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

Wilbert Sibanda

North-West University, VAAL Triangle Campus
VAN ECK BLVD, VANDERBIJLPARK, 1900,
South Africa

Philip Pretorius

North-West University, VAAL Triangle Campus
VAN ECK BLVD, VANDERBIJLPARK, 1900,
South Africa

ABSTRACT

A Two-Level Fractional Factorial design was employed to develop and optimize the combination of demographic characteristics that has the greatest effect on the spread of HIV in the South African population. HIV prevalence (dependent variable) was found to be highly sensitive to changes in the mother's age (15-55years) and level of education (Grades 0-13) (independent variables), using the 2006 South African Annual Antenatal HIV and Syphilis Seroprevalence data.

HIV prevalence was the optimization objective (R^2 = 0.842, Durbin-Watson Index = 3.440%, Coefficient of Variation (CV) = 52.169%). The Lagrangian technique produced no significant difference (P>0.05) between the experimental and predicted HIV prevalence values.

General Terms

Two level fractional factorial design

Keywords

Screening design, fractional factorial design, demographic characteristics, seroprevalence data.

1. INTRODUCTION

In South Africa, the National Department of Health (DoH) instituted a mechanism to annually monitor the HIV epidemic since 1990, through annual, nation-wide HIV and Syphilis seroprevalence surveys among pregnant women attending public sector antenatal clinics (Department of Health, 2010).

The purpose of the survey is to obtain information on the prevalence of the HIV and syphilis infection and to monitor trends over time. The annual antenatal HIV survey is the only existing national surveillance activity for determining HIV prevalence in South Africa and is therefore a vitally important tool to track the geographic and temporal trends of the epidemic (Department of Health, 2010).

Antenatal clinic data contains the following demographic characteristics for each clinic attendee; Mother's age (Mothage), population group (race), level of education (education), gravidity (number of pregnancies), parity (number of children born), partner's age (Fathage), name of hospital/clinic, HIV and syphilis (measured using RPR technique) results.

This study attempts to utilize experimental design techniques to develop a ranked list of important through unimportant demographic factors that affect the spread of HIV in the South African population. Finally, the study aims to develop an optimized combination of demographic characteristics that can result in the highest and or lowest prevalence of HIV. To achieve this objective, a resolution 4 fractional factorial design was employed. The latter constrained optimization technique will provide an efficient and economical method to acquire the necessary information to understand the relationship between the controllable and performance variables Benferoni et al., 2000; Furlanetto et al., 2003; Younes, Fotheringham and El-Dessouky, 2010.

2. LITERATURE REVIEW

A factorial design is an experiment in which only an adequately chosen fraction of the experimental combinations required for the complete factorial experiment is selected to be run.

Properly chosen fractional factorial designs for two-level experiments have the desirable properties of being both balanced and orthogonal. The main effects postulate states that single-factor effects tend to dominate two-factor interactions, and these in turn dominate three-factor interactions. Fractional design makes use of the above postulate by sacrificing high order interations to economize experiments. A major disadvantage of fractional factorial design is its failure to identify which factors or interactions control a process. Reducing the completeness of an experimental design leaves particular interactions invisible to later analysis.

The basic purpose of a fractional factorial design is to economically investigate cause-and-effect relationships of significance in a given experimental setting.

In general, designs of resolution three, and sometimes four, seek to screen out the few important main effects from the many less important others. On the other hand, designs of resolution five, and higher, are used for focusing on more than just main effects in an experimental situation. These designs enable the estimation of interaction effects and such designs are uagmented to a second-order design. Resolution refers to the lowest order interaction that is aliased with a single design factor.

3. FRACTIONAL FACTORIAL DESIGNS AND THEIR PROPERTIES

Fractional factorial designs are experimental designs consisting of a carefully chosen fraction of the experimental runs of a full factorial design. The fraction is chosen so as to exploit the sparsity-of-effects principle to expose information about the most important features of the problem studied, while using a fraction of the effort of a full factorial design in terms of experimental runs and resources.

Fractional designs are expressed using the notation $\mathbf{I^{k-p}}$, where **I** is the number of levels of each factor investigated, *k* is the number of factors investigated, and *p* describes the size of the fraction of the full factorial used.

A fractional factorial design is generated from a full factorial experiment by choosing an alias structure. The alias structure determines which effects are confounded with each other. Another important property of fractional factorial design is its resolution or the ability to separate main effects and low-order interactions from one another. Therefore, the resolution of the design is the minimum word length in the defining relation.

A quarter-fraction design, denoted as $\mathbf{2^{k-2}}$ consists of a fourth of the runs of the full factorial design. Quarter-fraction designs require two defining relations. The first defining relation returns half or the $\mathbf{2^{k-1}}$ design. The second defining relation selects half of the runs of the $\mathbf{2^{k-1}}$ design to give the quarter fraction.

4. METHOD

4.1. Sources of Data

Seroprevalence data utilized was obtained from the 2006 South African antenatal 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 (pps) 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.

HIV testing methodology used was the World Health Organization (WHO) recommended procedure for the antenatal surveys (World Global Programme on AIDS). The blood samples were tested using ELISA (Abbott Axysm system for HIV1/HIV2). Syphilis testing was conducted using Rapid Plasma Reagin (RPR) test (Brewer Diagnostic Kits, BBL Microbiology Systems, Maryland, USA) (Department of Health, 2010).

4.2. Generating the Experimental Design

4.2.1. Sampling

To facilitate the experimental design, a random sample of 330 was taken from the 33 034 subjects using ***SAS 9.1.3 Analytics platform*** (SAS Institute Inc., Cary, NC, 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.

4.2.2. Missing data

Out of the total data cases, 323 completed cases were selected out of 330 cases (97.88%) and the incomplete entries (7 cases – 2.12%) were discarded.

4.2.3. Variables

The variables used in the study were parity, gravidity, education, RPR, mothage, fathage and HIV status. 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.

4.2.4. Experimental Design Technique

In this study, the aim was to screen the overall main interaction effects among 6 independent variables (Table 1) in an economical manner. Under these circumstances (>5 factors are studied), the

Table1: The Fractional Factorial (Resolution 4) Matrix Design

Exp #	*Parity*	*Gravidity*	*Education*	*RPR*	*Mothage*	*Fathage*
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

fractional factorial resolution 4, design is highly recommended. Higher order linear full factorial and quadratic Box-Behnken designs would require 66 and 52 experimental runs, respectively, which is prohibitively time consuming. Two level designs are ideal for screening because these designs are simple and economical and they also give most of the information required to progress to a multilevel response surface to determine response behavior. Properly chosen fractional factorial designs for 2- level experiments have the desirable properties of being both balanced and orthogonal. Higher order terms such as x^2 are not estimated with fractional factorial designs. Essential Regression and Experimental Design, version 2.2 (Gibsonia, PA) was used to generate the statistical matrix, which required 18 experimental runs (16 model runs and 2 centre points) (Table 1).

The regression model (Equation 1) encompassing 7 linear terms was as follows;

Response = b0 + b1 * Parity + b2 * Gravidity + b3 * Education + b4 * RPR + b5 * Mothage + b6 * Fathage (1)

Where the response was the HIV status and the terms represent the 6 demographic characteristics in antenatal data; $b_0 \ldots b_6$ are the regression coefficients of the system.

The confounding rules were as follows;

Mothage = Parity * Gravidity * Education (2)

Fathage = Gravidity * Education * RPR

4.2.5. Choice of Levels for the Factors

Table 2: Factor Levels

Factor	Levels		Reason
	-1	1	
Parity (No. of children)	0	>1	41% (≈50% of attendee have 0 children
Gravidity (No. of pregnancies)	1	> 1	≈50% of attendees with 1 pregnancy
Education (Grades)	≤ 10	11, 12, 13	≈50% (49.9%) ≤ Grade 10
RPR (Syphilis)	0	1	
Mothage (years)	≤ 24	> 24	≈50% attendees ≤ 24 years
Fathage (years)	≤ 28	≥ 28	≈50% attendees ≤ 28 years

5. EXPERIMENTAL RESULTS AND ANALYSIS

SAS Design of Experiments was employed to develop a predictive model for HIV including all the demographic characteristics (Equation 3),

$HIV = b_0 - b_1 * Parit + b_2 * Gravid - b_3 * Educat - b_4 * RPR + b_5 * Mothage + b_6 * Fathage$ (3)

Where $\mathbf{b_0} = 0.31$, $\mathbf{b_1} = -0.1125$, $\mathbf{b_2} = 0.0465$, $\mathbf{b_3} = -0.1435$, $\mathbf{b_4} = -0.031$, $\mathbf{b_5} = 0.2775$ and $\mathbf{b_6} = 0.079$

Using the Lenth plot (Fig. 1), a method proposed by Lenth (1989), simultaneous margins of error (SME) around zero were computed to determine the relative contribution of each demographic characteristic to the HIV prevalence. The Lenth plot does this by determining the effect sizes which exceed the SME and these are considered active. A significance level (alpha) of 0.05 was adopted for the SME.

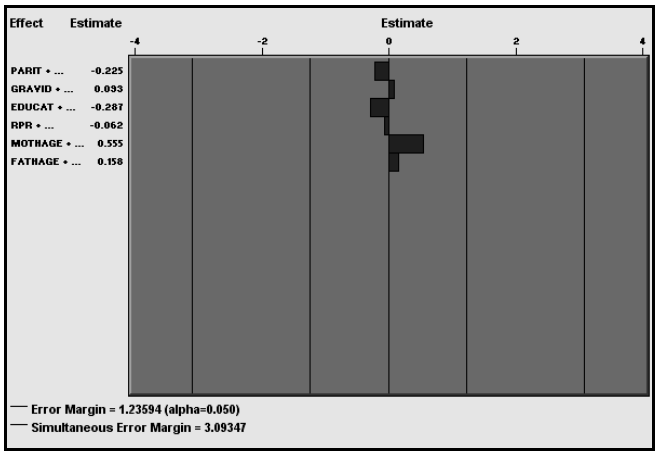

Figure 1. Lenth Plot

Only the effects within 2.5 times the preliminary estimate were included in the trimmed median in an attempt to include only the inactive effects in the estimate.

According to the Lenth plot, Mothage had the greatest influence on HIV prevalence followed by educational level, parity, fathage, gravidity and lastly syphilis (RPR). This scenario was further illustrated by the normal plot shown below (Fig. 2).

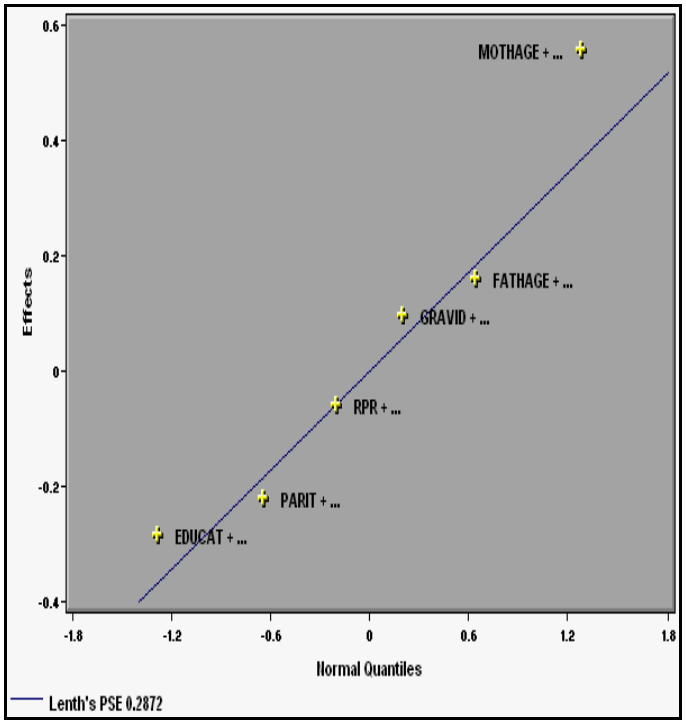

Figure 2. Normal Plot

To further investigate the differential effects of the demographic characteristics on the HIV prevalence, Essential Regression and Experimental Design[R] stepwise regression was employed. Forward selection added first the variable Mothage with the highest F-value when testing the significance of the regression. The next regressor variable added was education with the highest partial F-statistic, showing the highest partial correlation with the response HIV after accounting for effects of other variables already in the model. The forward regression converged to the equation of the response parameter provided in equation 2, since no additional regressor exceeded the predefined F_{in}.

Response (HIV) = b_0 + b_1 * Mothage + b_2 * Education (4)

Where b_0 = 0.363, b_1 = 0.340 and b_2 = – 0.160

Table 3: Correlation between Predicted and Experimental Responses Generated from the Fractional Factorial Design

Case	Mothage	Education	Resp_1	Predicted Resp_1	Residuals	Standardized Residuals	Cook's Distance
1	1	1	0.62	0.54	0.076	0.435	0.0750
2	-1	-1	0	0.18	-0.182	-1.037	0.2110
3	1	-1	1	0.86	0.137	0.779	1.0510
4	1	1	0.33	0.54	-0.214	-1.214	0.5850
5	-1	1	0	-0.14	0.137	0.779	1.0510
6	-1	-1	0.22	0.18	0.038	0.214	0.0090
7	-1	-1	0.19	0.18	0.008	0.043	0.0004

Table 4: Regression Diagnostics

Response Measured	R^2	Durbin-Watson Index (%)	CV (%)*
HIV Prevalence	0.842	3.440	52.169

CV indicates Coefficient of Variation

Table 5: Level of Significance of Regression Coefficients Generated in the Linear Model of Response Parameter at a 95% Confidence Interval (P<0.05)

Coefficient	Independent Variables	Regression coefficient
b0	-	0.363
b1	Mothage	0.340
b2	Education	-0.160

5.1 Model Adequacy

After calculating a model, a thorough analysis of the residuals is important to evaluate the adequacy of the regression. The methods used for residual analysis are a) normal probability plot of residuals, b) plot of residuals vs. predicted response and c) outlier analysis using threshold or cutoff values.

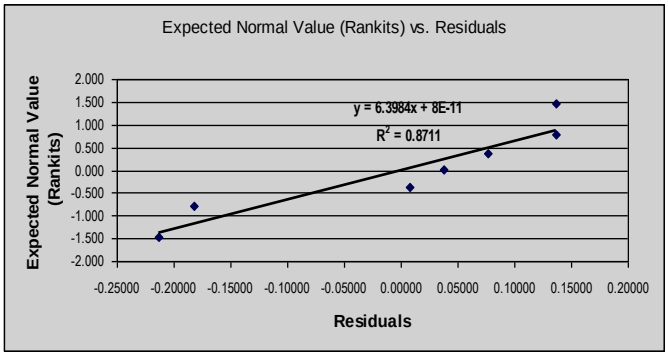

Figure 3: Plot of Expected Normal value (Rankits) against residuals.

The normal probability plot (Fig. 3) formed a straight line indicating that the residuals were perfectly normally distributed. In addition the plot of residuals against predicted response (Fig 4 yielded a horizontal band on both sides of the expected average for the residuals, zero. The plot (Fig. 5) of residuals against experimental cases indicated that they were no significant outliers in this study.

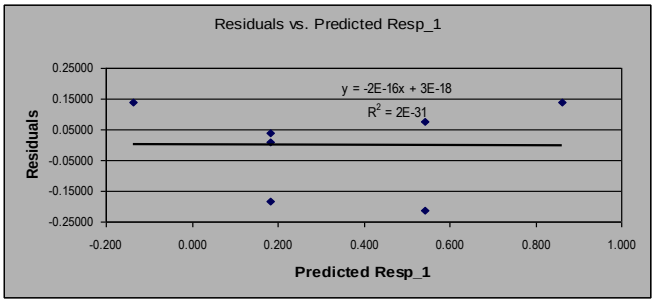

Figure 4: Plot of residuals against predicted response

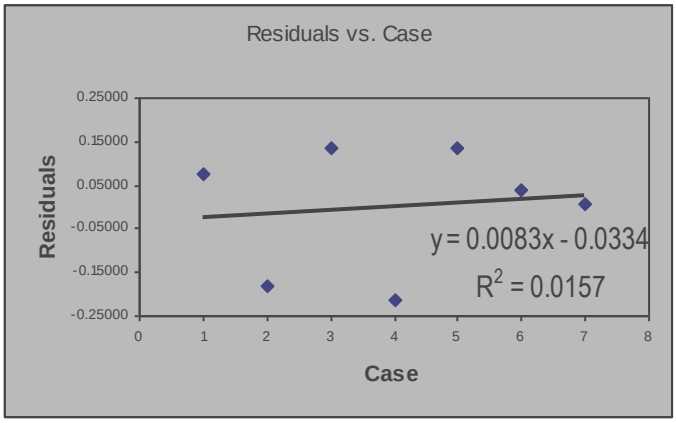

Figure 5: Plot of residuals against experimental cases

5.2 Constrained Optimization

In this study, optimization was accomplished by the Lagrangian approach originally introduced by Fonner and co-workers (Fonner, Buck and Banker, 1970).

Tables 4 and 5 provide a detailed analysis of the mathematical fit of the data for the response function and the significance level. This analysis is required for the final selection of the objective function to be applied in the constrained optimization. The regressor variables mothage and education were selected as optimization objectives.

The constraints imposed on the regressor variables were as follows;

$$-1 \leq \text{Mothage} \leq 1$$

$$-1 \leq \text{Education} \leq 1$$

Table 6 indicates the results obtained by constrained optimization. It is evident that the highest prevalence of HIV (86%) is attainable at the lowest level of education and highest maternal age.

Table 6, reflects the results based on the above constrained optimization.

Regressor variable	Level	HIV Prevalence (%)
Education	-1	86.29
Mothage	1	

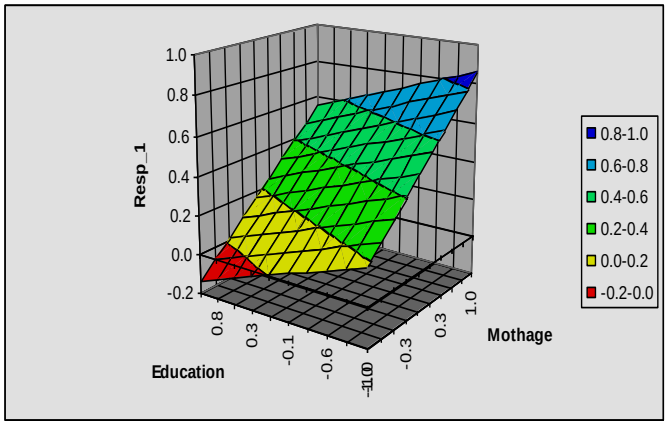

Figure 6: A typical surface response plot indicating the influence of increasing mothage and decreasing education on HIV prevalence.

Figure 6 illustrates a typical surface response plot depicting the effects of mothage and education on the HIV prevalence. Note that all other combinations of the surface plots were analyzed but not shown since those responses were statistically unstable. It was observed that as the mother's age increased, the HIV prevalence significantly increased. In addition, a decrease in education also significantly increased the HIV prevalence.

6. CONCLUSIONS

Application of the fractional factorial design assisted in the successful screening of demographic characteristics and optimization of the demographic characteristic combinations for the prediction of HIV prevalence. The measured HIV prevalence response was in close agreement with the predicted values of the optimized demographic characteristics combination, as shown in Fig. 6.

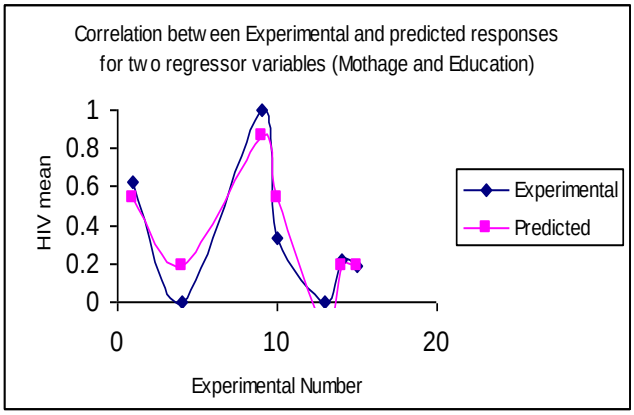

Fig. 6. Correlation between experimental data and values predicted by the response models

7. 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 Cathrine Tlaleng Sibanda and the National Department of Health (South Africa) for the antenatal seroprevalence data (2006-2007).

8. REFERENCES

[1] Benferoni M.C., Rossi S., Ferrari F., Stavik E., Pena-Romero A. and Caramella C. (2000) 'Factorial analysis of the influence of dissolution medium on drug release from carrageenan-diltiazem complexes', *AAPS PharmSciTech*, Vol. 1, No. 2, Article 15.

[2] Department of Health (2010) 'National Antenatal Sentinel HIV and Syphilis Prevalence Survey in South Africa, 2009'.

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

[4] Fonner D.E., Buck J.R. and Banker G.S. (1970) 'Mathematical optimization techniques in drug product design and process analysis', *J. Pharm Sci*, Vol. 59, pp. 587-1596.

[5] Furlanetto S., Maestrelli F., Orlandini S., Pinzauti S. and Mura P. (2003) 'Optimization of dissolution test precision for a ketoprofen extended-release product', *J. Pharm. Biomed Anal.* Vol. 32, pp. 159-165.

[6] World Global Programme on AIDS (1989).

[7] Younes B., Fotheringham A. and El-Dessouky H.M. (2010) 'Factorial optimization of the effects of extrusion temperature profile and polymer grade on as-spun aliphatic–aromatic copolyester fibers. I. Birefringence and overall orientation', Journal of Applied Polymer Science, Vol. 118, No. 3, pp. 1270-1277.