

Chapter 2: Methods used in published Articles

First Paper

2.4.1. Application of a screening design to rank the effect of demographic characteristics on the risk of acquiring HIV infection

2.4.1.1. Sources of Data

The screening design employed the 2006 South African annual antenatal HIV seroprevalence data. The data consisted of about 33 000 subjects that attended antenatal clinics for the first time across the nine provinces of South Africa. Antenatal surveys are anonymous, unlinked and cross-sectional studies conducted in the public health sector of South Africa (Department of Health, 2010). The choice of the first antenatal visit is made to minimize the chance for one woman attending two clinics and being included in the study more than once. The probability proportion to size sampling method was used to determine the sample size for the 2006 antenatal HIV survey. Provinces with the biggest population sizes of women in the reproductive age yielded the biggest sample sizes (Sibanda & Pretorius 2011:16).

3.4.1.2. Generating the Experimental Design

3.4.1.2.1. Sampling

A random sample of 330 subjects was taken from the 33 034 subjects using SAS 9.1.3 Analytics Platform (SAS Institute Inc., Cary, USA). In this technique each possible sample of n different units out of N has the same probability of being selected. The selection probability was therefore, $330/3304 = 0.00999$. Out of a sample of 330, 323 cases were found to be complete (97.88%), while 7 cases (2.12%) were found to exhibit missing data and thus discarded (Sibanda & Pretorius 2011:16).

3.4.1.2.2. Variables

The variables used in the study were parity, gravidity, education, mother's age, father's age, syphilis and HIV status (Sibanda & Pretorius 2011:16).

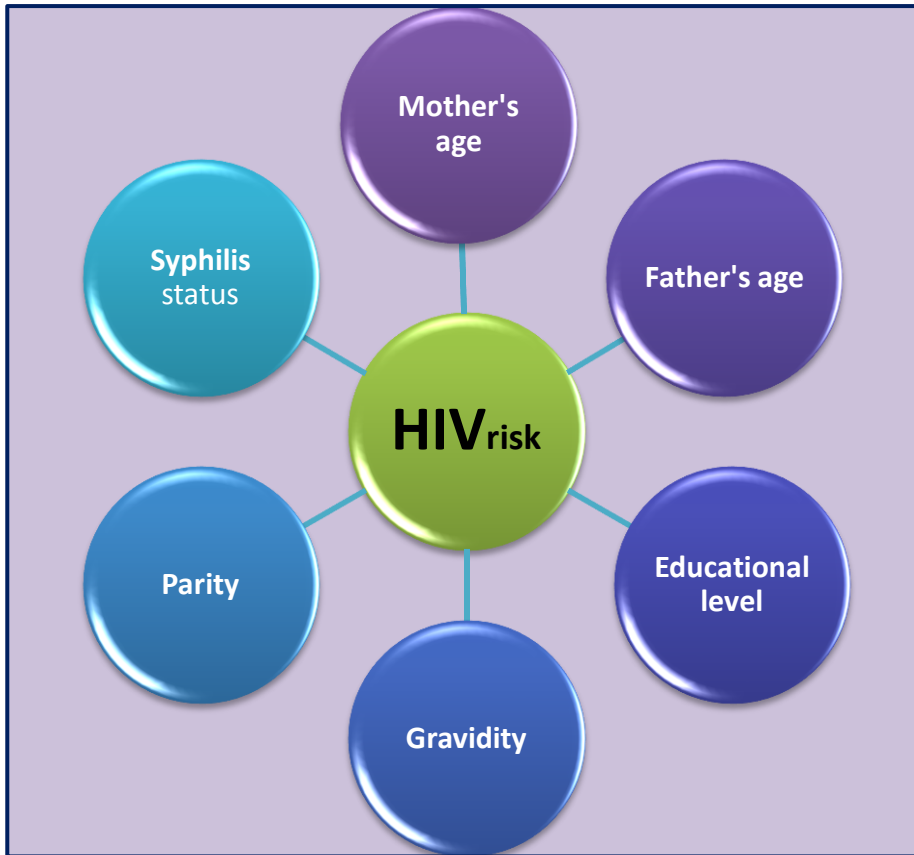


Fig. 3.2: Demographic characteristics studied by the screening design

The integer value representing educational level stands for the highest grade successfully completed, with 13 representing tertiary education. Gravidity as stated above represents the number of pregnancies, complete or incomplete, experienced by a female.

Parity represents the number of times the individual has given birth. Both of these quantities are important as they show 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 (Sibanda & Pretorius 2011:16).

3.4.1.2.3. The Design Matrix

The screening objective, using the fractional factorial design produced the coded design matrix as outlined in Table 1. A design matrix is an experiment that is useful for constructing and analyzing experiments.

Table3.1: The Fractional Factorial design matrix

Exp.No.	Parity	Gravidity	Education	Syphilis	Mother's age	Father's age
1	-1	-1	-1	-1	-1	-1
2	1	-1	1	1	-1	-1
3	0	0	0	0	0	0
4	1	1	-1	1	-1	-1
5	-1	1	-1	1	1	-1

6	-1	-1	1	1	1	-1
7	0	0	0	0	0	0
8	1	-1	-1	1	1	1
9	-1	-1	1	-1	1	1
10	-1	1	1	1	-1	1
11	-1	-1	-1	1	-1	1
12	1	1	-1	-1	-1	1
13	-1	1	-1	-1	1	1
14	1	1	1	-1	1	-1
15	-1	1	1	-1	-1	-1
16	1	-1	-1	-1	1	-1
17	1	1	1	1	1	1
18	1	-1	1	-1	-1	1

The fractional factorial resolution 4 design is highly recommended for studies with more than five factors. Two level (+1 and -1) designs are ideal for screening because these designs are simple and economical and give most of the information required prior to progressing to a multilevel response surface to determine response behavior. The fractional factorial design produced a regression model with seven linear terms as shown in equation 2.6, below.

$$HIV\ risk = b_0 + b_1*Parity + b_2*Gravidity + b_3*Education + b_4*syphilis + b_5*Mother's\ age + b_6*Father's\ age \quad (2.6)$$

The confounding rules were as follows;

- Mother's age = Parity * Gravidity * Education
- Father's age = Gravidity * Education * syphilis

Confounding occurs if the influence of two or more effects is inseparable. In other words, two or more effects are confounded if they use the same linear combination of responses. Confounding of factors is also termed aliasing of main effects (Sibanda & Pretorius 2011:17).

3.4.1.2.4. Selection of Factor Levels

For this current research, the factor levels were generated as shown in Table 3.2.

Table 3.2: Factor Levels

Factor	Levels		Reason
	-1	+1	
Parity	0	>1	41% (~50% of attendees have 0 children)

Gravidity	1	>1	~50% of attendees with 1pregnancy
Educational level (grades	≤ 10	11,12,13	~50% of attendees with ≤ 10 grade
Syphilis status	0	1	
Mother's age (years)	≤ 24	>24	~50% of attendees with ≤ 24 years
Father's age (years)	≤ 28	≥ 28	~50% of attendees with ≤ 28 years

Second Paper

3.4.2. Response surface modeling- Central composite face-centered design

3.4.2.1. 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. 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 (Sibanda & Pretorius 2012: 252-4).

3.4.2.2. Research Tools

This research utilized the following research tools:

- Design Expert V8 Software (StatEase Inc., 2011),
- SAS 9.3, an integrated system of software products provided by SAS Institute Inc.,
- Essential Regression and Experimental Design, version 2.2 (Gibsonia, PA) (Sibanda & Pretorius 2012: 252-5).

3.4.2.3. 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 (Sibanda & Pretorius 2012: 252-5).

3.4.2.4. Missing Data

Out of the total of 31 808 cases from the 2007 South African antenatal HIV seroprevalence database, 21 646 (68%) cases were found to be complete. 10 162 (32%) cases were in complete and thus discarded (Sibanda & Pretorius 2012: 252-5).

3.4.2.5. Variables

The variables used in the study were parity, education, mother's age, father's age and HIV status. Based on the results obtained in the screening design in section 3.4.1, where syphilis was found to be less important in influencing the risk of acquiring HIV, the response surface design

excluded syphilis as a variable. In addition, gravidity was removed because of its observed high correlation with parity (Sibanda & Pretorius 2012: 252-5).

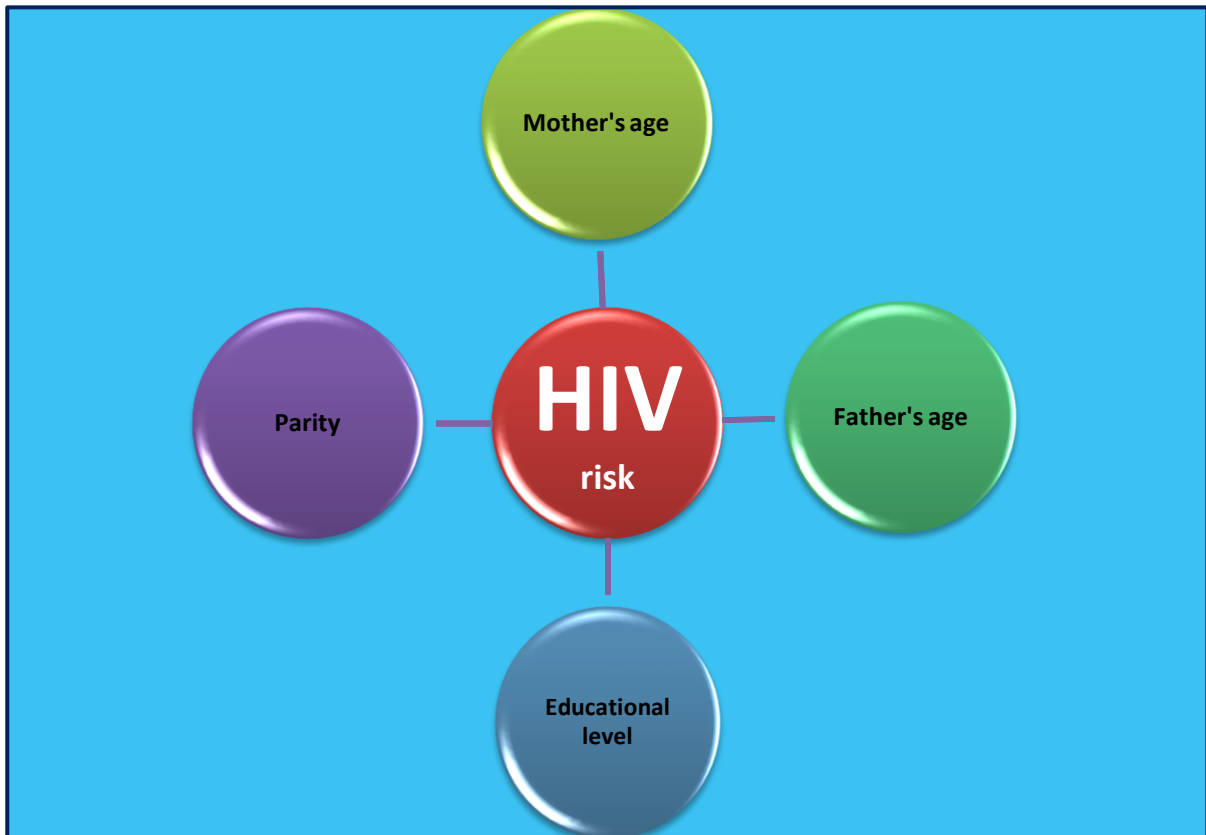


Fig. 3.3: Demographic characteristics studied by the Central Composite Face-centered design

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. The HIV status is binary coded; with 1 representing positive status and a 0 representing a negative status (Sibanda & Pretorius 2012: 252-5).

3.4.2.6. Experimental Design

In this study, the aim was to use a central composite face-centered design to study the individual and interaction effects of demographic characteristics on the HIV status of a pregnant mother using seroprevalence data. The central composite face-centered design with four factors and one response variable was developed. A two factor interaction design model was used, with 21 runs and no blocks. -1 and +1 denote the minimum and maximum levels of factors respectively (Sibanda & Pretorius 2012: 252-5).

Table 3.3: The Central composite matrix design

Run	Input Variables			
	Mother's age	Father's age	Education	Parity
1	1	-1	-1	1
2	0	0	0	0
3	-1	1	-1	1
4	-1	1	1	1
5	0	0	0	0
6	0	1	0	0
7	1	0	0	0
8	0	0	0	0
9	0	0	0	1
10	-1	-1	-1	-1
11	1	-1	1	1
12	-1	-1	1	-1
13	0	0	0	0
14	0	-1	0	0
15	0	0	1	0
16	0	0	0	-1
17	0	0	-1	0
18	-1	0	0	0
19	1	1	1	-1
20	1	1	-1	-1
21	0	0	0	0

3.4.2.6.1. Design matrix evaluation**3.4.2.6.1.1. Degrees of freedom**

In statistics, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary. Estimates of statistical parameters can be based upon different amounts of information or data. The number of independent pieces of information that go into the estimate of a parameter is called the degrees of freedom. In general, the degrees-of-freedom of an estimate of a parameter is equal to the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself. The degrees-of-freedom of the different errors from the design matrix are shown in Table 4.

$$\begin{aligned}
 & \sum(\text{observed value} - \text{fitted value})^2 && (\text{error}) \\
 = & \sum(\text{observed value} - \text{local average})^2 && (\text{pure error}) \\
 & + \sum \text{weight} \times (\text{local average} - \text{fitted value})^2 && (\text{lack of fit}) \\
 & && (2.7)
 \end{aligned}$$

As a rule of thumb, a minimum of 3 lack-of-fit degrees of freedom and 4 pure error degrees of freedom ensure a valid lack-of-fit test. Fewer degrees-of-freedom tend to lead to a test that may not detect lack-of-fit.

Table 3.4: Degrees-of-freedom of different errors

Error	Degrees of freedom
Model	10
Residuals	10
Lack of fit	6
Pure Error	4
Corr total	20

3.4.2.6.1.2. Standard Errors

The standard errors of the design are shown in fig. 3 and these errors are larger at the edges of the design. The flat bottom in this bowl-shaped surface of standard error is very desirable for an RSM design. The standard error plot shows how the variance associated with prediction changes over the design space. The shape depends only on input factors. It is evident that the central composite design provides relatively precise predictions over a broad area around the centre-point. The circular contours indicate the statistical property of rotatability, another desirable feature of central composite design (Sibanda & Pretorius 2012: 252-6).

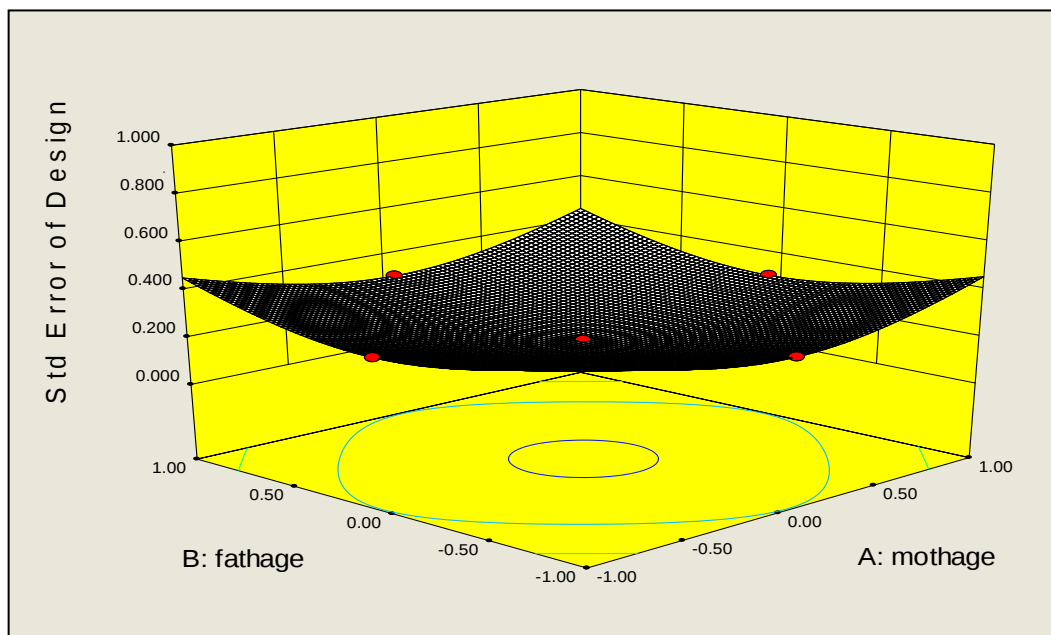


Fig. 3.4: Standard errors

3.4.2.6.1.3. Fraction of Design Space (FDS)

At the early stages of experimentation, screening and characterization, where factorials are the design of choice, the emphasis is on identifying main effects and interactions. For this purpose, power is an ideal metric to evaluate design suitability. However, when the goal is optimization,

as is the case in response surface modeling, the emphasis shifts to creating a fitted surface within a desired precision. Then fraction of design space (FDS) becomes a powerful tool for grading design suitability (Sibanda & Pretorius 2012: 252-8).

In simple terms, the FDS curve 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. The standard error of the predicted mean response at any point in the design space is a function of three things:

- i. the experimental error, expressed as standard deviation,
- ii. the experimental design,
- iii. where the point is located in the design space i.e. its coordinates (Sibanda & Pretorius 2012: 252-8).

The FDS plot for this specific design is shown in Fig. 13.

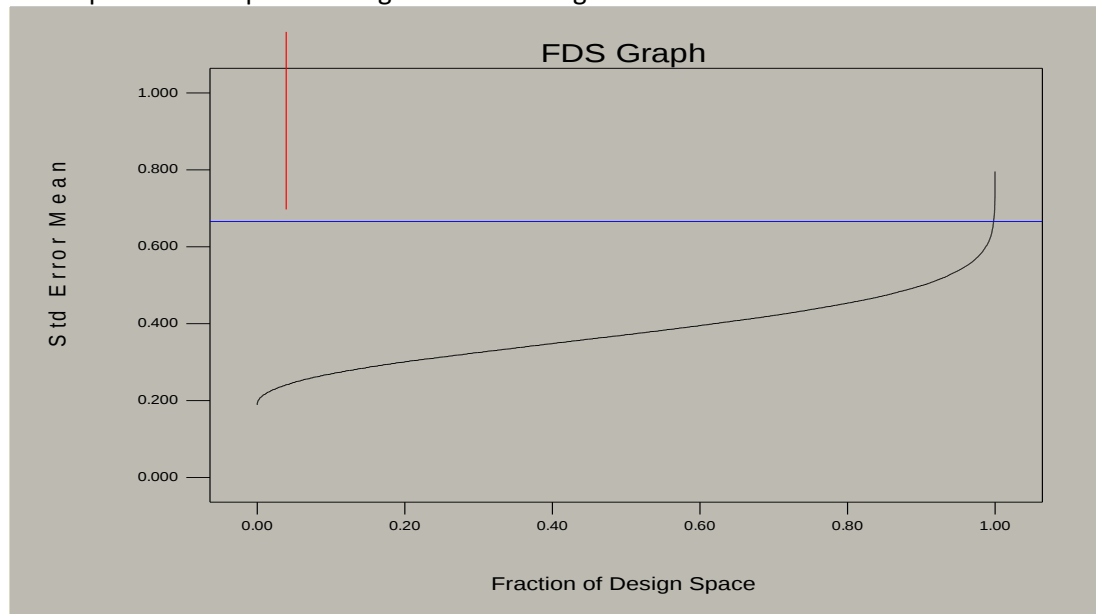


Fig. 3.5: Fraction of design space (FDS) plot of the standard error the design space

3.4.2.6. Choice of Levels for the factors

Quite often the appropriate selection of factor levels turns out to be a crucial problem, as the respective selection might determine whether an existing causal relation is detected by the study or not. In the case of a continuous independent variable, such as dose of a drug, it is possible to randomly select a given number of levels from the effective range of the drug. If the levels of a factor are selected by this kind of procedure the factor is called a random factor.

This kind of procedure may also be used for discrete independent variables with a large number of discrete levels. However, random factors are not typical of behavioural research. The experimenter rather tries to select the factor levels according to more or less rational criteria, thus

obtaining so-called fixed-factors (Sibanda & Pretorius 2012: 252-9). Factor levels should always be chosen such that an apparent difference in the dependent variable can be expressed for any two adjacent levels. If the functional relationship is being investigated between an independent and a dependent variable, there should be more levels in those regions where maxima or minima of the curve can be expected (Sibanda & Pretorius 2012: 252-9).

Table 3.5: Factor Levels

Factor	Levels		
	-1	0	1
Parity (No. of children)	0	1	≥ 2
Education (Grades)	≤ 8	9-11	12-13
Mother's age (years)	≤ 20	21-29	≥ 30
Father's age (years)	≤ 24	25-33	≥ 34

Third Paper

3.4.3. Comparison of Central Composite Face-centered (CCF) and Box-Behnken Designs (BBD) to study the effect of demographic characteristics on HIV risk in South Africa (Sibanda & Pretorius 2013)

3.4.3.1. 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. 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 (Sibanda & Pretorius 2013: 3).

3.4.3.2. Research Tools

This research utilized the following research tools:

- Design Expert V8 Software (Stat-Ease Inc., 2011),
- SAS 9.3, an integrated system of software products provided by SAS Institute Inc.,
- Minitab 16. Minitab Inc., United States (Sibanda & Pretorius 2013: 3).

3.4.3.3. Data Processing

The data was completely randomized, and this process was undertaken as a preprocessing technique to reduce bias in the design of experiment (Sibanda & Pretorius 2013: 3).

3.4.3.4. 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 (Sibanda & Pretorius 2013: 3).

3.4.3.5. Variables

The variables used in the study were mother's age, father's age, education, parity and HIV status (Sibanda & Pretorius 2013: 3).

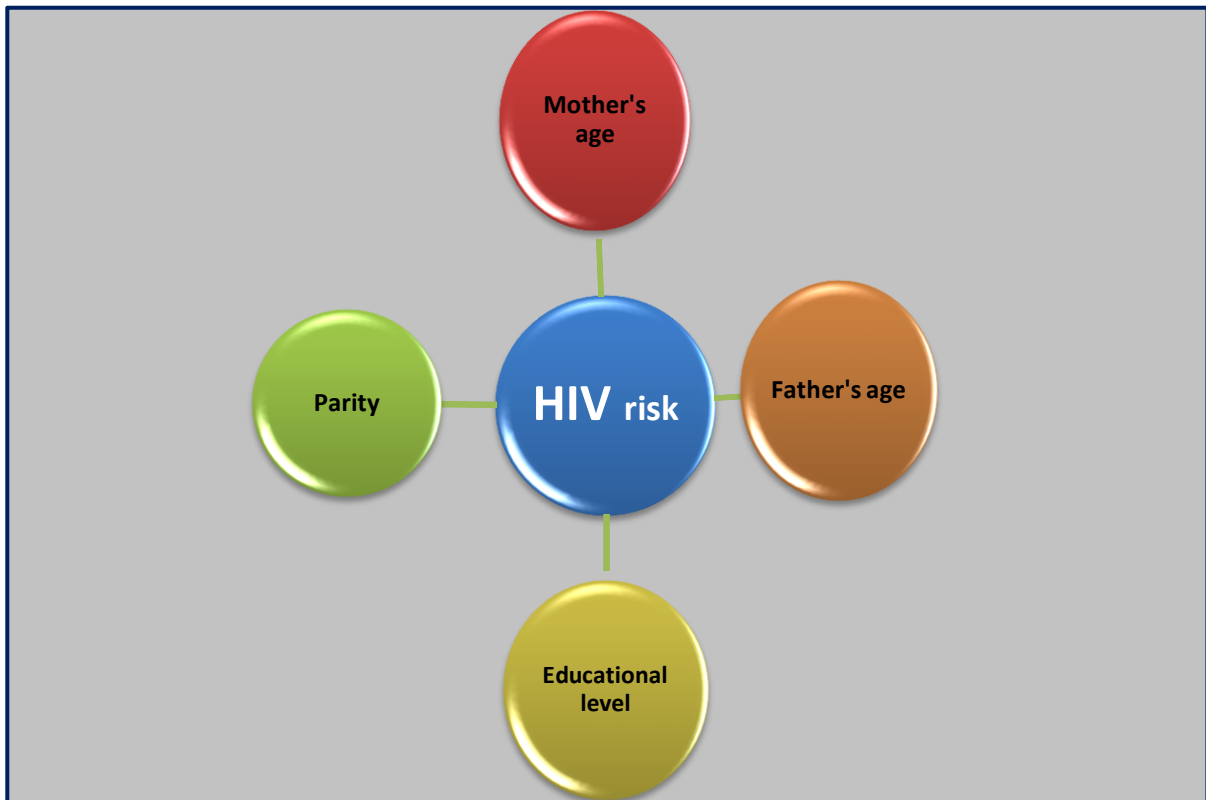


Fig. 3.6: Demographic characteristics studied by the Central Composite Face-centered design and Box-Behnken designs

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 (Sibanda & Pretorius 2013: 3).

3.4.3.6. Experimental Design

In this study, the aim was to use a Central Composite Face Centered (CCF) and a Box-Behnken design (BBD) to study the individual and interaction effects of demographic characteristics on the HIV status of a pregnant mother using seroprevalence data (Sibanda & Pretorius 2013: 4).

The central composite face-centered and Box-Behnken designs with four factors and one response variable were developed as shown in Tables 3.6 and 3.7. Based on sparsity-of-effects principle, two factor-interaction (2FI) design models were used, with 21 and 29 runs for the

central composite face-centered and Box-Behnken designs respectively. No blocking was used. -1 and +1 denote the minimum and maximum levels of factors respectively (Sibanda & Pretorius 2013:4).

Table 3.6: Central Composite Face-centered Design

	Factors			
Run	Mother's age	Father's age	Education	Parity
1	1	-1	-1	1
2	0	0	0	0
3	-1	1	-1	1
4	-1	1	1	1
5	0	0	0	0
6	0	1	0	0
7	1	0	0	0
8	0	0	0	0
9	0	0	0	1
10	-1	-1	-1	-1
11	1	-1	1	1
12	-1	-1	1	-1
13	0	0	0	0
14	0	-1	0	0
15	0	0	1	0
16	0	0	0	-1
17	0	0	-1	0
18	-1	0	0	0
19	1	1	1	-1
20	1	1	-1	-1
21	0	0	0	0

Table 3.7: Box-Behnken

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

3.4.3.7. Design Matrix Evaluation

3.4.3.7.1. 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 3.8. As a rule of thumb, a minimum of 3 lack-of-fit degrees of freedom and 4 pure error degrees of freedom ensure a valid lack of fit test. Fewer degrees of freedom tend to lead to a test that may not detect lack of fit (Sibanda & Pretorius 2013:4).

Table 3.8: Degrees of freedom of different errors

	Central composite face-cent design	Box-Behnken design
Model	10	10
Residuals	19	18
Lack-of-fit	14	14
Pure error	5	4
Corr Total	29	28

3.4.3.7.2. Standard errors

The standard errors of the CCF and BBD designs are shown in fig. 3.7(a) and (b), respectively. The BBD design has larger standard errors at the edges of the design space compared to the CCF and this could be attributed to the architecture of the designs and hence the BBD design is not capable of estimating the response parameter at the edges of the experimental space (Sibanda & Pretorius 2013: 4).

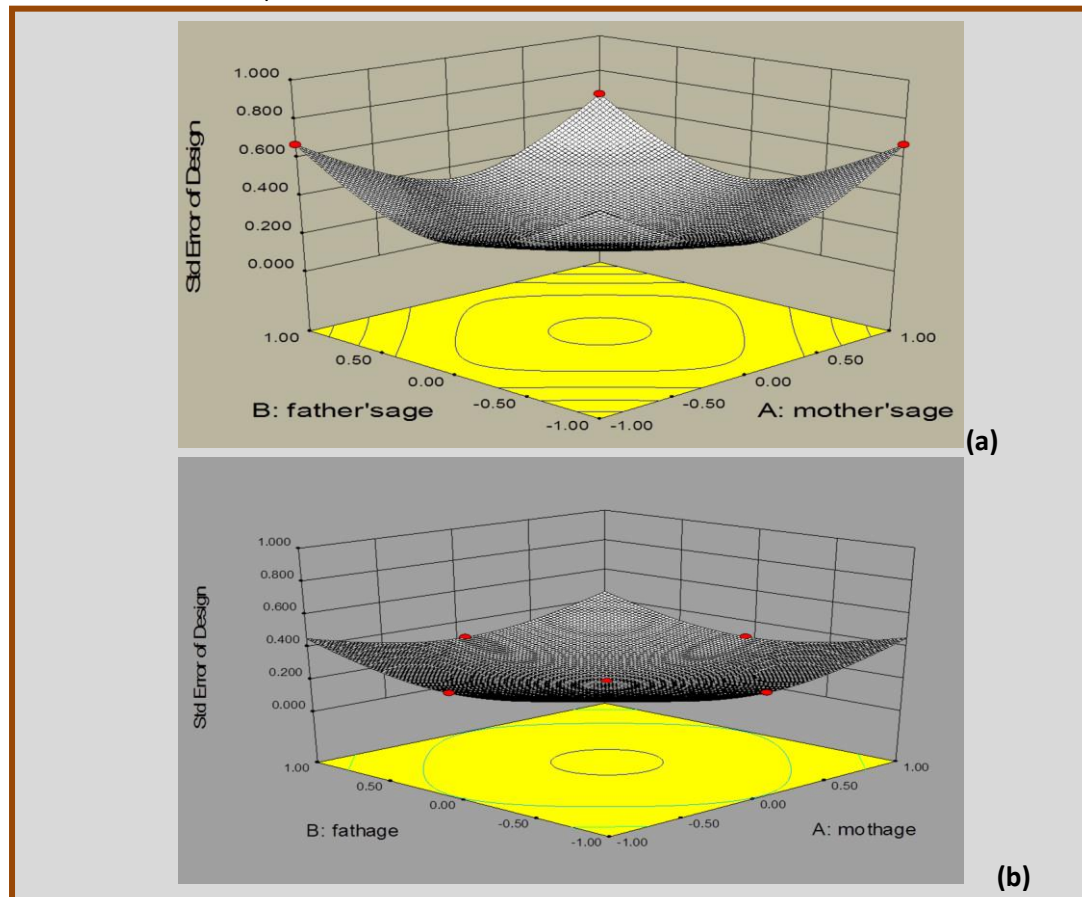
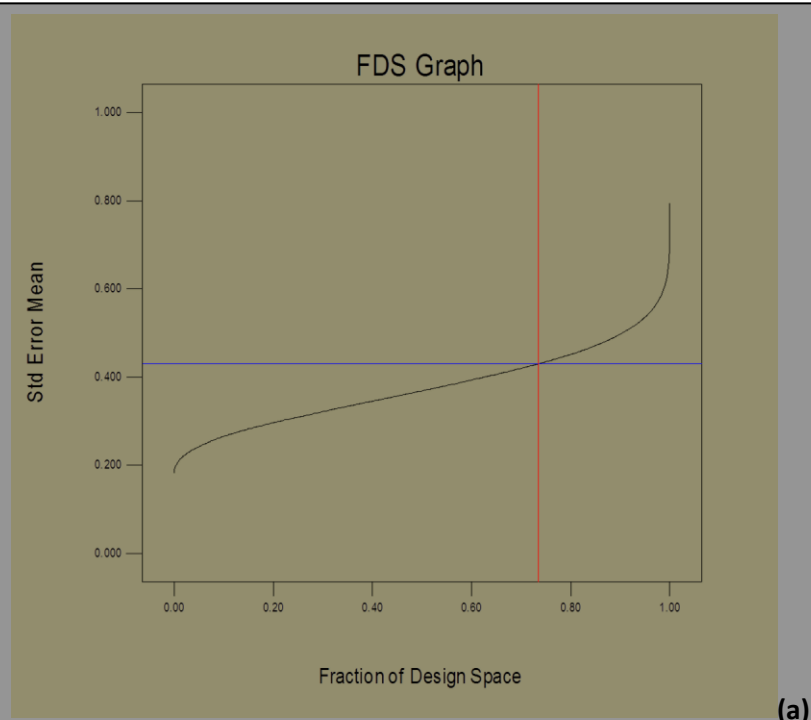


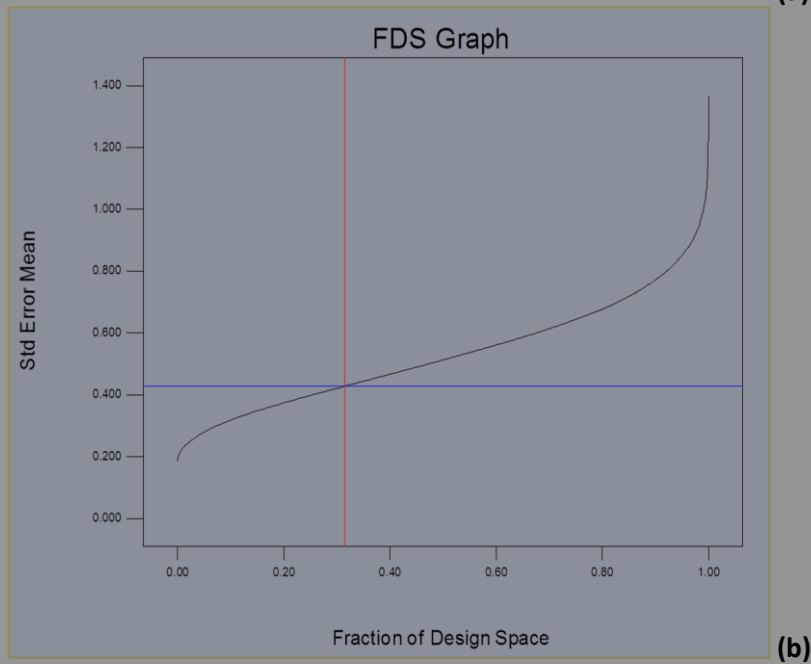
Fig. 3.7 (a and b): Standard error plot of the CCF and BBD designs respectively.

3.4.3.7.3. Fraction of Design Space

As stated above, FDS curve 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 (Sibanda & Pretorius 2013: 4).



(a)



(b)

Fig. 3.8 (a) and (b): FDS plots of standard errors of the central composite face-centered and Box-Behnken design spaces, respectively.

From Figs. 3.8 (a) and (b), it can deduced that only 73% of the central composite face-centered design space is precise enough to predict the mean within ± 0.90 , compared to 31% of the Box-Behnken (Sibanda & Pretorius 2013: 4).

3.4.3.7.4. Variance dispersion graphs (VDGS)

VDGs have recently become popular in aiding the choice of a response surface design. In addition, VDGs can be used to compare the performance of multiple design models such as linear models, linear models with interaction terms, linear models with quadratic terms or for full quadratic models. VDGs were developed by Giovannitti-Jensen and Myers in 1989 (Sibanda & Pretorius 2013: 4).

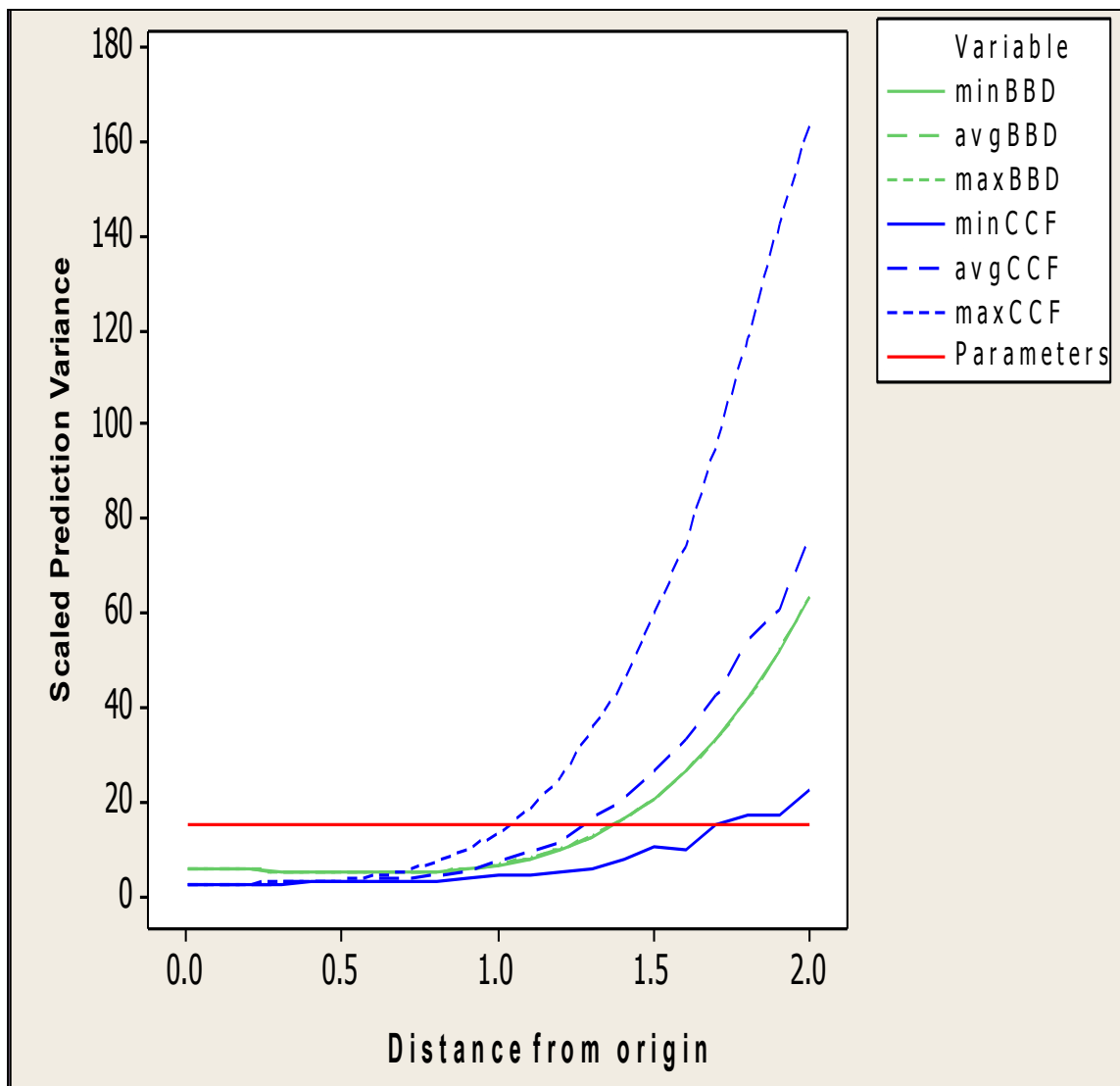


Fig. 3.9: VDG graphs of CCF and BBD designs

Comparison of the VDG graphs (**Fig. 3.9**) of the CCF and BBD designs illustrates the following points:

- i. Box-Behnken design has a greater deviation from rotatability than the central composite face-centered design.
- ii. Variance is close for the two designs near the center.
- iii. Variances of the central composite face-centered and the Box-Behnken designs between radii 0.5 and 2 appear to be better for the central composite face-centered design.
- iv. Maximum variance for the central composite face-centered and the Box-Behnken designs are 60 and 160 respectively (Sibanda & Pretorius 2013: 4).

3.5. Choice of factor levels for the CCF and Box-Behnken designs

TABLE 3.9: Factor Levels

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

Fourth Paper

3.4.4. Comparative study of the application of Box Behnken designs (BBD) and Binary Logistic Regression (BLR) to study the effect of Demographic Characteristics on HIV Risk in South Africa (Sibanda, & Pretorius 2012b).

3.4.4.1. Sources of data

This phase of the study sourced data from the 2007 South African antenatal data, supplied by the National Department of Health of South Africa. The data consisted of about 32 000 sub-

jects that attended antenatal clinics for the first time across the nine provinces of South Africa in 2007 (Sibanda & Pretorius 2012b).

3.4.4.2. Research Tools

This research utilized the following research tools:

- a) Design Expert V8 Software (StatEase Inc., 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 (Sibanda & Pretorius 2012b).

3.4.4.3. Data preprocessing

This involved completely randomizing the data to reduce bias in the design of experiment (Sibanda & Pretorius 2012b).

3.4.4.4. 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 (Sibanda & Pretorius 2012b).

3.4.4.5. Variables

The variables used in the study were mother's age, father's age, education, parity and HIV status as shown in Fig. 3.10 (Sibanda & Pretorius 2012b).

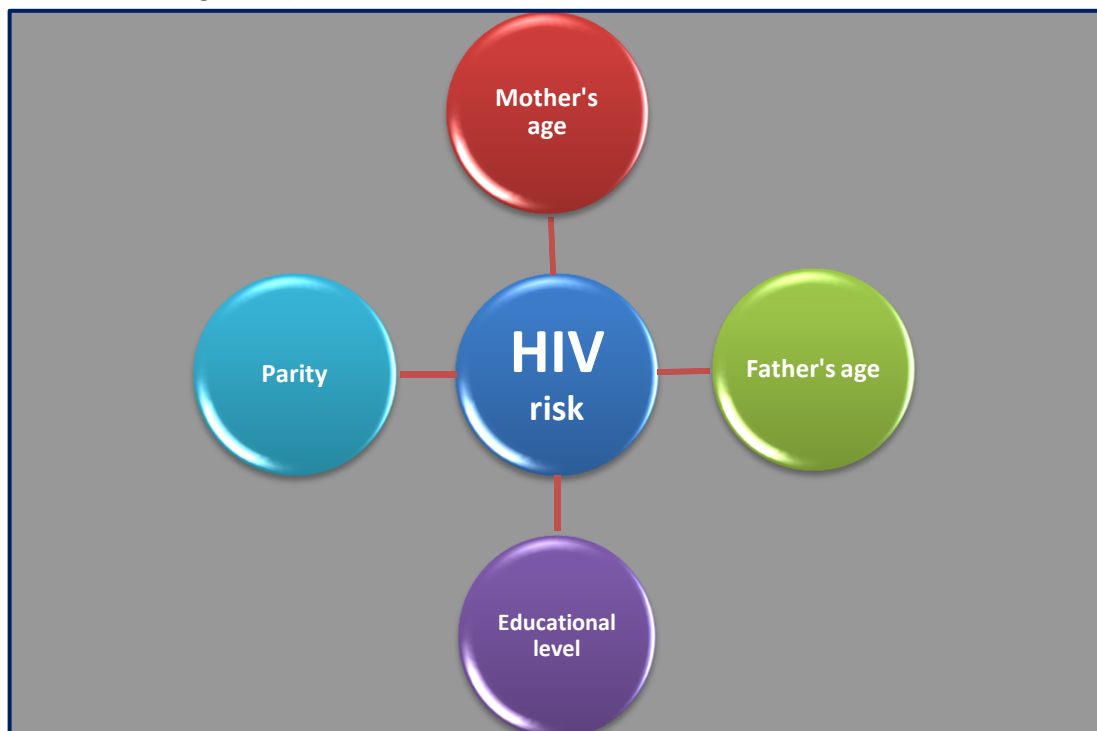


Fig. 3.10: Demographic characteristics studied by the Box-Behnken design and binary logistic regression methodology

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.

3.4.4.6. Experimental Design

In this study, the aim was to compare the outcomes from modeling the main and interaction effects of demographic characteristics using a Box-Behnken design (BBD) and a binary logistic regression (BLR) technique. A BBD design with four factors and one response variable was developed as shown in Table 3.10. 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 (Sibanda & Pretorius 2012b).

Table 3.10: The BBD Matrix Design with 4 factors, 1 response variable and 4 center points.

Run	Factors			
	Mother's age	Father's age	Education	Parity
1	0	0	0	0
2	0	1	-1	0
3	1	0	-1	0
4	1	0	1	0
5	-1	0	1	0
6	0	-1	0	1
7	0	0	0	0
8	0	0	0	0
9	-1	0	0	1
10	0	-1	1	0
11	0	-1	-1	0
12	-1	0	0	-1
13	0	0	-1	-1
14	0	0	-1	1
15	0	0	0	0
16	1	0	0	1
17	-1	-1	0	0
18	0	0	1	1
19	0	0	0	0
20	1	0	0	-1
21	0	0	1	-1
22	-1	1	0	0
23	0	1	0	1
24	0	-1	0	-1
25	1	-1	0	0

26	1	1	0	0
27	0	1	1	0
28	-1	0	-1	0
29	0	1	0	-1

3.4.4.7. Design matrix evaluation

3.4.4.7.1. 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 3.11.

As stated above, as a rule of thumb, a minimum of 3 lack-of-fit degrees of freedom and 4 pure error degree of freedom ensure a valid lack of fit test. Fewer degrees of freedom tend to lead to a test that may not detect lack of fit (Sibanda & Pretorius 2012b).

Table 3.11: Degrees of freedom for Box-Behnken design matrix

Model	10
Residuals	18
Lack-of-fit	14
Pure error	4
Corr Total	28

3.4.4.7.2. Standard errors

The standard errors of the Box-Behnken design are shown in fig. 3.11. The Box-Behnken design has large standard errors at the edges of the design space however this 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 (Sibanda & Pretorius 2012b).

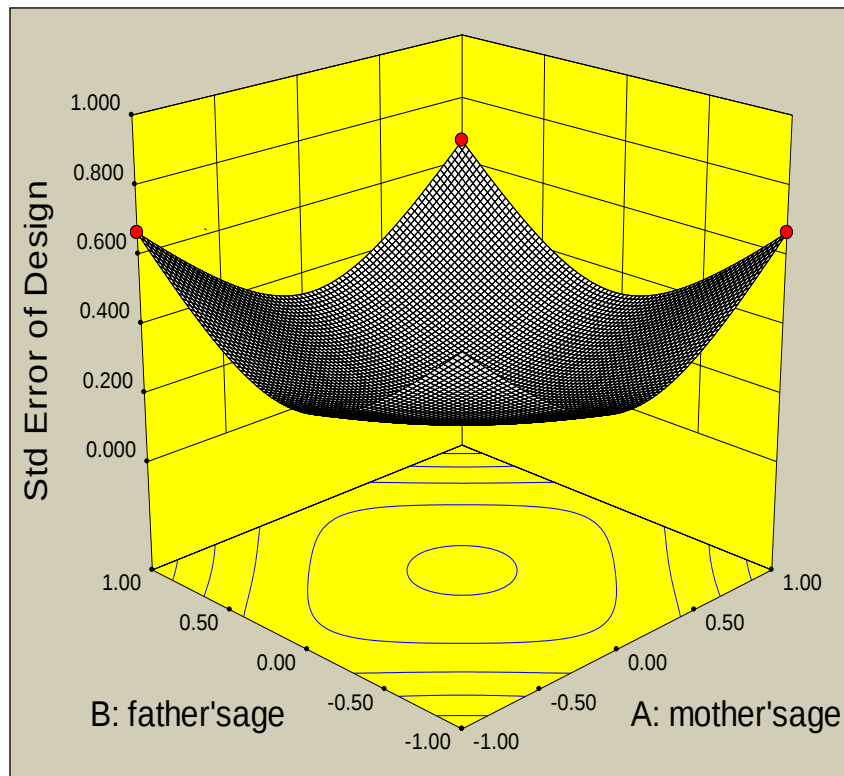


Fig. 3.11: 3D Plot of standard errors of the Box-Behnken design

3.4.4.7.3. Fraction of design space

FDS curve (Fig. 3.12) 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 (Sibanda & Pretorius 2012b).

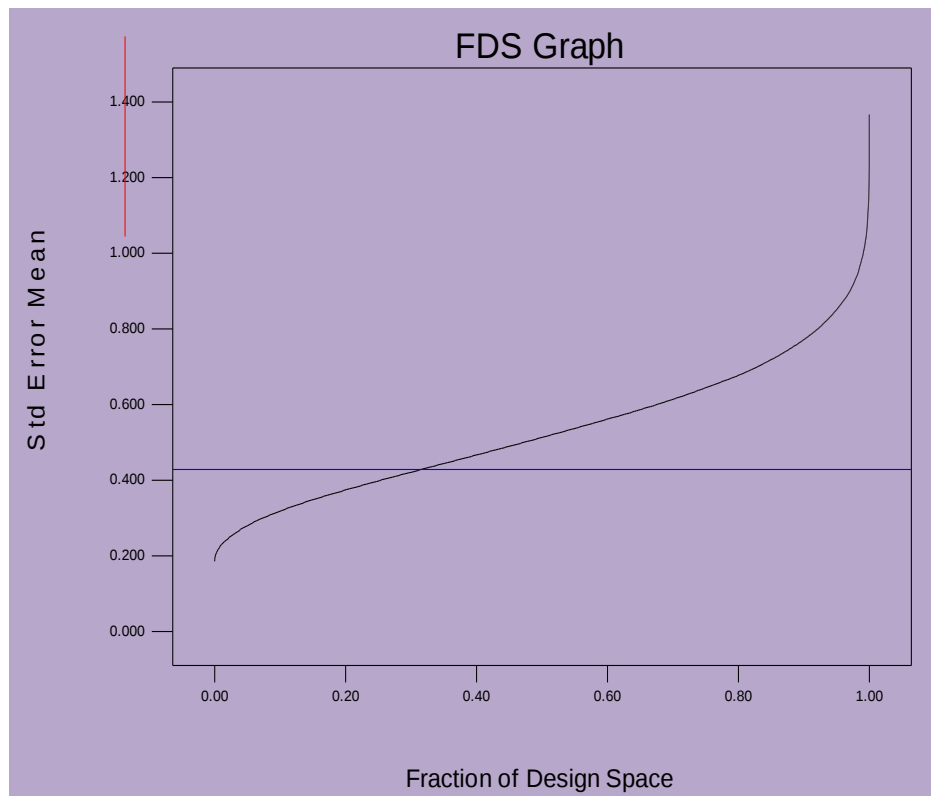


Fig. 3.12: FDS plot of the standard error over the BBD design space

From Fig. 3.12, it can be deduced that only 31% of the BBD design space is precise enough to predict the mean (Sibanda & Pretorius 2012b).

3.4.4.7.4. Choice of levels for the factors

Table 3.12: Factor Levels

Factors	Levels		
	-1	0	1
Mother's age	< 20	21-29	> 30
Father's age	< 24	25-33	> 34
Education	< 8	9-11	12-13
Parity	0	1	> 2

3.4.4.8. Data analysis methods

3.4.4.8.1. Goodness-of-fit

As discussed in the introduction to this thesis, deviance and Pearson's chi-squared goodness-of-fit will be 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, will be used to measure goodness-of-fit for logistic regression models (Sibanda & Pretorius 2012b).

3.4.4.8.2. Hypothesis testing

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. Like AIC, SC penalizes for the number of predictors in the model and the smallest SC is most desirable. In addition, three hypothesis testing techniques, namely Likelihood ratio (LR), Score and Wald Tests will be used to 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 (Sibanda & Pretorius 2012b).

3.4.4.9. Model adequacy checking

Model adequacy checking will be 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. A residual plot assists 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 (Sibanda & Pretorius 2012b).

3.4.4.9.1. 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. In addition, normal probability plots of residuals evaluate whether there are outliers in the dataset. Ideally, all points should lie on the diagonal, implying that the residuals constitute normally distributed noise (Sibanda & Pretorius 2012b).

3.4.4.9.2. 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 (Sibanda & Pretorius 2012b).

3.4.4.9.3. Deviance residuals

The deviance residual is the measure of deviance contributed from each observation. It is therefore, used to detect ill-fitting covariate patterns (Sibanda & Pretorius 2012b).

3.4.4.9.4. Pearson's residuals

Pearson residuals are used to detect ill-fitting covariate patterns. It is the raw residual divided by the square root of the variance function. The Pearson residual is the individual contribution to the Pearson χ^2 statistic. Like Deviance residuals, the Pearson residuals can be used to check the model fit at each observation for generalized linear models (Sibanda & Pretorius 2012b).

3.4.4.9.5. 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. This is a measure of the disposition of points on the x-space. Some observations tend to have disproportionate leverage on the parameter estimates (Sibanda & Pretorius 2012b).

3.4.4.10. Main effects model

A main effects plot 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. Assuming the sparsity-of-effects principle that states that a system is usually dominated by main effects and low-order interactions, interactions plot were also generated (Sibanda & Pretorius 2012b).

Fifth Paper

3.4.5. Novel application of multilayer perceptrons (MLP) neural networks to model HIV in South Africa using seroprevalence data from antenatal clinics (Sibanda, & Pretorius 2011b)

3.4.5.1. Design of study

The multilayer perceptron was used to study the relationship between the likelihood of being HIV positive or negative. Questions that needed to be addressed included:

- What is the appropriate neural network architecture for this particular data set?
- How robust is the neural network performance in predicting the HIV status in terms of sampling variability (Sibanda & Pretorius 2011b)?

For the first question, there are no definite rules to follow since the choice of the architecture also depends on the classification objective. For example, if the objective is to classify a given set of objects as well as possible, then a larger network may be desirable. However, if the network is to be used to predict the classification of unseen objects, then a larger network is not necessarily better. For the second question, a cross-validation approach was used to investigate the robustness of the neural networks in HIV status prediction (Sibanda & Pretorius 2011b).

3.4.5.2. Demographic characteristics studied

This study utilizes a total of six quantitative demographic characteristics as variables, namely parity, gravidity, mother's age, father's age, educational level and syphilis. The qualitative characteristics such as race and province were not included in this study (Sibanda & Pretorius 2011b).



Fig. 3.13: Demographic characteristics studied by MLP

Table 11: Specifications of demographic characteristics

Table 3.13: Specifications of variables for the MLP technique

Demographic characteristics	Specifications
Mother's age	13-45
Father's age	15-55
Education	0-13
Syphilis	0 (negative) & 1 (positive)
Gravidity	0-10
Parity	0-8
HIV status	0 (negative) & 1 (positive)

3.4.5.3. Five-fold validation

In 5-fold cross-validation, the original sample is randomly partitioned into 5 equal size subsamples. Of the 5 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 4 subsamples are used as training data. The cross-validation process is then repeated 5 times, with each of the 5 subsamples used exactly once as the validation data. The 5 results from the folds were then averaged to give single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once (Sibanda & Pretorius 2011b).

Table 3.14: Data Tagging

Group	Description
Training	Data used by the neural network to learn from
Cross validation	Data used to evaluate performance during learning
Testing	Data used to evaluate performance after training

3.4.5.4. Sensitivity analysis

Sensitivity analysis assesses the effect that each of the network inputs has on the network output, thus providing a feedback as to which input channels are the most significant. Sensitivity analysis provides an opportunity to prune the input space by removing the insignificant channels, reducing the size and complexity of the network. Sensitivity analysis is therefore a method for extracting the cause and effect relationship between the inputs and outputs of the network (Sibanda & Pretorius 2011b).