frac{-R^2_{Y.X_aX_b}-R^2_{Y.X_b}+R^2_{Y.X_aX_c}-R^2_{Y.X_c}}{2}$	$\frac{-R^2_{Y.X_aX_b}-R^2_{Y.X_a}+R^2_{Y.X_bX_c}-R^2_{Y.X_c}}{2}$	$\frac{-R^2_{Y.X_aX_c}-R^2_{Y.X_a}+R^2_{Y.X_bX_c}-R^2_{Y.X_b}}{2}$
X_aX_b	2	-	-	$R^2_{Y.X_aX_bX_c} - R^2_{Y.X_aX_b}$
X_aX_c	2	-	$R^2_{Y.X_aX_bX_c} - R^2_{Y.X_aX_c}$	-
X_bX_c	2	$R^2_{Y.X_aX_bX_c} - R^2_{Y.X_bX_c}$	-	-
Average (CD_2)		$R^2_{Y.X_aX_bX_c} - R^2_{Y.X_bX_c}$	$R^2_{Y.X_aX_bX_c} - R^2_{Y.X_aX_c}$	$R^2_{Y.X_aX_bX_c} - R^2_{Y.X_aX_b}$
Overall (GD)		$GD(X_a)$	$GD(X_b)$	$GD(X_c)$

where,

$$GD(X_a) = \frac{-2R^2_{Y.X_a}-R^2_{Y.X_b}-R^2_{Y.X_c}+R^2_{Y.X_aX_b}+R^2_{Y.X_aX_c}-2R^2_{Y.X_bX_c}+2R^2_{Y.X_aX_bX_c}}{6}$$

$$GD(X_b) = \frac{-R^2_{Y.X_a}+2R^2_{Y.X_b}-R^2_{Y.X_c}+R^2_{Y.X_aX_b}-2R^2_{Y.X_aX_c}+R^2_{Y.X_bX_c}+2R^2_{Y.X_aX_bX_c}}{6}$$

$$GD(X_c) = \frac{-R^2_{Y.X_a}-R^2_{Y.X_b}+2R^2_{Y.X_c}-2R^2_{Y.X_aX_b}+R^2_{Y.X_aX_c}+R^2_{Y.X_bX_c}+2R^2_{Y.X_aX_bX_c}}{6}$$

General dominance analysis is the weakest among three but it is easily handled out. For this reason, we use general dominance analysis. General dominance analysis is same as J. Lindman, Merenda and Gold (1980) method. For reducing complication we use J. Lindman, Merenda and Gold (1980) method.

Chapter 3

Data and variable

3.1 Source of Data

The Bangladesh Demographic and Health Survey (BDHS) is a part of the worldwide Demographic and Health Survey (DHS) program, which is planned to take out data on basic nationwide demonstrators of social prosperity including family planning, fertility and maternal and child health.

The Bangladesh Demographic and Health Survey (BDHS), 2017-18 is the national level Demographic and Health Survey (DHS), which is covered a household survey of ever married women aged 15-49. It is to be mention that some articles have already been published by using this data set (Hossain et al., 2020). The (BDHS) is conducted in Bangladesh under the dominance of the National Institute of Population Research and training (NIPORT) of the ministry of Health and Family Welfare and accomplished by M itra and Associates of Dhaka.

The primary motives of the BDHS, 2017-18 are to give up-to-date information on fertility preferences; approval, awareness, use of family planning methods and fertility and childhood mortality levels; maternal and child health which are included breastfeeding

practices, newborn care and nutrition levels; community level dataon accessibility and availability of health and family planning services; and knowledge and behaviors unto HIV/AIDS. The Maryland, the Inner City Fund (ICF) International of Rockville, the United State Agency (USA) provided the technical assistance to perform the survey. The United State Agency (USA) also provided the financial support for the survey.

There are seven administrative divisions in Bangladesh.The sampling frame contains details on the EA's location, type of residence and the expected amount of residential households.This design resulted in the selection of 20,250 residential households. After removing three clusters (one urban and two rural) that had been completely eroded by floodwater, the survey was completed in 672 clusters.

Any analysis based on the 2017-18 BDHS data requires the use of sampling weights to ensure that the survey results are accurately represented at the national and divisional levels.The BDHS sampling weights for 2017-18 are not expected to cause any significant differences in the overall survey indicators.

3.2 Variables

The response variable of interest is the number of antenatal care (ANC) visits.

The covariates considers in this study are:

1. Place of residence
2. Region of mother
3. Mother's educational level
4. Wealth index
5. Birth order
6. Mothers age at birth
7. Media exposure
8. Decision making
9. Wanted pregnancy
10. Problem of getting health facility

In our covariates some variables are directly obtained from the BDHS data. But some of them has to modified. Media exposure variable modified from three variable (Newspaper, magazine, radio). If respondents use any of the component for ANC visit then she is exposed by media, otherwise not.

Bangladesh is divided into three region. These are South (Barisal, Chittagong), Central (Dhaka, Mymensingh, Sylhet) and North (Rajsahi, Rangpur).

In decision making variable where respondent is involved in making decision that category is denoted as respondent category and other categoery is denoted as others.

Age variable is divided into three category where less than 20 is one and other two are age between 20 to 35 and greater than 35.

Table 3.1: Variables and their categories

Variables	Categories
Place of residence	1 = Urban
	2 = Rural
Region	1 = South
	2 = Central
	3 = North
Mother's educational level	0 = No education
	1 = Primary
	2 = Secondary
	3 = Higher
Wealth index	0 = Poor
	1 = Middle
	2 = Rich
Birth order	1 = Ist birth
	2 = 2nd and 3rd birth
	3 = Above 3rd birth
Mothers age at birth	1 = <20
	2 = 20-35
	3 = >35
Media exposure	1 = Exposed
	0 = Non-exposed
Decision making	1 = Respondents
	0 = Other
Wanted pregnancy	1 = Yes
	0 = No
Problem of getting health facility	1 =Problem
	0 =No problem

Chapter 4

Results

4.1 Univariate analysis

For the purposes of explanatory analysis we consider the sample characteristics of the covariates, i.e., univariate frequency distribution and corresponding percentage distribution of the category of the covariates. The interpretation is given below according to the covariates.

Number of ANC visits are represented in the figure below:

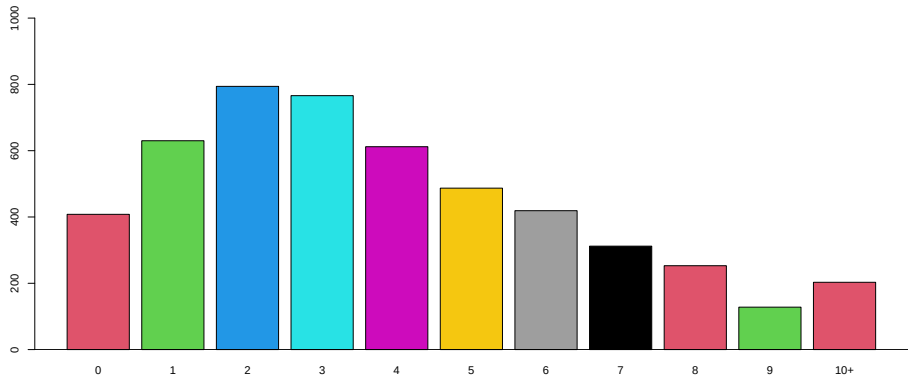

Figure 4.1: Bar diagram of number of ANC visits

Region variable is split into two groups urban and rural. More than half of the respondents (65.6%) are rural, with the remainder (34.4%) being urban. Bangladesh is divided into three region. These are South (Barisal, Chittagong), Central (Dhaka, Mymensingh, Sylhet) and North (Rajsahi, Rangpur). From the table, it is oveserved that out of 42.1% mothers are from south while 36.3%, 21.7% are from central and north. Here, respondents education are categorized into four category (no education, primary, secondary and higher). Most (47%) of the mother's educational status is secondary while 6.2% mother's educational status is no education, 27.8% status are primary. Poor and rich family's mothers share almost equal percentage(40%). Middle family have only 18.1%. 38.2% mothers give 1st birth while 49.7% for 2nd and 3rd birth and 12.1% for above 3rd birth. From the table it is observed that 28.2% mothers age are less than 20 while 68% for between 20 to 35 ages and 3.8% for above 35 years. Exposure of media variable is categorized into yes and no, where if any mother read newspaper, magazine or listening radio, they are categorized as yes,otherwise no. 64.3% respondents are exposed by media and 35.7% respondents are non-exposure. In this study, it is also observed that 80.8% of mothers takes part in deciding on their health care and 28% are not involve any decision making. 78.9% mothers did not want their current child and only 21% mother wanted their current child. For getting health facilities for self,distance was a problem for 40.7% mothers and for 59.3% mother ,distance was not a problem.

Table 4.1: Frequency distribution and corresponding percentage distribution of the categories of the covariates

Covariate	Frequency	Percentage
Place of residence		
urban	1725	34.4
rural	3287	65.6
Region		
South	2109	42.1
Central	1817	36.3
North	1086	21.7
Mother's education level		
No education	312	6.2
Primary	1392	27.8
Secondary	2402	47.9
Higher	906	18.1
Wealth index		
Poor	2096	41.8
Middle	905	18.1
Rich	2011	40.1
Birth order		
Ist birth	1915	38.2
2nd and 3rd birth	2493	49.7
Above 3rd birth	604	12.1
Mothers age at birth		
<20	1413	28.2
20-35	3407	68.0
>35	192	3.8
Media exposure		
Non-exposed	1788	35.7
Exposed	3224	64.3
Decision making		
Other	1402	28
Respondent	3610	72
Wanted pregnancy		
No	1058	21.1
Yes	3954	78.9
Problem of getting health facility		
No problem	2972	59.3
Problem	2040	40.7

4.2 Bivariate analysis

The table represents the chi-square test and the p-value was used to determine whether a specific explanatory variable was remarkably associated with the number of antenatal care visits. It is clear from the table that all explanatory variables are found to be significantly connected with the number of antenatal care visits except decision making. As for all cases p-value is very small (<0.001). So, there is significant relation between the number of ANC visits and Covariates. But in case of decision making p-value is greater than 0.05. decision making variable is insignificantly connection with the number of ANC visits.

Table 4.2: Bivariate associations between the number of antenatal care visits and different predictors in Bangladesh.
.

Characteristics	χ^2	P-value
Place of residence	7195.263	<0.001
Region	75.257	<0.001
Mother's educational level	688.049	<0.001
Wealth index	525.587	0.009
Birth order	253.993	<0.001
Mothers age at birth	61.615	0.005
Media exposure	429.93	<0.001
Decision making	27.651	0.068
Wanted pregnancy	61.626	<0.001
Problem of getting health facility	105.861	<0.001

4.3 Application of Poisson and Negative binomial (NB) regression to the number of ANC visits

We have two different count regression models, namely Poisson and Negative binomial regression models for the number of ANC visits in Bangladesh.

4.3.1 Poisson regression model

Table 6.3 presents the estimates of Poisson regression model along with p-values. This table also displays the incidence rate ratio (IRR) of each categories of the covariates :

Table 4.3: Estimates, p-values and incidence rate ratio (IRR) of Poisson regression model for the determinants of number of ANC visits.

.

Covariate	Estimate	p-value	IRR
Intercept	0.666	<0.001	
Place of residence			
Rural	-0.141	<0.001	0.86
Urban (ref)	-	-	-
Region			
Central	0.076	<0.001	1.07
North	0.169	<0.001	1.18
South (ref)	-	-	-
Mother's education level			
Primary	0.260	<0.001	1.29
Secondary	0.432	<0.001	1.54
Higher	0.555	<0.001	1.74
No education (ref)	-	-	-
Wealth index			
Middle	0.101	<0.001	1.10
Rich	0.179	<0.001	1.19
Poor (ref)	-	-	-
Birth order			
2nd and 3rd birth	-0.069	<0.001	0.93
Above 3rd birth	-0.240	<0.001	0.78
1st birth (ref)	-	-	-
Mother's age at birth			
<20	0.095	<0.001	1.10
20-35	0.121	<0.001	1.12
>35	-	-	-
Media exposure			
Exposed	0.223	<0.001	1.25
Non-exposed (ref)	-	-	-
Wanted pregnancy			
Yes	0.100	<0.001	1.10
No (ref)	-	-	-
Problem of getting health facility			
Problem	-0.049	<0.001	0.95
No problem (ref)	-	-	-

All the covariates have significant influence on the average number of ANC visits The likelihood of number of ANC visits of rural women were 14% (p-value=<0.001) less than those of urban women and it is found to be significant. Women who live in central region had 1.07 times as likely to have mean number of ANC visits as women who live in south region. Also, women of north region had 18% more IRR of having average number of ANC visits than those were from south region. Respondents with primary education have 29% (IRR=1.29) higher ANC visits than the respondents with no education. Respondents with secondary education have 54% (IRR=1.54) higher ANC visits than the respondents with no education. Respondents with higher education have 74% (IRR=1.74) higher than the respondents with no education. All the education status have significant effect (p-value=<0.001) on the number of ANC visits. On an average, the number of having ANC visits is 10% (IRR=1.10) higher for women whose wealth index status is middle compared to whose status is poor. The average number of ANC visits is 19% (IRR=1.19) higher for women whose wealth index status is rich compared to whose status is poor. The incidence rate for having ANC visits is .93 times lower for women at their 2nd and 3rd birth than the women at their 1st birth and this is higly significant.The incidence rate for having ANC visits is .78 times lower for women at their above 3rd birth than women at their 1st birth and this is also higly significant. When mother's age at their current birth was less than 20 have 10% (IRR=1.10) higher ANC visits compared to those whose age was greater than 35. When mother's age at their current birth was between 20 and 30 have 12% (IRR=1.12) higher compared to those whose age was greater than 35. The excess to media is 1.25 times higher for having ANC visits than the women who are not exposed by media. The ANC visits by mothers who want their current child are 1.10 times higher for ANC visits than the mother who do not want their current child. The effect is significant. The expected

number of women for getting health facility distance was a problem is 5% (IRR=.95) lower than the women where distance was not a problem for getting medical help.

4.3.2 Overdispersion tests

For detecting overdispersion, there are some common statistical test such as Z-score, Lagrange multiplier tests. Using Poisson regression model result are concluded below.

Score test

Table 6.4 presents the results of score test for detecting overdispersion to the number of ANC visits in Bangladesh. The p-value indicate that the data has overdispersion (p-value=<0.001) and hence we should use Negative Binomial model to identify the determinants of the number of ANC visits in Bangladesh.

Table 4.4: Score tets detection overdispersion to the number of ANC visits in Bangladesh

Methods	Z-score	S.E	p-value
Dean and Lawless	2.356	0.33	<0.001
Winkelman	1.666	0.023	<0.001
Cameron and Trivedi	3.332	0.04	<0.001

Lagrange mutiplier test

Lagrange multiplier test statistics is 24793.99 and the corresponding p-value is 0.000. So, the data has significant overdispersion.

4.3.3 Negative binomial regression model

Table 6.5 presents the estimates of Negative Binomial regression model along with p-values. This table also displays the incidence rate ratio (IRR) of each categories of the covariates. All the covariates have significant influence on the average number of ANC visits The likelihood of number of ANC visits of rural women were 14% (p-value=<0.001) less than those of urban women and it is found to be significant. Women who live in central region had 1.08 times as likely to have mean number of ANC visits as women who live in south region. Also, women of north region had 20% more IRR of having average number of ANC visits than those were from south region. Respondents with primary education have 30% (IRR=1.30) higher ANC visits than the respondents with no education. Respondents with secondary education have 54% (IRR=1.54) higher ANC visits than the respondents with no education. Respondents with higher education have 75% (IRR=1.75) higher than the respondents with no education. All the education status have significant effect (p-value=<0.001) on the number of ANC visits. On an average, the number of having ANC visits is 11% (IRR=1.11) higher for women whose wealth index status is middle compared to those whose status is poor. The average number of ANC visits is 32% (IRR=1.32) higher for women whose wealth index status is rich compared to whose status is poor. The incidence rate for having ANC visits is .93 times lower for women at their 2nd and 3rd birth than women at their 1st birth and this is highly significant.The incidence rate for having ANC visits is .78 times lower for women at their above 3rd birth than women at their 1st birth and this is also higly significant. When mother's age at their current birth was less than 20 have 9% (IRR=1.09) higher ANC visits compared to those whose age was greater than 35. When mother's age at their current birth was between 20 and 30 have 12% (IRR=1.12) higher compared to those whose age was

Table 4.5: Estimates, p-values and incidence rate ratio (IRR) of Negative Binomial regression model for the determinants of number of ANC visits.

.

Covariate	Estimate	p-value	IRR
Intercept	0.649	<0.001	
Place of residence			
Rural	-0.139	<0.001	0.86
Urban (ref)	-	-	-
Region			
Central	0.085	<0.001	1.08
North	0.188	<0.001	1.20
South (ref)	-	-	-
Mother's educational level			
Primary	0.263	<0.001	1.30
Secondary	0.434	<0.001	1.54
Higher	0.560	<0.001	1.75
No education (ref)	-	-	-
Wealth index			
Middle	0.109	<0.001	1.11
Rich	0.185	<0.001	1.32
Poor (ref)	-	-	-
Birth order			
2nd and 3rd birth	-0.068	<0.001	0.93
Above 3rd birth	-0.241	<0.001	0.78
1st birth (ref)	-	-	-
Mother's age at birth			
<20	0.093	<0.001	1.09
20-35	0.120	<0.001	1.12
>35	-	-	-
Media exposure			
Exposed	0.224	<0.001	1.25
Non-exposed (ref)	-	-	-
Wanted pregnency			
Yes	0.102	<0.001	1.10
No (ref)	-	-	-
Problem of getting health facility			
Problem	-0.049	0.01	0.95
No problem (ref)	-	-	-

greater than 35. The excess to media is 1.25 times higher for having ANC visits than the women who are not exposed by media. The ANC visits by mothers who want their current child are 1.10 times higher for ANC vists than the mother who do not want their current child. The effect is significant. The expected number of women for getting health facility distance was a problem is 5% (IRR=.95) lower than the women where distance was not a problem for getting medical help.

Table 4.6: Variable importance obtained using Poisson regression model

Variable	Place of residence	Region	Mother's educational level	Wealth index	Birth order	Age of mothers during birth	Media exposure	Wanted pregnancy	Problem of getting health facility
percentages	9.67	4.27	28.78	19.90	9.35	3.12	20.39	1.78	2.69

4.3.4 Result obtained from Dominance analysis

Poisson regression

After applying dominance analysis in Poisson regression model, from table 4.6 we find that 28.78% variation for ANC visits are explained by mother's educational level. mother;s education mostly influence the dependent variable than others. That means if educational achievement of mothers can be increase, than ANC visits will also increase. Media is the second most important variable whose influence are 20.39%. Wealth index role is also noticeable than the others which is 19.90%. 9.35% variation can be explained by birth order. Age, wanted pregnancy and problem of getting health facility has very little contribution which are respectively 3.12%, 1.78%, 2.69%.

Table 4.7: Variable importance obtained using negative binomial regression model

Variable	Place of residence	Region	Mother's educational level	Wealth index	Birth order	Age of mothers during birth	Media exposure	Wanted pregnancy	Problem of getting health facility
percentages	9.09	4.76	28.80	19.72	9.35	3.12	20.39	1.78	2.69

Negative binomial regression

After applying dominance analysis in negative binomial regression model, from the table 4.7 we find that 28.80% variation for ANC visits are explained by mother's educational level. mother;s education mostly influence the dependent variable than others. That means if educational achievement of mothers can be increase, than ANC visits will also increase. Media is the second most important variable whose influence are 20.39%. Wealth index role is also noticeable than the others which is 19.72%. 9.35% variation can be explained by birth order. Age, wanted pregnancy and problem of getting health facility has very little contribution which are respectively 3.12%, 1.78%, 2.69%.

After combining two result obtained from Poisson and NB model we noticed that the two model demonstrated almost the same result.

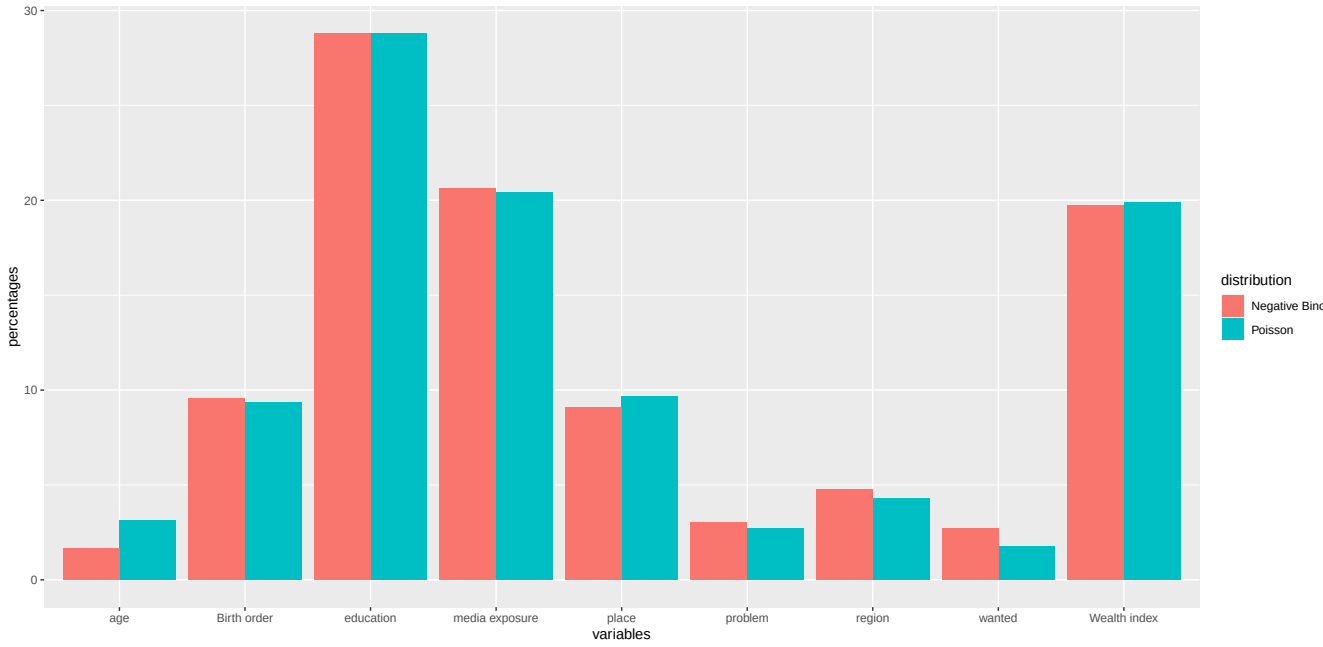

Figure 4.2: Relative importance among the predictors

Chapter 5

Conclusion and discussion

5.1 Discussion

One of the fundamental distribution for the count data is the Poisson distribution. The mean of this distribution is assumed to be equivalent to the variance. But in real life data may have overdispersion or underdispersion.

After analysis several literature we choose ten independent variable and as our target is find the variable importance for count data ,we use one dependent variable which has no category. Our required explanatory variable are: "Place of residence","Region","educational level","Wealth index","Birth order","Mothers age at birth","Media exposure","Wanted pregnancy","Problem of getting health facility". In our bivariate analysis decision making variable is found statistically insignificant.

Our study indicate that the likelihood of average number of ANC visits of rural women were less than those of urban women. This result corroborates with another study in Bangladesh conducted by (Hossain et al., 2020) and a study in Ethiopia by (Terefe &

Gelaw, 2019).

Respondents with primary, secondary and higher likelihoods of average number of ANC visits than the respondents with no education. And this study supports the previous study in Bangladesh.(Hossain et al., 2020).

Our study reflect that on an average, the number of having ANC visits is higher for women whose wealth index status is middle compared to whose status is poor. The average number of ANC visits is higher for women whose wealth index status is rich compared to whose status is poor. Another study in Ethiopia indicate that the number of having ANC visits is higher for women whose wealth index status is middle and rich compared to poor (Fenta, Ayenew, & Getahun, 2021). In 2014 in Bangladesh the number of having ANC visits equal for women whose wealth index status is middle compared to whose status is poor. The average number of ANC visits is higher for women whose wealth index status is rich compared to whose status is middle (Hossain et al., 2020).

In my study The incidence rate for having ANC visits is lower for women at their 2nd and 3rd birth than the women at their 1st birth.The incidence rate for having ANC visits is lower for women at their above 3rd birth than women at their 1st birth and this is also higly significant. In 2014 in our country the expected number of having ANC visits is higher at their 1st birth than 2nd and 3rd birth .The expected number of having ANC visits is lower for above 3rd birth than the women at their 2nd and 3rd birth (Hossain et al., 2020).

In my study the excess to media is higher for having ANC visits than the women who are not exposed by media. The same result is demonstrated by the previous study in Bangladesh by (Hossain et al., 2020) and in Ethiopia by (Fenta et al., 2021).

In my study when mother's age at their current birth was less than 20 is higher ANC visits compared to those whose age was greater than 35. when mother's age at their current birth was between 20 and 30 is lower compared to those whose age was between 20 and 30. But in 2014 in our country when mother's age at their current birth was less than 20 is lower ANC visits compared to those whose age was between 20 and 30. when mother's age at their current birth was greater than 35 is lower compared to those whose age was between 20 and 30 (Hossain et al., 2020).

5.2 Recommendation

The relative importance analysis shows that education is the most important predictor of the number of ANC visits, so government and policy-makers have to emphasize on education. Media exposure is the second most important predictor of the number of ANC visits. So, radio television and newspaper should publish the importance of good maternal health regularly.

5.3 Further scope of the study

Poisson regression is the most elementary approach for handling count data. None of the studies rank the order of the predictors of the determinants of the number of ANC. dominance analysis technique has been widely used by researchers to examine the predictor importance more accurately in the linear regression. Budescu (1993) introduced

one of the most popular methods, dominance analysis. Further we can use this analysis for zero truncated , zero-inflated and hurdle regression etc model. Government and policy maker can use this approach for determine the important predictor for their further decision making.

References

Abbas, A. M., Rabeea, M., Hafiz, H. A. A., & Ahmed, N. H. (2017). Effects of irregular antenatal care attendance in primiparas on the perinatal outcomes: a cross sectional study. *Proceedings in Obstetrics and Gynecology*, *7*(2).

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological methods*, *8*(2), 129.

Azen, R., & Traxel, N. (2009). Using dominance analysis to determine predictor importance in logistic regression. *Journal of Educational and Behavioral Statistics*, *34*(3), 319–347.

Bekalo, D. B., & Kebede, D. T. (2021). Zero-inflated models for count data: an application to number of antenatal care service visits. *Annals of Data Science*, *8*(4), 683–708.

Bhowmik, K. R., Das, S., & Islam, M. A. (2020). Modelling the number of antenatal care visits in bangladesh to determine the risk factors for reduced antenatal care attendance. *PloS one*, *15*(1), e0228215.

Budescu, D. V. (1993). Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychological bulletin*, *114*(3), 542.

DeJardine, Z. V. C. (2013). Poisson processes and applications in hockey. *Lakehead University, Thunder Bay, Ontario, Canada. URL https://www. lakeheadu.*

ca/sites/default/files/uploads/77/docs/DejardineFinal. pdf.

Duodu, P. A., Bayuo, J., Mensah, J. A., Aduse-Poku, L., Arthur-Holmes, F., Dzomeku, V. M., ... Nutor, J. J. (2022). Trends in antenatal care visits and associated factors in ghana from 2006 to 2018. *BMC pregnancy and childbirth*, *22*(1), 1–14.

Fenta, S. M., Ayenew, G. M., & Getahun, B. E. (2021). Magnitude of antenatal care service uptake and associated factors among pregnant women: analysis of the 2016 ethiopia demographic and health survey. *BMJ open*, *11*(4), e043904.

Hossain, Z., Akter, R., Sultana, N., & Kabir, E. (2020). Modelling zero-truncated overdispersed antenatal health care count data of women in bangladesh. *PloS one*, *15*(1), e0227824.

Hossain, Z., et al. (2021). Analyzing overdispersed antenatal care count data in bangladesh: Mixed poisson regression with individual-level random effects. *Austrian Journal of Statistics*, *50*(4), 78–90.

Islam, U. N., Sen, K. K., & Bari, W. (2022). Living standard and access to tetanus toxoid immunization among women in bangladesh. *BMC Public Health*, *22*(1), 1–11.

Jo, Y., Alland, K., Ali, H., Mehra, S., LeFevre, A. E., Pak, S. E., ... Labrique, A. B. (2019). Antenatal care in rural bangladesh: current state of costs, content and recommendations for effective service delivery. *BMC health services research*, *19*(1), 1–13.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data* (Vol. 1230). Springer.

McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. Routledge.

NIPORT, I. (2020). Bangladesh demographic and health survey 2017-18. *Dhaka and Maryland*.

Sarker, B. K., Rahman, M., Rahman, T., Rahman, T., Khalil, J. J., Hasan, M., ... others (2020). Status of the who recommended timing and frequency of antenatal care visits in northern bangladesh. *PLoS One*, *15*(11), e0241185.

Sultana, M., Mahumud, R. A., Ali, N., Ahmed, S., Islam, Z., Khan, J. A., & Sarker, A. R. (2017). Cost of introducing group prenatal care (gpc) in bangladesh: a supply-side perspective. *Safety in Health*, *3*(1), 1–8.

Sultana, N., & Bari, W. (2017). Analyzing overdispersed antenatal care visits of pregnant women in bangladesh: Negative binomial regression model. *Dhaka University Journal of Science*, *65*(2), 133–137.

Terefe, A. N., & Gelaw, A. B. (2019). Determinants of antenatal care visit utilization of child-bearing mothers in kaffa, sheka, and bench maji zones of snnpr, southwestern ethiopia. *Health Services Research and Managerial Epidemiology*, *6*, 2333392819866620.

Wilson, J. R. (1989). Chi-square tests for overdispersion with multiparameter estimates. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *38*(3), 441–453.

Workie, M. S., & Lakew, A. M. (2018). Bayesian count regression analysis for determinants of antenatal care service visits among pregnant women in amhara regional state, ethiopia. *Journal of Big Data*, *5*(1), 1–23.

Appendix

R codes used in this thesis paper

```
### R code for poisson regression model##


rm(list=ls())

library(haven)

k<-read_sav('C:/thesis/bdhs 2017-18/main1.sav')

k

attach(k)

names(k)


library(MASS)

library(corpcor)

varlist<-c("ANC", "DIVISION_NEW","Wealth_index","place","Birth_order","media_expo

library(yhat)

k$ANC=as.numeric(k$ANC)

k$DIVISION_NEW=as.factor(k$DIVISION_NEW)

k$Wealth_index=as.factor(k$Wealth_index)

k$Birth_order=as.factor(k$Birth_order)
```

```
k$place=as.factor(k$place)

k$media_exposure=as.factor(k$media_exposure)

k$education=as.factor(k$education)

k$age=as.factor(k$age)

k$problem=as.factor(k$problem)

k$wanted=as.factor(k$wanted)

lm.out<-glm(ANC~DIVISION_NEW+Wealth_index+place+Birth_order+media_exposure+educat

a<-summary(lm.out)

a

b<-a$coefficients

c<-b[,1]

c

exp(c)

library(modEvA)

Dsquared(lm.out)

dv<-"ANC"

ivlist<-c("DIVISION_NEW","Wealth_index","place","Birth_order","media_exposure","e

ilist <- unlist(ivlist)

cols <- length(ilist)

dv<-as.factor(dv)

prmtn <- function(n) {

##function permutations from package e1071

if (n == 1)

return(matrix(1))

else if (n < 2)
```

```
stop("n must be a positive integer")
z <- matrix(1)
for (i in 2:n) {
x <- cbind(z, i)
a <- c(1:i, 1:(i - 1))
z <- matrix(0, ncol = ncol(x), nrow = i * nrow(x))
z[1:nrow(x), ] <- x
for (j in 2:i - 1) {
z[j * nrow(x) + 1:nrow(x), ] <- x[, a[1:i + j]]
}
}
dimnames(z) <- NULL
z
}

#prmtn<-prmtn(cols)[i,]
aps_formula <- function(dv, ilist, prm)
{
fd <- paste(dv, "~", sep="")
p <- sort(prm)
fi <- paste(ilist[p], collapse="+")
formula <- paste(fd, fi, sep="")
}
```

```
lmgm <- function(dataMatrix, dv, ivlist)
{

dom <- matrix(, nrow = dim(prmtn(cols))[1], ncol=cols)
order_v <- vector(mode = "character", length = dim(prmtn(cols))[1])
r_sq <- vector(mode="numeric")
tmp_r <- vector(mode = "numeric", length = cols)
b <- vector(mode = "numeric", length = cols)



count <- 0
for (i in 1:dim(prmtn(cols))[1])
{
prmtn<-prmtn(cols)[i,]
order_v[i] <-  paste(ivlist[prmtn], collapse="")
for(j in 1:cols)
{

f <- aps_formula(dv, ivlist, prmtn[1:j])

if(j == 1) a <- 0
else a <- b[j-1]
if(is.na(r_sq[f]))
```

```
{
r_sq[f]<-Dsquared(glm(f,dataMatrix,family=poisson()))
count <- count+1
}
b[j] <- r_sq[f]
tmp_r[j] = b[j] - a
}
dom[i, ] <- tmp_r[order(prmtn)]
}
rownames(dom) <- order_v
CR=apply(dom,2,mean)
return(list(DA = dom, CR = CR, C = count))
}



g<-lmgm(k,dv,ivlist)
g
A<-sort(g$CR)
apply(g$DA,1,sum)
poisson<-(g$CR/sum(g$CR))*100

## R-code for finding dispersion statistics value##



#From model fitting
```

```
rss<-sum(residuals(lm.out,type="pearson")^2)

disp=rss/lm.out$df.residual

round(disp,3)


#Using formula

mu<-predict.glm(lm.out,type="response")

pearson<-sum((ANC-mu)^2/mu)

dispersion<-pearson/lm.out$df.residual

round(dispersion,3)


#R-code for calculating Z-score test


case(1): Dean and lawless (1993)

z_dean=((ANC-mu)^2-ANC)/(sqrt(2)*mu)

z_score=lm(z_dean~1)

summary(z_score)



case(2):Winkelmann (2008)

z_win=((ANC-mu)^2-ANC)/(2*mu)

z_win=lm(z_win~1)

summary(z_win)



case(3):Cameron (2008)
```

```
z_cam=((ANC-mu)^2-ANC)/mu

z_cam=lm(z_cam~1)

summary(z_cam)



##R-code for calculating Lagrange multiplier test

n=nrow(k)

lag_multi=((sum(mu^2)-n*mean(mu))^2)/(2*sum(mu^2))

lag_multi

pchisq(lag_multi,1,lower.tail=F)



##R-code for Negative binomial model

rm(list=ls())

library(haven)

k<-read_sav('C:/thesis/bdhs 2017-18/main1.sav')

k

attach(k)

names(k)


library(MASS)

library(corpcor)


varlist<-c("ANC", "DIVISION_NEW","Wealth_index","Birth_order","media_exposure","e
```

```
library(yhat)

k$ANC=as.numeric(k$ANC)

k$DIVISION_NEW=as.factor(k$DIVISION_NEW)

k$Wealth_index=as.factor(k$Wealth_index)


k$Birth_order=as.factor(k$Birth_order)

k$place=as.factor(k$place)

k$media_exposure=as.factor(k$media_exposure)

k$education=as.factor(k$education)

k$age=as.factor(k$age)

k$problem=as.factor(k$problem)

k$wanted=as.factor(k$wanted)



lm.out<-glm.nb(ANC~DIVISION_NEW+Wealth_index+place+Birth_order+media_exposure+edu

a<-summary(lm.out)

a

b<-a$coefficients

c<-b[,1]

c

exp(c)

library(modEvA)

Dsquared(lm.out)
```

```
dv<-"ANC"

ivlist<-c("DIVISION_NEW","Wealth_index","place","Birth_order","media_exposure","e

ilist <- unlist(ivlist)

cols <- length(ilist)

dv<-as.factor(dv)

prmtn <- function(n) {

##function permutations from package e1071

if (n == 1)

return(matrix(1))

else if (n < 2)

stop("n must be a positive integer")

z <- matrix(1)

for (i in 2:n) {

x <- cbind(z, i)

a <- c(1:i, 1:(i - 1))

z <- matrix(0, ncol = ncol(x), nrow = i * nrow(x))

z[1:nrow(x), ] <- x

for (j in 2:i - 1) {

z[j * nrow(x) + 1:nrow(x), ] <- x[, a[1:i + j]]

}

}

dimnames(z) <- NULL

z

}
```

```
#prmtn<-prmtn(cols)[i,]
aps_formula <- function(dv, ilist, prm)
{
fd <- paste(dv, "~", sep="")
p <- sort(prm)
fi <- paste(ilist[p], collapse="+")
formula <- paste(fd, fi, sep="")
}




lmgm <- function(dataMatrix, dv, ivlist)
{

dom <- matrix(, nrow = dim(prmtn(cols))[1], ncol=cols)
order_v <- vector(mode = "character", length = dim(prmtn(cols))[1])
r_sq <- vector(mode="numeric")
tmp_r <- vector(mode = "numeric", length = cols)
b <- vector(mode = "numeric", length = cols)



count <- 0
for (i in 1:dim(prmtn(cols))[1])
{
```

```
prmtn<-prmtn(cols)[i,]

order_v[i] <-  paste(ivlist[prmtn], collapse="")

for(j in 1:cols)

{


f <- aps_formula(dv, ivlist, prmtn[1:j])


if(j == 1) a <- 0

else a <- b[j-1]

if(is.na(r_sq[f]))

{

r_sq[f]<-Dsquared(glm.nb(f,dataMatrix))

count <- count+1

}

b[j] <- r_sq[f]

tmp_r[j] = b[j] - a

}

dom[i, ] <- tmp_r[order(prmtn)]

}

rownames(dom) <- order_v

CR=apply(dom,2,mean)

return(list(DA = dom, CR = CR, C = count))

}
```

```
n<-lmgm(k,dv,ivlist)

n

neg<-(n$CR/sum(n$CR))*100

neg



#For graphical Representation

hist(Wealth_index)

hist(place,main="Histogram of palce of residence")

hist(region)

hist(education,main="Mother's educational achievement")

hist(Birth_order,main="Birth order")

hist(age,main="Mother's age at birth")

hist(media_exposure,main="Media exposure")

hist(problem,main="Distance for getting health facility")

hist(wanted,main="Wanted pregnancy")



#From Dominance analysis

#Poisson distribution

percentage<-c(9.67,4.27,28.78,19.90,9.35,3.12,20.39,1.78,2.69)

variables<-c("Place","Region","educational","Wealth index","Birth order","Age","E

a<-data.frame(variables,percentage)

barplot(a$percentage,names=a$variables,col = "cyan3")
```

```
#Negative binomial distribution

percentages<-c(9.09,4.76,28.80,19.72,9.55,1.67,20.63,2.73,3.02)

b<-data.frame(variables,percentages)

barplot(b$percentages,names=b$variables,col = "coral2")
```