Journal of Applied Medical Sciences, vol. 1, no. 2, 2012, 15-40
ISSN: 2241-2328 (print version), 2241-2336 (online)
Scienpress Ltd, 2012

Comparative Study of the Application of Box Behnken Design (BBD) and Binary Logistic Regression (BLR) to Study the Effect of Demographic Characteristics on HIV Risk in South Africa.

Wilbert Sibanda[1] and Philip Pretorius[2]

Abstract

In this study, a Box Behnken Design (BBD) and a Binary Logistic Regression (BLR) were applied to study the effects of demographic characteristics on the risk of HIV in South Africa. The demographic characteristics studied for each pregnant mother attending an antenatal clinic in South Africa, were mother's age, partner's age (father's age), mother's level of education and parity. Using the 2007 South African antenatal seroprevalence data, the BBD design showed that HIV status of a pregnant woman was highly sensitive to changes to her age and educational level. These results were independently confirmed by the BLR model. Individually the father's age and parity had no significant effect on the HIV status. However, the latter two demographic characteristics showed significant effects on the HIV risk in two way interactions with other demographic characteristics. The results from the BBD provided the following summary statistics, R^2 = 0.99 and two-factor interactions (2FI) model F-value of 88.29. The latter value of 88.29 for the BBD 2FI model is significant with only a 0.01% chance that this value could be due to noise. An adeq.precision value of 31.33 was obtained for the BBD further confirming that the 2FI model could be used to navigate the experimental design space. Finally, the 3-D response surface plots of HIV risk against mother's age and her education were created.

[1] DST/NWU Preclinical Platform, North-West University, South Africa.
e-mail: Wilbert.sibanda@nwu.ac.za
[2] School of Information Technology, North-West University, South Africa.
e-mail: Philip.pretorius@nwu.ac.za

Article Info: *Received :* November 2, 2012. *Revised :* December 12, 2012.
Published online : December 27, 2012

Keywords: Response Surface design, Central Composite design, Face Centered, Demographic, Seroprevalence.

1 Introduction

The South African HIV antenatal HIV seroprevalence survey is the largest in the world with a sample size of 36 000 subjects [1]. The antenatal HIV data assists in monitoring the HIV infection trends within the republic of South Africa. A thorough understanding of the epidemic leads to a fruitful utilization of resources and development of innovative approaches to curb the spread of the disease. In South Africa, the antenatal clinic surveys are conducted annually in October to obtain an estimate of the prevalence levels of that year [1].

The first case of HIV infection in South Africa was observed in 1982 [2]. Initially the epidemic was confined to the gay community. However, since 1982 the epidemic has grown in leaps and bounds to be the leading killer in South Africa. In 2007, an estimated 1.7 million people in the sub-region were nearly infected with HIV, the majority of them being women [13].

The control of HIV in South Africa is multi-disciplinary involving government, research and academic institutions, civil society, non-governmental organizations, community-based organizations and the private sector. All these organizations work together to alleviate health and social consequences of HIV and AIDS.

As stated in the abstract, the raw antenatal clinic data contains the following demographic characteristics for each pregnant woman; age, population group, educational level, gravidity, parity, partner’s age, name of clinic, HIV and syphilis results [4].

This research paper follows on our previous work that explored the application of Response Surface Methodologies (RSMs) to study the relationship between demographic characteristics and HIV risk. Two RSMs techniques were used namely Central Composite Face-Centered (CCF) and Box-Behnken Designs (BBD). The two RSM designs demonstrated that the mother’s age had the greatest influence on the HIV risk of antenatal clinic attendees.

This work aims to compare the BBD and BLR techniques in predicting and determining the effect of demographic characteristics on HIV prevalence in South Africa.

1.2 Literature Review

1.2.1 Logistic Regression

Logistic regression facilitates the investigation of the relationship between a response and a set of explanatory variables. The response can be dichotomous in nature. The logistic regression is a generalized linear model, which is a type of binomial regression. The purpose of the logistic regression is to transform the limited range of a probability, restricted to the range 0 to 1 into the full range $-\infty$ to $+\infty$, which makes the transformed value more suitable for fitting using a linear function [14]. In epidemiology, such a probability between 0 and 1 gives the risk of an individual getting a disease.

The assumptions for logistic regression

- The outcome must be discrete

- There needs to be enough responses as a ratio of variables to reduce standard errors and enhance maximum likelihood estimation
- The regression equation should have a linear relationship with the logit form of the discrete variable
- There should be an absence of multi-collinearity
- There should be no outliers
- There should be independence of errors

Logistic Regression Model

The logistic model is designed to ensure that whatever estimate of risk is obtained, it will always be some number between 0 and 1. This therefore means that for the logistic model, the risk estimate cannot be above 1 or below 0.

The Shape of Logistic Model

As shown in Fig. 1, as z moves from $-\infty$ towards $+\infty$, the value of f(z) hovers close to zero for a while, then starts to increase dramatically towards 1, and finally leveling off around 1.The result is an elongated, S-shaped diagram.

The S-shape of the logistic function is favored by epidemiologists, provided the variable z stands for an index that combines combinations of several risk factors, and f (z) represents the risk for a given value of z.

Fundamentally, the S-shape of f(z) indicates that the effect of z on an individual's risk is minimal for low z's until some threshold is attained. The risk then rapidly increases over a certain range of intermediate z-values, and then remains extremely high around 1 as soon as z gets large enough.

The threshold idea is believed by epidemiologists to apply to a variety of disease conditions. In general, an S-shaped model is believed to be ideal for multivariate nature of epidemiological research [16].

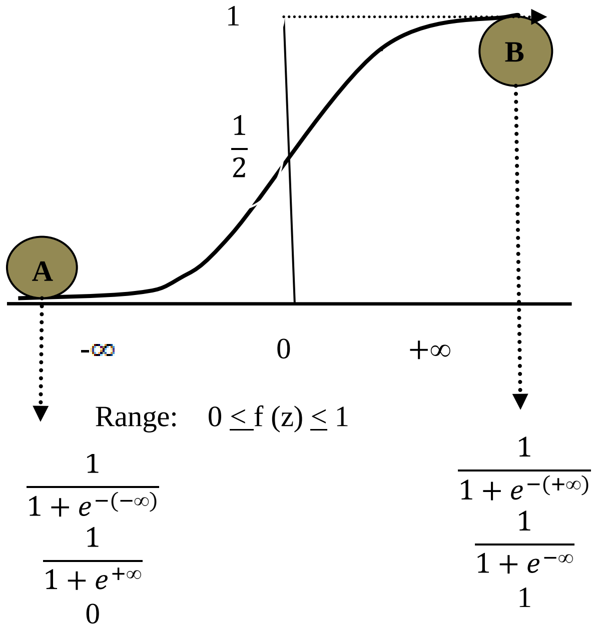
1
$\frac{1}{2}$
B
A
-∞
0
+∞
Range: $0 \leq f(z) \leq 1$
$\frac{1}{1+e^{-(-\infty)}}$
$\frac{1}{1+e^{+\infty}}$
0
$\frac{1}{1+e^{-(+\infty)}}$
$\frac{1}{1+e^{-\infty}}$
1

Figure 1: Shape of Logistic Model

Development of Logistic Model from Logistic Function

The logistic function is expressed as:

$$Z=\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \quad (1)$$

z is the linear sum α plus β_1 times x_2 and so on to β_k times x_k where x's are independent variables of interest and α and β_1 are constant terms representing unknown parameters. Therefore z is an index that combines the x's.

$$\begin{tikzcd}
z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \arrow[d] \\
f(z) = \dfrac{1}{1+e^{-(z)}}
\end{tikzcd}$$

Figure 2: Logistic model

The independent variables x_1, x_2 and so on up to x_k on a group of subjects, with disease status either 1 (with disease) or 0 (without disease). Using the logistic function to describe a disease over a period of time T_0 to T_1, in a disease-free individual with independent variables x_1, x_2 up to x_k measured at time T_0.

The probability being modeled can be denoted by the conditional probability statement $P(D=1|x_1, x_2, \ldots x_k)$. The model is defined as logistic if the expression for the probability for the probability α plus the sum from i equals 1 to k of β_i times x_i. The terms α and β_i in this model represent unknown parameters that we need to estimate based on data obtained on the X's and on D (disease outcome) for a group of subjects.

Therefore, a knowledge of the parameter α and β_i and determination of values x_i through x_k for a particular disease-free individual, this formula can be used to obtain probability that a given individual would develop a disease over a given period of time.

Letting $P(D=1|x_1, x_2, \ldots x_k)$ be P(x) where x is a collection of variables x_1 through $x_{k.}$ Therefore the logistic model formula is:

$$P(x) = \frac{1}{1+e^{-(\alpha + \Sigma\beta iXi)}} \quad (2)$$

Logit Transformation

An alternative way of writing the logistic regression is the logit form of the model. To obtain a logit form from the logistic model, a transformation is required. The logit transformation logit p(x) is given by the natural log (to base e) of the quantity p(x) divided by one minus p(x), where p(x) represents the logistic model. The transformation facilitates the calculation of a number called logit p(x) for an individual with independent variables given by x.

$$\text{logit } p(x) = ln_e\left[\frac{P(X)}{1+P(X)}\right]$$

$$\text{where } p(x) = \frac{1}{1+e^{-(\alpha + \Sigma\beta iXi)}}$$

$$1-p(x) = 1 - \frac{1}{1+e^{-(\alpha + \Sigma\beta iXi)}}$$

$$= \frac{e^{-(\alpha + \Sigma\beta iXi)}}{1+e^{-(\alpha + \Sigma\beta iXi)}}$$

$$\frac{P(x)}{1+P(x)} = \frac{\frac{1}{1+e^{-(\alpha + \sum\beta iXi)}}}{\frac{e^{-(\alpha + \sum\beta iXi)}}{1+e^{-(\alpha + \sum\beta iXi)}}}$$

$$= e^{(\alpha + \sum\beta iXi)}$$

$$ln_e[\frac{P(x)}{1-P(x)}] = ln_e[\alpha + \sum\beta ixi]$$

$$\textit{Therefore } p(x) = ln_e[\frac{P(x)}{1-P(x)}]$$

$$= \textit{log odds}$$

$$= \alpha + \sum\beta ixi \qquad (3)$$

Where odds is the ratio of probability that some event will occur over the probability that the same event will occur. Therefore $\frac{P(x)}{1-P(x)}$, describes the odds (risk) for developing a disease for an individual with independent variables specified by x.

Statistical Inferences from Logistic Regression

Maximum Likelihood (ML) Estimation

This is a statistical method for estimating the parameters in a mathematical model. ML is preferred for non-linear models such as the logistic models regression. Furthermore, ML estimates require no restrictions on the characteristics of the independent variables.ML value is therefore a numerical value of the likelihood function L when the ML estimates are substituted for the corresponding parameter values. There are three test procedures for ML namely the Likelihood Ratio (LR), Wald and the Score Tests.

Likelihood Ratio test

The difference between log likelihood statistics for two models, one of which is a derivation of the other has an approximate chi-square distribution in large samples. This type of statistic is called a likelihood ratio (LR) or LR statistic. The degrees of freedom (df) for this chi-square test are equal to the difference between the number of parameters in the two models.

$$-2lnL1 - (-2lnL2) = -2\ [\frac{L1}{L2}]$$

LR approximates the χ^2 variable with df of 1 and provided the sample is large enough. Furthermore, LR is like the F-statistic as it compares two models such as main effects against main effects and interactions.

For large contribution to the model L_2 is much larger than L_1, then

$$\frac{L1}{L2} \approx 0$$

$$Ln\ [\frac{L1}{L2}] \approx ln\ [0] = -\infty$$

$$LR = -2\ [\frac{L1}{L2}] \approx +\infty \qquad (4)$$

Therefore for highly significant additions to the model, LR is large and positive.

For no contribution to the model

$L_2 \approx L_1$

$\frac{L1}{L2} \approx 1$

$LR = -2ln\ (1) = -2\ x\ 0 = 0$

Therefore if the addition to the model is insignificant,

$$LR \approx 0 \tag{5}$$

0 (Not significant) < LR < ∞ (Significant)
Therefore LR approximates χ^2 if the sample (n) is large.

Wald Test
This is another hypothesis testing technique in logistic regression. The Wald test statistic is calculated by deviding the estimated coefficient of interest by its standard error. This test statistic has approximately a normal (0,1), or Z distribution in large samples. This square of this Z statistic is approximately a chi-square statistic with one degree of freedom. The likelihood ratio statistic and its corresponding squared Wald statistic give approximately the same value in very large samples.

$$LR \approx Z^2_{wald} \text{ in large samples} \tag{6}$$

Score Test
This is another method for hypothesis testing. The score test is designated to evaluate whether a model constrained by the proportional odds assumption is significantly different from the corresponding model in which the odds ratio parameters are not constrained by the proportional odds assumption. The test statistic is distributed approximately chi-square, with degrees of freedom equal to the number of odds ratio parameters. The score test gives the same numerical chi-square values as the LR and the Wald statistics.

1.2.2 Response Surface Methodology (RSM)

RSM is a collection of statistical and mathematical methods that are useful for modelling and analyzing design. RSM experiments are designed to allow forthe estimation of interaction and even quadratic effects, and thus provide an idea of the local shape of the response surface being investigated.

Linear terms alone produce models with response surfaces that are hyperplanes. The addition of interaction terms allows for warping of the hyperplane. Squared terms produce the simplest models in which the response surface has a maximum or minimum, and so an optimal response.

RSM comprises of fundamentally three techniques [5], namely:

i. Statistical experimental design
ii. Regression modelling
iii. Optimization

The detailed outline of the steps involved in the design of experiments using RSM is clearly indicated in figure 3.

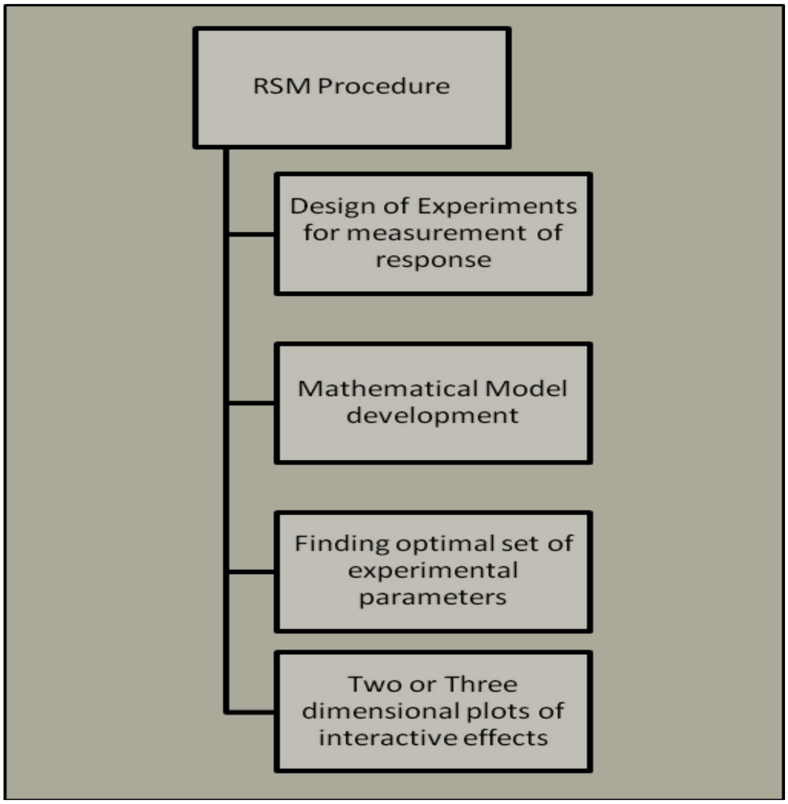
RSM Procedure
Design of Experiments for measurement of response
Mathematical Model development
Finding optimal set of experimental parameters
Two or Three dimensional plots of interactive effects

Figure 3: Design procedure of an RSM

An example of an RSM is the Box-Behnken Design.

Box Behnken (BBD) Design

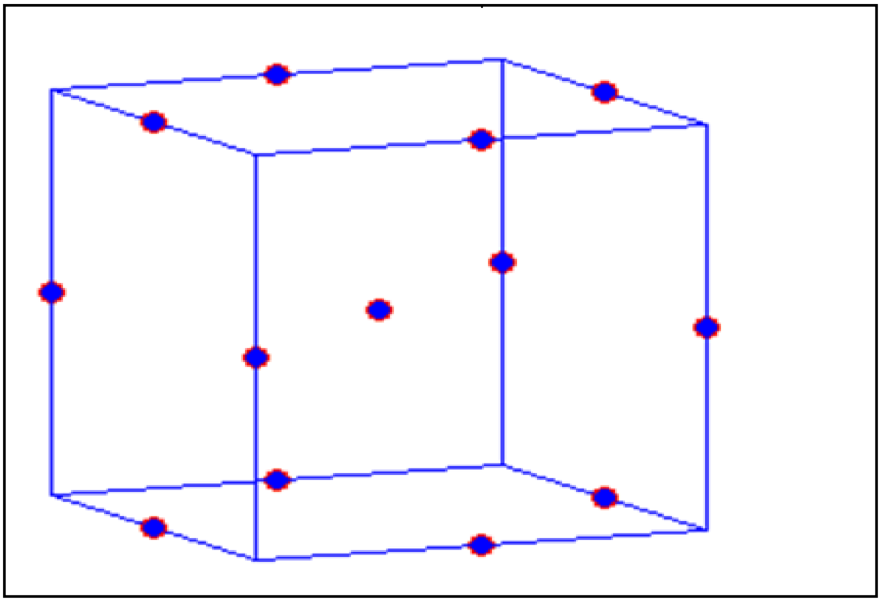

Figure 4: BBD Design [10]

The Box-Behnken design (Fig. 4) is an independent quadratic design, that does not contain an embedded factorial or fractional factorial design. The Box-Behnken design is characterized by treatment combinations at the midpoints of edges of the experimental space and the centre. These designs are rotatable and require 3 levels of each factor. The designs have limited capability for orthogonal blocking compared to the central composite designs [11].

2 Experimental Methodology

2.1 Sources of Data

Seroprevalence data studied was obtained from the 2007 South African antenatal data, supplied by the National Department of Health of South Africa [1]. The data consisted of about 32 000 subjects that attended antenatal clinics for the first time across the nine provinces of South Africa in 2007.

2.2 Research Tools

This research utilized the following research tools:

a) Design Expert V8 Software (StatEaseInc, 2011)
b) SAS 9.3, an integrated system of software products (SAS Institute Inc).
c) Essential Regression and Experimental Design, version 2.2 (Gibsonia, PA)
d) Minitab 16. Minitab Inc., United States.

2.3 Sampling Procedure

To facilitate the experimental design, the data was completely randomized, and this process was undertaken as a preprocessing technique to reduce bias in the design of experiment.

2.4 Missing Data

Out of the total of 31 808 cases from the 2007 South African antenatal seroprevalence database, 21 646 (68%) cases were found to be complete. 10 162 (32%) cases were incomplete and thus discarded.

2.5 Variables

The variables used in the study were mother's age, father's age, education, parity and HIV status. The integer value representing level of education stands for the highest grade successfully completed, with 13 representing tertiary education. Parity represents the number of times the individual has given birth. Parity is important as it shows the reproductive activity as well as reproductive health state of the women. The HIV status is binary coded; a 1 represents positive status, while a 0 represents a negative status.

2.6 Experimental Design

In this study, the aim was to use a Box-Behnken design (BBD) and a binary logistic regression to study the individual and interaction effects of demographic characteristics on the HIV status of a pregnant mother using seroprevalence data. A BBD designs with four factors and one response variable was developed as shown in Table 1. Based on sparsity-of-effects principle, two factor-interaction (2FI) design models were used, with 29 runs and no blocks. -1 and +1 denote the minimum and maximum levels of factors respectively.

Table 1: The BBD Matrix Design with 4 factors, 1 response variable and 4 center points.

	Factors				**Response**
Run	Mother's age	Father's age	Education	Parity	HIV
1	0	0	0	0	0.33
2	0	1	-1	0	-
3	1	0	-1	0	-
4	1	0	1	0	0.33
5	-1	0	1	0	-
6	0	-1	0	1	0.32
7	0	0	0	0	0.33
8	0	0	0	0	0.33
9	-1	0	0	1	-
10	0	-1	1	0	-
11	0	-1	-1	0	-
12	-1	0	0	-1	0.17
13	0	0	-1	-1	-
14	0	0	-1	1	-
15	0	0	0	0	0.33
16	1	0	0	1	-
17	-1	-1	0	0	0.28
18	0	0	1	1	-
19	0	0	0	0	0.33
20	1	0	0	-1	-
21	0	0	1	-1	0.33
22	-1	1	0	0	0
23	0	1	0	1	0.33
24	0	-1	0	-1	0.32
25	1	-1	0	0	-
26	1	1	0	0	-
27	0	1	1	0	-
28	-1	0	-1	0	0.11
29	0	1	0	-1	0.27

3 Design Matrix Evaluation

3.1 Degrees of Freedom

Design matrix evaluation showed that there were no aliases for the 2FI model and the degrees of freedom for the matrix are shown in Table 2. As a rule of thumb, a minimum of 3 lack-of-fit df and 4 pure error df ensure a valid lack of fit test. Fewer df tend to lead to a test that may not detect lack of fit [8].

Table 2: Degrees of Freedom for BBD Matrix

Model	10
Residuals	18
Lack of Fit	14
Pure Error	4
Corr total	28

3.2 Standard Errors

The standard errors of the BBD design are shown in fig. 5. The BBD design has large standard errors at the edges of the design space. The BBD design is not capable of estimating the response parameter at the edges of the experimental space. It is therefore advisable to work well within the design margins to achieve a greater degree of accuracy.

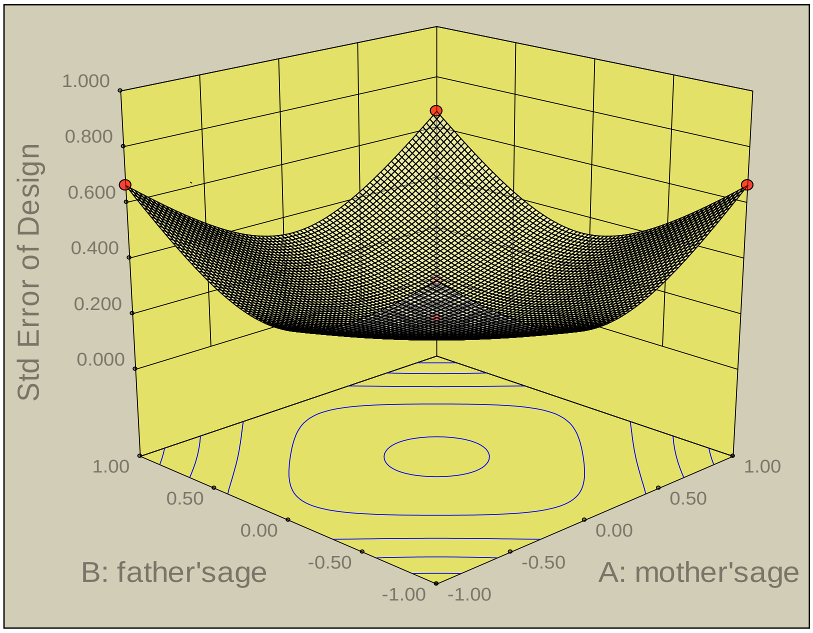

Figure 5: 3D Plot of standard error of BBD design

3.3 Variance Inflation Factor (VIF)

The variance inflation factor (VIF) quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity.

Therefore, VIF values should be ideally 1 and values greater than 10 indicate that coefficients are poorly estimated due to multicollinearity [8]. The VIF values in Table 3, indicate that coefficients of individual demographic characteristics and their interactions are estimated adequately without multicollinearity for the BBD design.

Table 3: Signal to noise ratio with the BBD design matrix

Term	**BBD**	
	VIF	**R_i^2**
A	1.0	0.0
B	1.0	0.0
C	1.0	0.0
D	1.0	0.0
E	1.0	0.0
AB	1.0	0.0
AC	1.0	0.0
AD	1.0	0.0
AE	1.0	0.0
BC	1.0	0.0
BD	1.0	0.0
BE	1.0	0.0
CD	1.0	0.0
CE	1.0	0.0
DE	1.0	0.0

3.4 R_i- Squared

In general, high R_i-squared values mean the terms are correlated with each other, leading to poor model. For this experiment, low R_i-squared values were obtained for individual factors and their interactions as shown in Table 3.

3.5 Fraction of Design Space (FDS)

FDS curve (fig. 6) is the percentage of the design space volume containing a given standard error of prediction or less. Flatter FDS curve means that the overall prediction error is constant. In general the larger the standard error of prediction, the less likely the results can be repeated, and the less likely that a significant effect will be detected.

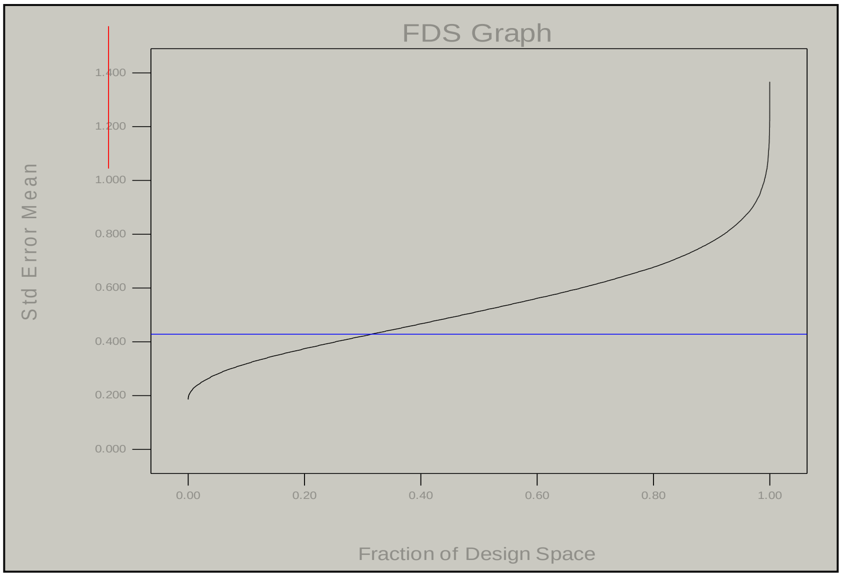

Figure 6: FDS plot of the standard error over the BBD design space

Fig. 6, it indicates that only 31% of the BBD design space is precise enough to predict the mean within +90.

3.6 Variance Dispersion Graphs (VDGs)

Variance dispersion graphs (VDGs) have recently become popular in aiding the choice of a response surface design [12]. Furthermore, variance dispersion graphs can be used to compare the performance of multiple designs for a specific models such as linear model, linear model with interaction terms, linear models with quadratic terms or for full quadratic model. VDGs were developed by Giovannitti-Jensen and Myers in 1989 [12].

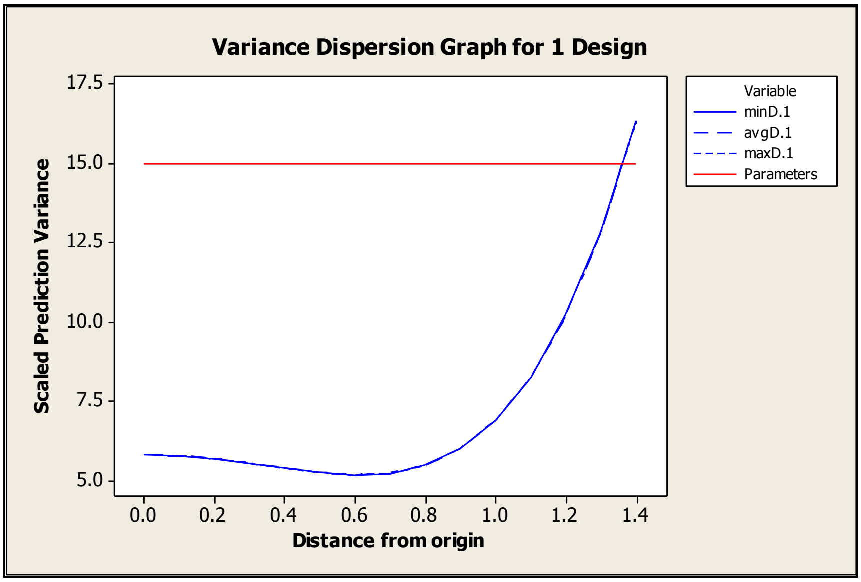

Figure 7: VDG graphs of BBG design

3.7 Choice of Levels for the Factors

Table 4: Factor Levels

	Levels		
Factor	**-1**	**0**	**1**
Parity	0	1	≥ 2
Education (Grades)	≤ 8	9-11	12-13
Mother'sage (years)	≤ 20	21-29	≥ 30
Father'sage (years)	≤ 24	25-33	≥ 34

4 Main Results

4.1 Model Fit Statistics

4.1.1 Box Behnken Design

a) Sequential Model Sum of Squares

This technique shows the effect of increasing terms to the complexity of the total model.

Table 5: Sequential Model Sum of Squares for the BBD Design

Source	Sum of Squares	F-value	P-value
Mean vs. Total	1.13		
Linear vs. Mean	0.097	5.37	0.014
2FI vs. Linear	0.044	49.83	0.000

For the BBD, the probability (P-value) is lowest for the 2FI models at a significance level of 0.05.

b) Model Summary Statistics

Table 6: Model Summary Statistics

Source	Standard Deviation	R^2	R^2 Adjusted	PRESS	Adeq. Precision
Linear	0.067	0.68	0.56	0.15	
2FI	0.013	0.95	0.98	-	31.33

The R^2 statistics of the linear models are considerably lower than those of the two-factor interactions (2FI) models. Therefore, the 2FI model has the lowest standard deviation, high R^2 and low Predicted Residual Sum of Squares (PRESS), implying that the 2FI model best fits the data. Statistically, high R^2 values imply that a large proportion of variation in the observed values is explained by the model.

Adeq.precision is used to measure the signal to noise ratio and a ratio greater than 4 is desirable indicating model can be used to navigate the design space. The BBD design has an adeq. precision of 31.33 indicating an adequate signal.

4.1.2 Binary Logistic Regression

In logistic regression, Deviance and Pearson's chi-squared goodness-of-fit are measures used to compare the overall difference between observed and fitted values. In addition, information criteria such as AKAIKE Information Criterion (AIC), Schwartz Criterion (SC) and negative log-likelihood, are used to measure goodness-of-fit for logistic regression models.

a) Pearson's Chi-Square test

Table 7: Pearson values

Criterion	**Value**	**DF**	**Value/DF**	**Pr > ChiSq.**
Pearson	47.14	25	1.89	0.0047

Pearson's chi-square statistic includes the test for independence in two-way contingency tables. This technique has been extended from generalized linear model theory to test for adequacy of the current fitted model. Given a model with responses y_i, weights w_i, fitted means μ_i, variance function $v(\mu)$ and dispersion $\Phi = 1$, the Pearson goodness-of-fit is;

$$\chi^2 = \sum \frac{wi(yi-\mu i)^2}{v\,(\mu i)} \tag{7}$$

If the fitted model is correct and observations y_i are approximately normal, then χ^2 is approximately distributed as χ^2on the residual degrees of freedom for the period. Therefore, Table 7 shows a desirable goodness-of-fit for the logistic regression model.

b) Residual Deviance

Table 8: Deviance values

Criterion	**Value**	**DF**	**Value/DF**	**Pr > ChiSq.**
Deviance	46.75	25	1.87	0.0052

The other goodness-of-fit test is the residual deviance. This is the log-likelihood ratio statistic for testing the fitted model against the saturated model in which there is a regression coefficient for every observation.

$$D^2 = 2\{ln[L_s(\beta)] - ln\ [L_m(\beta)]\} \tag{8}$$

where;

$\ln [L_m(\beta)]$ = maximized log-likelihood of the fitted model

$\ln [L_s(\beta)]$ = maximized log-likelihood of the saturated model

The deviance quantity compares the values predicted by the fitted model and those predicted by the most complete model we could fit. A very large D^2 value is evidence for model lack-of-fit. However under specificity regularity conditions D^2 converges asymptotically to a χ^2 distribution with h degrees of freedom, where h is the difference between the number of parameters in the saturated model and the number of parameters in the model being considered;

$$D^2 \approx \chi^2_h \tag{9}$$

The null hypothesis can be tested as follows;

H_0: $\beta_h = 0$

H_0 is rejected when:

$$D^2 > \chi^2_{1-\alpha} \tag{10}$$

α is fixed level of significance

If the null hypothesis cannot be rejected, it can be concluded that the fitting of the model of interest is substantially similar to that of the most completed model that can be built. Therefore, from the Deviance table (Table 8) above, it can be concluded that the saturated model has a desirable goodness-of-fit. The saturated model represents the largest model that can be fitted and leads to perfect prediction of the outcome of interest.

c) AKAIKE Information Criterion (AIC)

Table 9: AKAIKE Information Criterion (AIC)

Criterion	Intercept Only	Intercept and Covariates
AIC	18887.57	18189.01

This technique was introduced by AKAIKE in 1973, as tool of optimal model selection [15].

$$AIC = -2logL + 2((k-1) + s) \tag{11}$$

Where k is the number of levels of the dependent variable and s is the number of predictors in the model. Therefore as the number of independent variables k included in the model increases, the lack-of-fit term increases while the penalty decreases.

AIC is used for the comparison of models from different samples. The model with the lowest AIC is considered best as it minimizes the difference from the given model to the 'true' model. From Table 9, it is evident that the model with intercept and covariates better fits the data compared to the intercept only model.

d) Schwarz Criterion (SC)

Table 10: Schwarz Criterion (SC)

Criterion	Intercept Only	Intercept and Covariates
SC	18895.28	18227.53

Schwarz criterion was developed in 1978 as a model selection criterion. The model was derived from a Bayesian modification of the AIC criterion [15].

Schwarz criterion is defined as;

$$SC= -2logL + ((k-1) + S) * log (\sum f_i) \tag{12}$$

Where f_i's are the frequency values of the i^{th} observation, k is the number of levels of the dependent variable and S is the number of predictors in the model. Like AIC, SC penalizes for the number of predictors in the model and the smallest SC is most desirable. Table 10 shows that the intercept and covariates model is better than the intercept only model.

e) -2 logL

Table 11: 2Log L

Criterion	Intercept	Intercept and Covariates
-2Log L	18885.57	18179.01

This is negative two times the log likelihood. The -2logL is used in hypothesis testing for nested models.

The intercept only model is the logistic regression estimate when all variables in the model are estimated at zero. As shown in Table 11, the model with independent variables and the intercepts has lower -2logL value indicating that it is better than intercept only model. Furthermore, the values for the three measures of model-fit are similar.

4.2. ANOVA for 2FI Response Surface

4.2.1 Box Behnken Design

The ANOVA table (Table 12) for the BBD design confirms the adequacy of the 2FI model. The model F-value of 88.29 for the BBD is significant with only a 0.01% chance of this value being due to noise.

The BBD design confirms that the mother's age has the greatest effect on the HIV status of antenatal clinic attendees. The mother's educational level was the second most important individual factor. Also of note is the fact that the interaction of the mother's age with father's age and educational level significantly affects the HIV status of the antenatal clinic attendees.

Table 12: ANOVA Results (BBD design)

Source	Sum of Squares	F-value	P-value
Model	0.014	88.29	<0.0001
Mother's age	0.054	301.92	<0.0001
Father's age	0.0004	2.25	0.194
Education	0.0046	25.60	0.004
Parity	0.0009	5.06	0.074
Mother's age*Father's age	0.023	126.75	<0.0001
Mother's age*Education	0.017	94.10	0.0002
Mother's age*Parity	0.001	6.51	0.051
Father's age*Parity	0.0009	5.06	0.074
Education*Parity	0.0004	22.56	0.0051

4.2.2 Binary Logistic Regression (BLR)

a) Likelihood Ratio (LR), Wald and Score Tests

The results of the LR, Wald and the Score Test for testing the joint significance of the explanatory variables are included in Table 13, below.

Table 13: ANOVA Results (BLR)

Test	Chi-Square	DF	Pr>Chi-Square
Likelihood Ratio	726.07	3	<0.0001
Score	670.25	3	<0.0001
Wald	628.26	3	<0.0001

The likelihood ratio of 726.07 therefore confirms that the fitted model with intercept and covariates is important and has a significant effect on the basic model with no predictors. The three hypothesis testing techniques (LR, Score and Wald Tests) confirm the effect of the addition of covariates to the basic model with intercept only. In general, for large samples the LR is approximately equal to the Wald score, as shown in Table 13.

4.3 Model Adequacy Checking

Model adequacy checking is conducted to verify whether the fitted model provides an adequate approximation to the true system and to verify that none of the least squares regression assumptions are violated. Extrapolation and optimization of a fitted response surface will give misleading results unless the model is an adequate fit.

There are many statistical tools for model validation, but the primary tool for most process modeling applications is graphical residual analysis.

The residual plots assist in examining the underlying statistical assumptions about residuals. Therefore residual analysis is a useful class of techniques for the evaluation of the goodness of a fitted model.

4.3.1 Residual Analysis

Box Behnken Design

a) Normality Probability Plot of Residuals

A normal probability plot of residuals can be used to check the normality assumption. If the residuals plot approximates a straight line, then the normality assumption is satisfied.

Furtherrmore, the normal plot of residuals as shown in Fig. 11, evaluates whether there are outliers in the dataset. All the points lie on the diagonal, implying that the residuals constitute normally distributed noise.

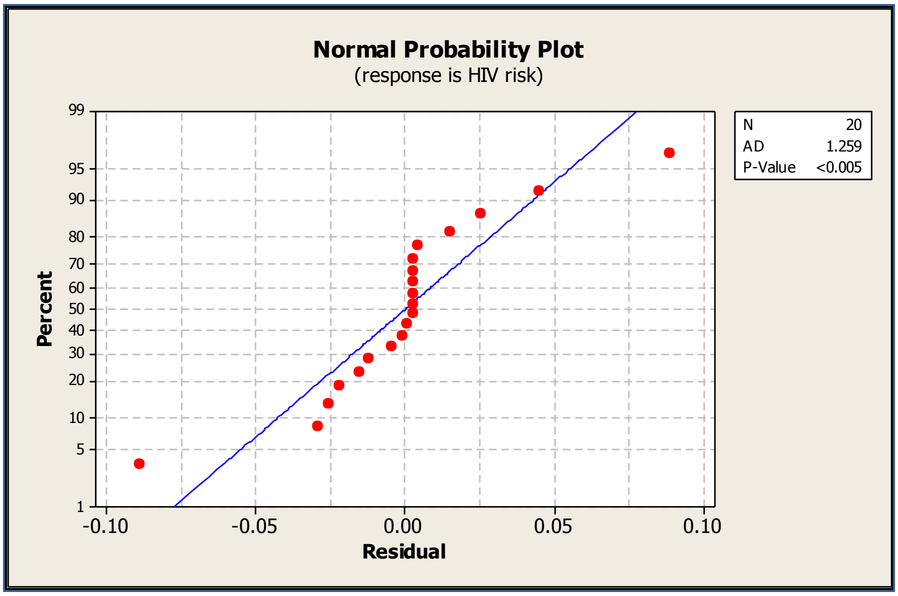

Figure 8:Normal Plot of Residuals for the BBD Design

b) Plot of Residuals vs Fitted Response
The residuals should scatter randomly suggesting that the variance of the original observations is constant for all values of the response. However, if the variance of the response depends on the mean level of the response, the shape of the plot tends to be funnel-shaped, suggesting a need for a transformation of the response variable.

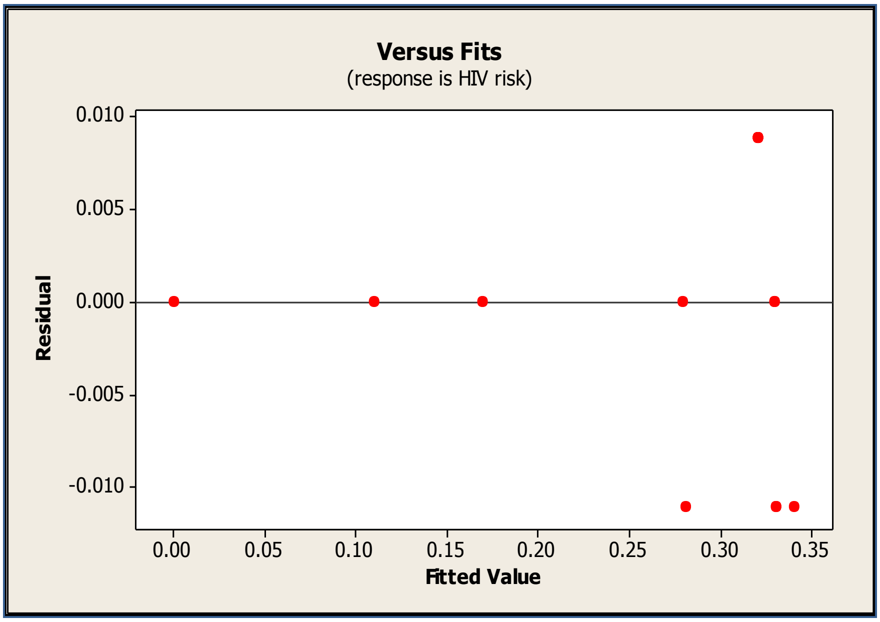

Figure 9: Plot of Residuals vs Fiited Response for the BBD design

The plot of the residuals vs fitted response for the BBD design (Fig. 9) suggests that variance of the original observations is constant.

c) Plot of Residuals vs Observation order
Non random patterns on these plots indicate model inadequacy. This might require transformation to stabilize the situation.

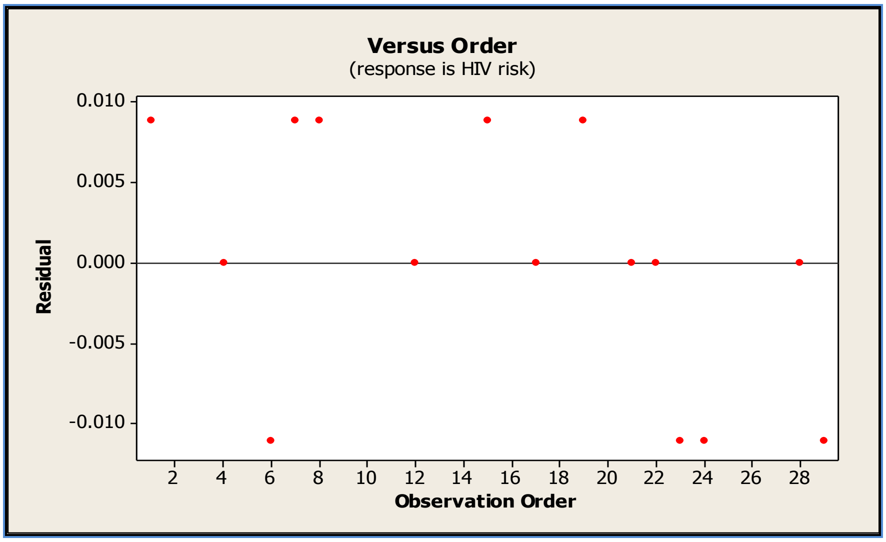

Figure 10: Plot of Residuals vs Observation order for the BBD design

Binary Logistic Regression

The stepwise regression technique on SAS calculates a residual Chi-square score statistic and reports the statistic, its degrees of freedom and the p-value. The residual chi-square is the chi-square score statistic testing the null hypothesis. For the stepwise regression procedure, a residual chi-square test is conducted for each parameter added onto the model.

The null hypothesis is that the addition of each parameter has no effect on the model and this is due to resdual influence. However the results in the Table 14 below indicates that residual effect has no effect on the addition of parameter to the model.

Table14: Residual Analysis

Parameter added	Residual Chi-Square		
	Chi-Square	DF	Pr>Chi-Square
Intercept	684.44	10	<0.0001
Mother's age	79.51	9	<0.0001
Education	40.28	8	<0.0001
Mother's age*Education	14.10	7	0.0495

a) Deviance Residuals

The deviance residual is the measure of deviance contributed from each observation and is given by;

$r_{Di} = sign\ (r_i)\ \sqrt{(d_i)}$ (13)

where D_i is the individual deviance contribution. The deviance residuals can be used to check the model fit at each observation for generalized linear models. Therefore deviance residual is used to identify poorly fitting observations. Observations with a deviance residual in excess of two may indicate lack-of-fit. Fig. 11, shows that there is no lack-of-fit.

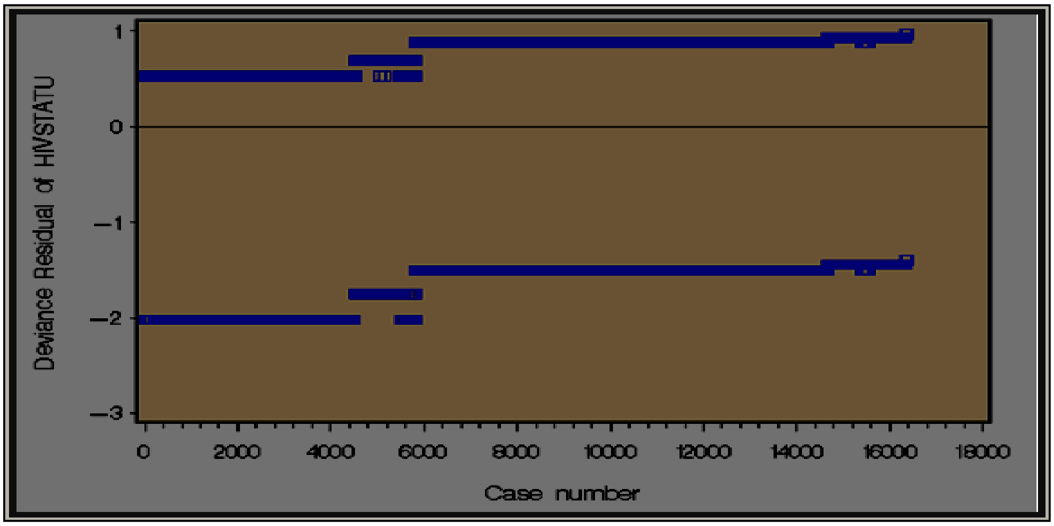

Figure 11: Deviance Residuals from the Logistic regression

b) Pearson Residuals

The Pearson residual is the raw residual divided by the square root of the variance function v (μ). The Pearson residual is the individual contribution to the Pearson χ^2 statistic.

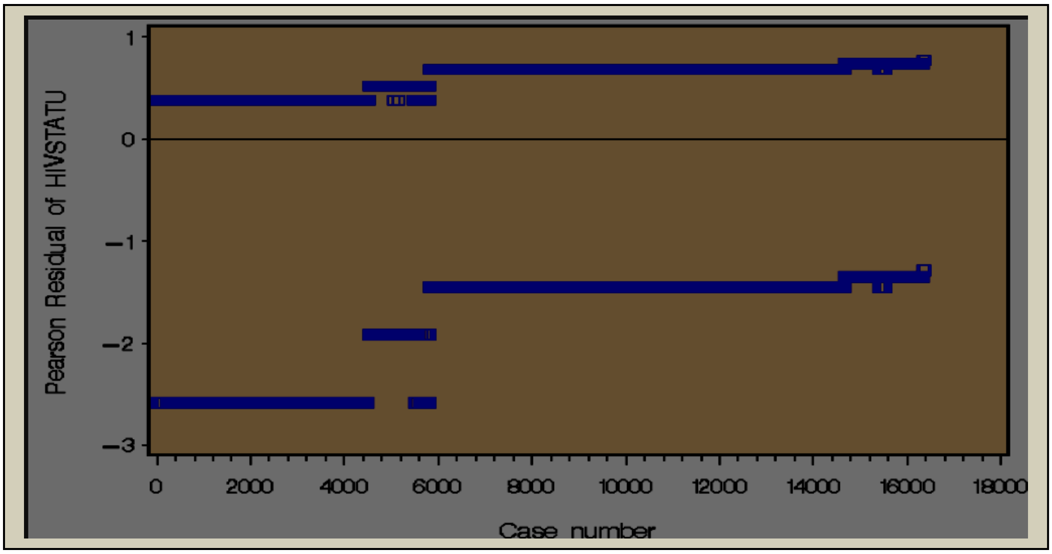

Figure 12: Pearson Residuals from the Logistic regression

Like Deviance residuals, the Pearson residuals can be used to check the model fit at each observation for generalized linear models.

4.3.2 Influence Diagnosis

Parameter estimates or predictions may depend more on the influential subset than on the majority of the data. It is therefore important to locate these influential points and assess their impact on the model. The Leverage of points test was used for the influential diagnosis.

Box Behnken Design

a) Leverage Points
This is a measure of the disposition of points on the x-space. Some observations tend to have disproportionate leverage on the parameter estimates, the predicted values and the summary statistics [5]. Figure 13 shows that the leverage of the points is below one.

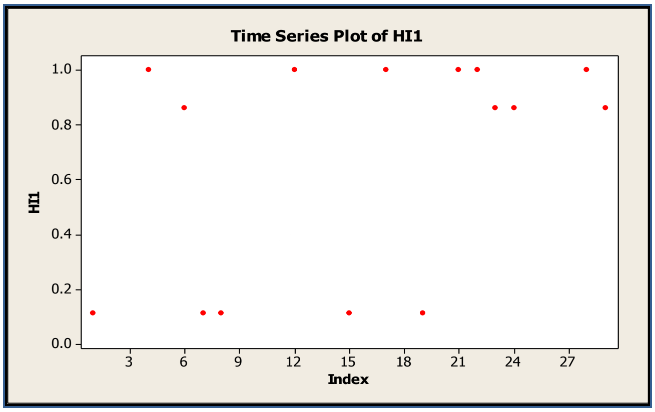

Figure 13: Plot of Leverage of points for the BBD design

Binary Logistic Regression

a) Leverage Plot
The leverage plot (Fig. 14) for the logistic regression shows that very few observations have a disproportionate leverage on the parameter estimates on the predicted values.

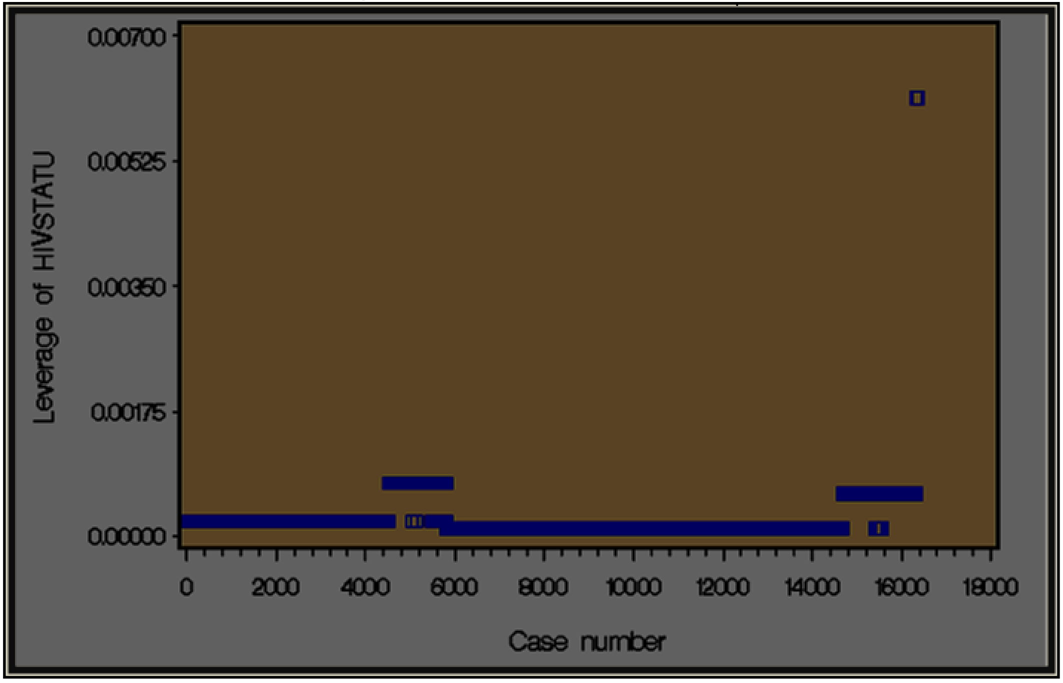

Figure 14: Plot of Leverage Points for the BLR

5. Final Equations of the Response Surface Models

5.1 Box Behnken Design

Table 15: Final Equation from BBD Response Surface

Box Behnken Design HIV =	Factors
+0.32	
+0.18	Mother’s age
-0.02	Father’s age
-0.07	Education
+0.02	Parity
+0.13	Mother’s age*Father’s age
-0.10	Mother’s age*Education
+0.05	Mother’s age*Parity
+0.02	Father’s age*Parity

5.1.1 Main Effects Model

A main effects plot (Fig.16) is a plot of the means of the response variable for each level of a factor, allowing for the determination of which main effects are important. From the main effects plot, it is evident that HIV risk increases steeply as the mother's age and her educational level increase from the low level to the middle level.

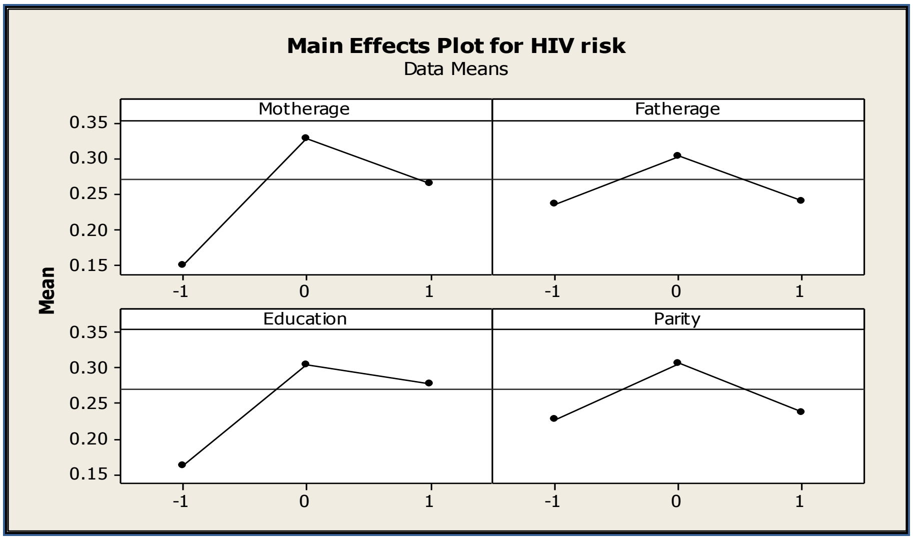

Figure 16: Main Effects Plot

The co-efficient plot derived from the final response surface equation in Table 15, clearly indicates that the mother's age and her educational level are the single most important determinants of the HIV status of an antenatal clinic attendees. Coefficient plots represent the relative importance of each variable on the model equation.

5.1.2 Interaction Effects Model

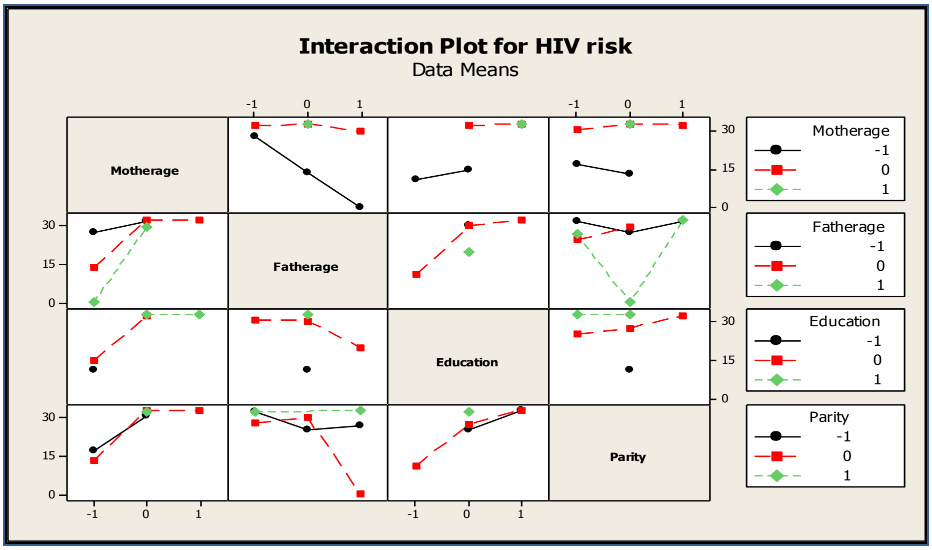

Figure 17: Interactions Plot

Assuming the sparsity-of-effects principle that states that a system is usually dominated by main effects and low-order interactions, an interactions plot as shown in Fig. 17 was generated. On the basis of sparsity-of-effects principle, the research assumed that main effects and two-factor interactions are the most significant responses in this experimental design. In other words, higher order interactions such as three factor interactions are rare. This phenomenon is sometimes referred to as the hierarchical ordering principle.
The interactions plot derived from the design shows that the interaction of the mother's age with the other demographic characteristics has a significant effect on the HIV risk of pregnant mothers. These results are confirmed by the co-efficient plot of the main and interactions effects (Fig. 18).

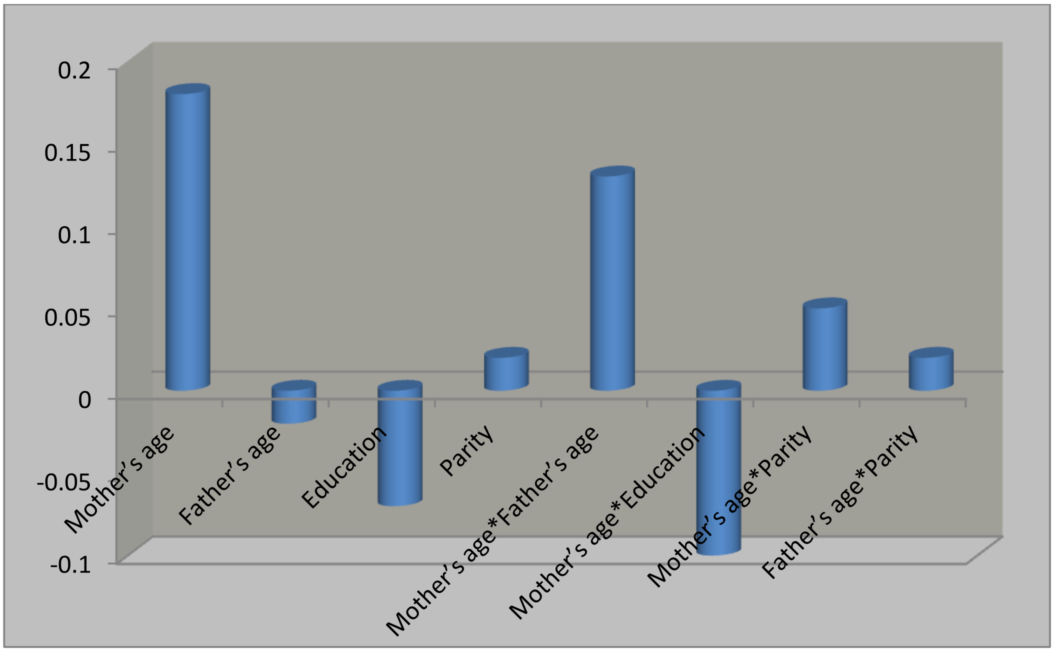

Figure 18: Coefficient Plot of Main and Interaction Effects

5.2 Logistic Regression

5.2.1 Main Effects Model

The main effects model was produced using HIV status as the response variable at two levels, HIV negative (0) and HIV positive (1). The model generated was based on the binary logit with Fisher's scoring as the optimization technique. In total 12 071 HIV positive and 4 312 HIV negative individuals were studied. Maximum likelihood technique was employed to develop estimates of the intercept and the model parameters.

Table 13: Maximum Likelihood Estimates

Parameter	Estimate	Standard Error	Wald $_{\chi}2$	Pr> $_{\chi}2$
Intercept	0.77	0.02	1189	<0.001
Parity	-0.06	0.03	3.65	0.06
Mother's age	-0.76	0.06	144.48	<0.001
Father's age	0.07	0.03	5.64	0.018
Education	-0.28	0.05	28.21	<0.001

The co-efficient plot (Fig. 19) derived from the maximum likelihood estimate, confirms the results obtained by the Box Behnken Design, that the mother's age and her educational level are the single most important determinants of the HIV status of an antenatal clinic attendee.

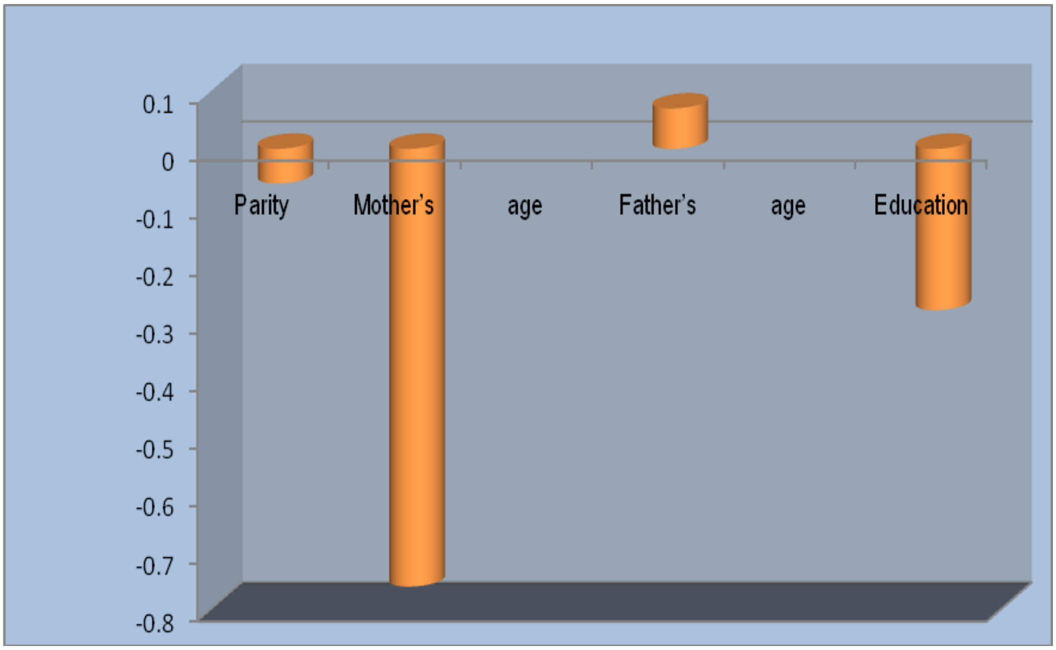

Figure 19: Coefficient Plot of the Main Effects

5.2.2 Interaction Effects Model using Stepwise Regression

However the use of stepwise regression only found the mother's age and her educational level to be worth including in the model at a significance level of 0.05. Furthermore, stepwise regression demonstrated that the interaction of the mother's age and her educational level had a significant effect on the HIV risk.

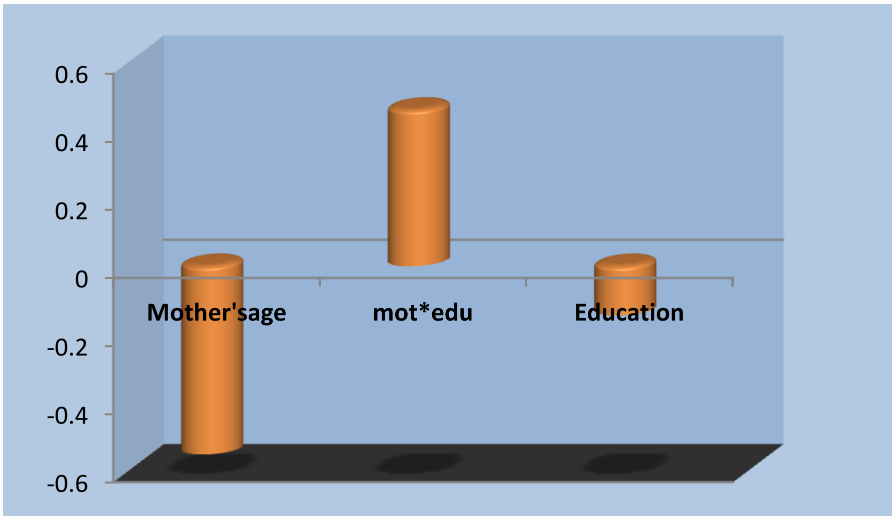

Figure 20: Coefficient Plot of the Main and Interaction Effects based on ML method

6 3D Response Surface plot

Fig. 21 shows the 3D plots of the influences of mother's age and her educational level on the HIV risk of pregnant mothers. The response surface plots indicate that the HIV risk increases with the age of the mother, however the increase in HIV risk is lower for the educated woman compared to their less educated counterparts. The latter observation could be attributed to increased HIV/AIDS awareness in the educated groups.

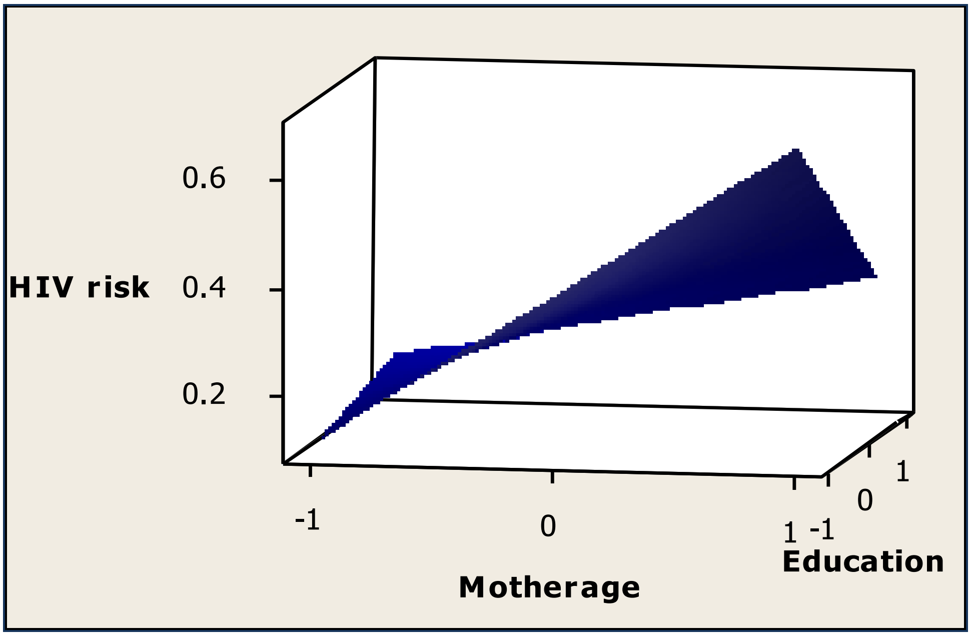

Figure 21: 3D Response Surface plot of BBD Design

7 Discussion

A Box Behnken Design was compared with a Binary Logistic Regression with respect to the capability to determine the effect of demographic characteristics on HIV risk. The two techniques illustrated that the mother's age and her educational level had the greatest effect on her HIV status.

ACKNOWLEDGEMENTS: Wilbert Sibanda acknowledges doctoral funding from South African Centre for Epidemiological Modelling (SACEMA), Medical Research Council (MRC) and North-West University. Special thanks to CathrineTlalengSibanda and the National Department of Health (South Africa) for the antenatal seroprevalence data (2006-2007).

References

[1] Department of Health. National Antenatal Sentinel HIV and Syphilis Prevalence Survey in South Africa, (2010).

[2] Department of Health. Protocol for implementing the National Antenatal Sentinel HIV and Syphilis Prevalence Survey, South Africa, (2010).

[3] W. Sibanda and P. Pretorius, Application of Two-level Fractional Factorial Design to Determine and Optimize the Effect of Demographic Characteristics on HIV Prevalence using the 2006 South African Annual Antenatal HIV and Syphilis Seroprevalence data, *International Journal of Computer Applications*, **35** (12), (2011).

[4] D.C. Montgomery. *Design and Analysis of Experiments*. Wiley, New York, 2008.

[5] R.H Myers and D.C. Montgomery. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Wiley-Interscience, New York, 2002.

[6] B. Mutnury. *Modeling and Characterization of High Speed Interfaces in Blade and Rack Servers Using Response Surface Model*. Electronic Components and Technology Conference (ECTC), IEEE 61st.

[7] Z. Zhang. *Comparison about the Three Central Composite Designs with Simulation*. International Conference on Advanced Computer Control Advanced Computer Control (ICACC '09). (2008).

Design Expert 8.0.71. *StatEase software*.

[8] W. Sibanda, P. Pretorius and A. Grobler, Response Surface Modeling and Optimization to Elucidate the Differential Effects of Demographic Characteristics on HIV Prevalence in South Africa. *IEEE/ACM Interantional Conference on Advances in Social Networks Analysis and Mining*, (2011).

[9] MathWorks, Inc. 1994-2012.

[10] NIST/SEMATECH, *e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/*, (2012).

[11] L.A. Trinca and S.G. Gilmour. Dispersion Variance Dispersion Graphs for Comparing Response Surface Designs with Applications in Food Technology. *Appl. Statist*. 48, (1999), 441-455.

[12] AIDS Epidemic Update. Joint United Nations Programme on HIV/AIDS (UNAIDS) and World Health Organization (WHO), (2009).

[13] P.D. Alison, Logistic Regression using SAS: Theory and Application. SAS Institute Inc, (2009).

[14] D.J Beal, Information Criteria Methods in SAS for Multiple Linear Regression Models, SESUG Proceedings © SESUG Inc.