

Modelling Cases of Spontaneous Abortion Using Logistic Regression

Edwin Kung'u Kagereki*, Anthony Wanjoya, Thomas Mageto

Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:

kagerekie@gmail.com (E. K. Kagereki), awanjoya@gmail.com (A. Wanjoya), tttmageto@gmail.com (T. Mageto)

*Corresponding author

To cite this article:

Edwin Kung'u Kagereki, Anthony Wanjoya, Thomas Mageto. Modelling Cases of Spontaneous Abortion Using Logistic Regression. *International Journal of Data Science and Analysis*. Vol. 5, No. 6, 2019, pp. 143-147. doi: 10.11648/j.ijdsa.20190506.16

Received: November 22, 2019; **Accepted:** December 16, 2019; **Published:** December 25, 2019

Abstract: Spontaneous abortion is the expulsion of a foetus before the 28th week of gestation. Studies approximate that 10-25% of pregnancies are lost due to miscarriages. This phenomenon's aetiology remains a mystery hence uncertainty of detecting its cause. Furthermore, most pregnant women realize they have conceived later in the gestation period and some start antenatal care late during the pregnancy. In Kenya, total fertility rate has decreased for the last three decades from 8.1 to 3.9. However, with the decrease of total fertility rate, prevalence of maternal mortality and morbidity factors has greatly impacted on the pregnancy. Among them is spontaneous abortion. This study used secondary data from Kenyatta national hospital and employed logistic regression to model miscarriage's risk factors, investigate socio demographic and lifestyle factors, to investigate interactions among identified risk factors and fit a predictive model. Significant socio demographic factors identified were age and recurrent miscarriage. A woman who had experienced prior miscarriage had a 7.5-fold risk. Lifestyle factors identified were body mass index, diabetes mellitus and HIV. Underweight women had a 13.2-fold risk. There were significant interactions between gravidity and previous miscarriage; diabetes and body mass index. A predictive model was fit. The model has a good measure of separability, 80% classification accuracy and it is significant.

Keywords: Spontaneous Abortion, Logistic Regression, Risk Factor

1. Introduction

The expectation of a pregnant woman is to hold her a baby when gestation period is over. However, for some pregnant women, this expectation is deprived from them. This is through pregnancy loss, which is among adverse pregnancy complications. Spontaneous abortion, also known as miscarriage is the main type of pregnancy loss [16]. Spontaneous abortion is an irreversible process which affects 10-25% of clinically recognized pregnancies [14]. However, the actual rate of miscarriage is even higher, as many women have very early miscarriages without ever realizing that they are pregnant. About 80% of miscarriages happen in the first trimester, but the risk of it declines as gestation time progresses.

Miscarriage is classified according to its frequency of occurrence experienced by pregnant women. The classification includes sporadic and recurrent miscarriage.

Sporadic miscarriage is single occurrence of pregnancy loss and affects 50% of women. Recurrent miscarriage is three or more occurrence of pregnancy losses and affects 1% of women. Furthermore, spontaneous abortion can be subdivided into threatened abortion, incomplete and complete miscarriages. Threatened miscarriage is abnormal vaginal bleeding with or without abdominal pain, and affects 20% of pregnancies. 5.5%-42.7% risk for subsequent complete miscarriage has been associated with spotting [10]. Inevitable miscarriage arises when severe cramps persist and accompanied by opening of the cervix.

Spontaneous abortion has negative effects on the affected pregnant woman. This can be extended to her partner. It magnifies emotional distress, grief and anxiety which eventually can lead to depression to affected women [4]. This burdens heavily the affected woman, draining her energy and will to live. Moreover, due to the magnification of emotional distress and grief, breakage of families and divorces can arise from it. Miscarriage is an economical burden to couples who

want to salvage the pregnancy and to the affected women going for therapy and counselling sessions. Etiology of spontaneous abortion is unknown [6]. It is not possible to state a cause with surety due to its heterogeneous and complexity of its etiology. However, over the past years, studies have found some factors which facilitate and promote occurrence of miscarriages. Common risk factors are extreme of age (both paternal and maternal), diabetes mellitus, alcohol consumption, smoking, caffeine intake, extremes of body mass index (BMI), anti-phospholipid syndrome, hypertension, low serum progesterone levels, infections and stress.

2. Methodology

2.1. Study Area and Data Source

Kenya is a country located in the East Africa. Kenyatta national hospital is located in Nairobi county. The data was obtained from Kenyatta national hospital.

2.2. Logistic Regression

Logistic function was invented for description of populations' growths. Logistic regression is a special model of generalized linear models with a link function as logarithm of odds. It was first suggested by [3] in the analysis of biological experiments, where later [5] diversified its implementation. It is popular in studies where the outcome variable (y) is binary or dichotomous in nature. Its popularity is because there is no necessity of assumptions [1]. In this study, dependent variable (Y) is binary and Y=1 if response is "yes" and Y=0 if response is "no". Let π be probability of an event,

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^n \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \quad (3)$$

Maximizing the log likelihood, we get two (p+1) equations which are non-linear and iterative solutions to them will be achieved using R statistical software. Thus fitted values of the logistic model are;

$$\hat{\pi}(x_i) = \frac{\exp(\hat{g}(x_i))}{1 + \exp(\hat{g}(x_i))} \quad (4)$$

With confidence level

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\beta}_j)} \quad (5)$$

For $j=0, 1, \dots, p$.

2.5. Model Assessment

It is obligatory to assess adequacy and appropriate of the model. To achieve this, likelihood ratio test will be conducted. We shall use likelihood ratio test to assess importance of each explanatory variable. The test statistic is:

$$D = -2 \sum_{i=1}^n \left[y_i \ln \left(\frac{\hat{\pi}_i}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \hat{\pi}_i}{1 - y_i} \right) \right] \quad (6)$$

$$\pi = \frac{\exp(\beta_0 + \sum_{i=1}^n \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^n \beta_i x_i)} \quad (1)$$

Where $\beta_i, i=0, 1, \dots, n$.

Logistic regression model is defined as:

$$g(x) = \ln \left[\frac{\pi}{1 - \pi} \right] = \beta_0 + \sum_{i=1}^n \beta_i x_i \quad (2)$$

2.3. Selection of Variables

2.3.1. Univariate Analysis

For nominal and ordinal variables, we shall use contingency table of outcome Y versus the independent variables and among independent variables themselves. Then use likelihood chi-square test and Fisher's Exact. For continuous variables, univariate logistic regression will be done. Therefore, any candidate with p -value < 0.05 was statistically significant.

2.3.2. Multivariate Analysis

Forward stepwise method was done so as to include variables in the model. At every stage, test of significance, based on likelihood ratio test, was conducted for variable inclusion. The process stopped when all significant variables ($p < 0.05$) were included in the model. All variables were subjected to step wise regression using R.

2.4. Estimating Parameters

The model coefficients are unknown quantities and so, we will be needed to estimate them. Method of maximizing likelihood function is to be used in estimating parameters. The likelihood for a given model is interpreted as the joint probability of the observed outcomes expressed as a function of the chosen regression model [7]. Log likelihood function is to be maximized. It is in the form:

Hosmer-Lemeshow test is commonly used to assess goodness fit of the model. We shall group the percentiles of estimated probabilities in ten groups. Grouping method is most preferred to fixed cut points especially when the estimated probabilities are small. Hosmer-Lemeshow statistic has chi-square distribution with 8 degrees of freedom.

3. Results and Discussions

3.1. Results

Table 1 indicates frequencies of those who experienced miscarriage (YES) and those who did not (NO). From the table below, miscarriage cases recorded were 12.3%, less compared to those who did not experience at all. This is consistent to [14] findings.

Table 1. Descriptive Statistics.

Event	Frequency	%
YES	24	12.3
NO	171	87.9

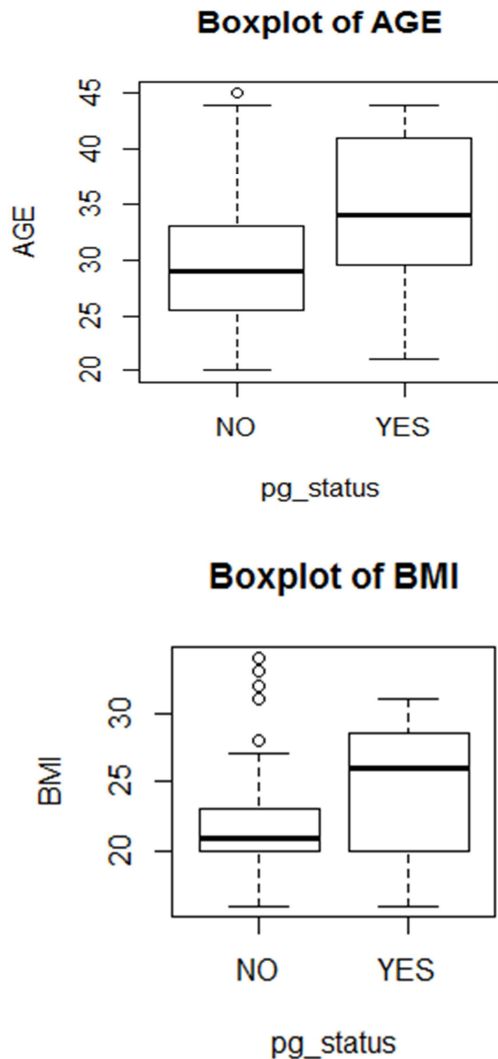


Figure 1. Box plots of BMI and age.

Figure 1 shows that the average age of women who experienced miscarriage was higher unlike those who did not. Overall, the mean of women was 30 years with the youngest woman being 20 as the eldest being 45. Age was categorized into three groups; 20-30, 30-40 and 40+. Also the figure shows that the average BMI of women who experienced miscarriage was higher unlike those who did not. The mean of BMI of women was 22.21. This continuous variable was sub divided into underweight, overweight, obese and normal. Gravidity is continuous variable where its mean is 1.35 \pm 1. Maximum recorded gravidity was 5 as 0 was the least among the women. It was also sub divided into two categories, that is nulligravid and multigravid. The rest predictor variables are dichotomous in nature with two factor levels.

Women above 40 years had the highest risk followed by those between 30-40 years as compared to youngest category. Those who had previously miscarried have a 5.4-fold risk than those who never had. Multigravid women have an almost 3-fold risk as compared to nulligravid women. It followed that obese women were far much vulnerable as 12

of them lost their pregnancies. This constituted to half of the miscarriages. They have highest risk followed by those underweight and overweight women. Lifestyle factors included smoking, alcohol consumption, caffeine intake, diabetes and living with HIV virus. Diabetic women and HIV infected women have about 3-fold risk. This summarized in table 2.

Table 2. Potential risk factors for miscarriage.

Variables	Crude OR (95% CI)	p-value
Age		
<30	1	
30-40	2.3 (0.8-7.1)	0.00104
40+	11.3 (3.4-40.5)	
Gravidity		
0	1	
1+	2.98 (0.97-12.99)	0.076
Marital status		
single	1	
cohabit	2 (0.6-8.8)	0.274
Previous miscarriage		
No	1	
Yes	5.4 (2.1-13.8)	0.00418
Nausea		
No	1	
Yes	0.8 (0.3-2.3)	0.611
BMI		
<20	3.8 (0.7-15.5)	
20-24	1	
25-29	0.96 (0.1-4.2)	0.00099
30+	19.5 (6.5-64.3)	
Smoking		
No	1	
Yes	0.6 (0.3-1.5)	0.317
Caffeine		
No	1	
Yes	0.7 (0.3-1.8)	0.498
Alcohol		
No	1	
Yes	0.7 (0.3-1.6)	0.4
Diabetes		
No	1	
Yes	3.1 (1.3-7.4)	0.009
HIV		
No	1	
Yes	2.98 (1.2-7.2)	0.011

Overall significant variables are those with $p < 0.05$. These are age, previous miscarriage, BMI, diabetes and HIV status. Interactions among predictor variables were also assessed. Significant interactions were observed among gravidity and previous miscarriage ($r=0.26$, $p= 0.000$) and BMI and diabetes ($r=0.31$, $p=0.000$).

A multiple logistic regression was conducted to obtain adjusted odds ratio of variables. We observed that women above 40 had the greatest risk followed by those between 20-30 years of age compared to the youngest age group. Those who had previously experienced miscarriage have almost 8-fold risk compared to those who never experienced it. Obese women have the highest risk and underweight women have 13-fold risk compared to women with normal BMI. Alcohol increased the risk thrice, however this study found it not statistically significant. Diabetic women have almost 10-fold

risk. HIV positive women have a 6-fold risk. This summarized in table 3.

Table 3. Adjusted odds ratio of risk factors for miscarriage.

Variables	Crude OR (95% CI)	p-value
Age		
<30	1	
30-40	5.7 (0.7-46.8)	0.00905
40+	100.9 (7.9-1416.3)	
Gravidity		
0	1	
1+	0.9 (0.1-8.0)	0.9779
Marital status		
single	1	
cohabit	4.6 (0.4-49.4)	0.2075
Previous miscarriage		
No	1	
Yes	7.5 (1.2-45.1)	0.0283
Nausea		
No	1	
Yes	0.5 (0.1-2.6)	0.3958
BMI		
<20	13.2 (1.5-112.6)	
20-24	1	
25-29	0.7 (0.1-6.3)	0.00155
30+	124.3 (14.3-1077.3)	
Smoking		
No	1	
Yes	0.3 (0.1-1.5)	0.1460
Caffeine		
No	1	
Yes	0.6 (0.1-2.6)	0.4893
Alcohol		
No	1	
Yes	2.9 (0.7-12.5)	0.1548
Diabetes		
No	1	
Yes	9.9 (1.9-50.3)	0.0054
HIV		
No	1	
Yes	6.2 (1.3-30.4)	0.0248

3.2. Predictive Model

Statistically significant variables were included in the multiple logistic model through stepwise regression on the basis of likelihood ratio test. Only five variables were found to be significant ($p < 0.05$); age, previous miscarriage, BMI, diabetes and HIV status. The logistic regression output is shown on the table below.

Table 4. Logistic regression output.

Variable	β	S.E	z	$p > z $
Intercept	-11.6495	2.44355	-4.767	$1.87e^{-6}$
Age	0.11499	0.04489	2.561	0.01042
Previous miscarriage	1.16628	0.60047	1.942	0.05210
BMI	0.20552	0.06878	2.988	0.00281
Diabetes	1.25252	0.51822	2.417	0.02207
HIV status	1.27691	0.55781	2.289	0.01565

Some statistics were calculated to determine the accuracy and nature of the model. Table 5 summarizes the information about the statistics.

Table 5. Characteristics of predictive model.

Statistic	Value	p-value
Log likelihood	-53.46207	$2.925e^{-7}$
Hosmer-Lemeshow	7.8064	0.4526
Classification accuracy	0.801	
auc	0.855	

From Table 5, we can observe that the model is significant. p -value of Hosmer-Lemeshow statistic is greater than 0.05. This is interpreted as the model not being a poor fit. Area under curve for ROC measures the level of separability. We can observe that the value of area under curve tends towards one. It has misclassification error of 20% which is good for the model.

3.3. Discussion

From this study, we found significant miscarriage risk factors to be age, previous miscarriage, body mass index, and diabetes and HIV status.

Maternal age is a major risk factor. Maternal aging is associated with increased infertility, miscarriage and poor prognosis of pregnancy. Studies on miscarriages all conclude that indeed it is a major risk factor. From this study, we found that women had a higher risk as compared to the younger expectant women. Moreover, advanced age has been linked to affecting ovarian aging which enhances rate of meiotic errors in the oocyte [9]. These errors result to fetal anomalies, which contributes more than 50% of spontaneous abortions. Maternal age and previous miscarriages independently decrease live birth rate of subsequent pregnancy [15]. Recurrence of miscarriage is a risk factor which affects 1% of pregnancies. In this study, women with a history of miscarriages had a higher significant risk compared to those with lack of the history. Nevertheless, a significant interaction between gravidity and history of miscarriage was observed.

Low BMI ($<20 \text{ kgm}^{-2}$) is risk factor for miscarriage in early weeks of gestation [2]. Obesity is mostly associated with lower progesterone levels among expectant women [8]. Miscarriage risk is higher in underweight and obese women than those with normal BMI [12]. This was evident from our study.

HIV infection increases 6% risk of miscarriage annually [13]. From the study, it is evident that HIV infection is a significant risk factor. It is evident, diabetes mellitus is a significant risk factor. More so, there is a significant positive interaction between diabetes and BMI. This is consistent with previous studies.

Majorly, logistic regression is used in predicting response variable rather than estimating probability. Therefore, it was the best approach to adopt in this study. Its classification accuracy is best achieved when applied to small and moderate sized datasets [11]. Assessing the predictive model, we found that it is a significant model of no poor fit with

classification accuracy is 80%.

4. Conclusion and Recommendation

Miscarriage is a frequent adverse outcome of pregnancy, as more than 12% of all recognized pregnancies end in a spontaneous abortion. Age and previous miscarriage were found to be significant socio-demographic risk factors. Women age forty and above are more vulnerable to experience miscarriage. Hence, women need to be advised to have children at legal desirable younger age other than when they are old. Also, recurrence of miscarriage increases the risks. It was found that gravidity increases recurrent miscarriage.

Significant lifestyle risk factors were body mass index, diabetes and HIV. Low BMI and obesity are contributory risk factors for miscarriages. This is evident from this study. Moreover, there is a positive relationship between BMI and diabetes. This implies that body mass index is a causative condition of diabetes mellitus. The latter is a significant miscarriage's risk factor. Diabetic women are prevalent in experiencing miscarriage. It is evident HIV increases miscarriage risk.

In studies where the response variable is binary and dataset is small to moderate sized, logistic regression is the best approach to adopt. This study exhibited the above characteristics. Since it is a challenging task to estimate probability of pregnancies being lost due to miscarriage, logistic regression was used to predict response variable. Application of logistic regression in building predictive model was a success. This is because our predictive model has a classification accuracy of 80% and a good measure of separability of 85.5%. More so, it is not of poor fit.

References

- [1] Akgül A. and Çevik O., 2003, Statistical Analysis Methods "Management Implementation is SPSS", Emek Offset, Ankara.
- [2] Arck, P. C., Rucke, M., Rose, M., Szekeres-Bartho, J., Douglas, A. J., Pritsch, M., Blois, S. M., Pincus, M. K., Barenstrauch, N., Dudenhausen, J. W., et al. (2008). Early risk factors for miscarriage: a prospective cohort study in pregnant women. *Reproductive biomedicine online*, 17 (1): 101-113.
- [3] Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39 (227): 357-365.
- [4] Callander, G., Brown, G. P., Tata, P., and Regan, L. (2007). Counterfactual thinking and psychological distress following recurrent miscarriage. *Journal of Reproductive and Infant Psychology*, 25 (1): 51-65.
- [5] Cox, D. (1970). *The analysis of binary data*, Methuen & co. Ltd., London, pages 48-52.
- [6] Cramer, D. W. and Wise, L. A. (2000). The epidemiology of recurrent pregnancy loss. In *Seminars in reproductive medicine*, volume 18, pages 331-340.
- [7] Dietz K., Gail M., Krickeberg K., Samet J. and Tsatis A., (2005). *Regression Methods in Biostatistics; Linear, Logistic, Survival, and Repeated Measures Models*. Statistics for Biology and Health. ISBN 0-387-20275-7.
- [8] Goh, J. Y., He, S., Allen, J. C., Malhotra, R., and Tan, T. C. (2016). Maternal obesity is associated with a low serum progesterone level in early pregnancy. *Hormone molecular biology and clinical investigation*, 27 (3): 97-100.
- [9] Hassold, T. and Hunt, P. (2001). To err (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews Genetics*, 2 (4): 280.
- [10] Kouk, L. J., Neo, G. H., Malhotra, R., Allen, J. C., Beh, S. T., Tan, T. C., and Ostbye, T. (2013). A prospective study of risk factors for first trimester miscarriage in Asian women with threatened miscarriage. *Singapore Med J*, 54 (8): 425-431.
- [11] Lim, T.-S., Loh, W.-Y., and Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine learning*, 40 (3): 203-228.
- [12] Metwally, M., Saravelos, S. H., Ledger, W. L., and Li, T. C. (2010). Body mass index and risk of miscarriage in women with recurrent miscarriage. *Fertility and sterility*, 94 (1): 290-295.
- [13] Otieno, G., Okanda, J., Kinuthia, J., John-Stewart, G., Akelo, V., and Kohler, P. (2018). Prior miscarriage prevalence higher among hiv positive than in hiv-negative women: A community survey in rural kenya. *J Pediatr Womens Healthcare*. 2018; 1 (2), 1011.
- [14] Regan, L. and Rai, R. (2000). Epidemiology and the medical causes of miscarriage. *Best practice & research Clinical obstetrics & gynaecology*, 14 (5): 839-854.
- [15] Sugiura-Ogasawara, M., Ozaki, Y., Kitaori, T., Suzumori, N., Obayashi, S., and Suzuki, S. (2009). Live birth rate according to maternal age and previous number of recurrent miscarriages. *American Journal of Reproductive Immunology*, 62 (5): 314-319.
- [16] Warakamin, S., Boonthai, N., and Tangcharoensathien, V. (2004). Induced abortion in Thailand: current situation in public hospitals and legal perspectives. *Reproductive Health Matters*, 12 (sup24): 147-156.