Chapter 4: Discussion

Index
4.1. Screening design-fractional factorial
4.2. Response surface methodology-Central composite face-centered design
4.3. Comparison of two response surface methodologies
4.4. Comparison of response surface methodology with a binary logistic regression
4.5. Application of multi-layer perceptron (MLP) to model HIV data
4.6. Comparison of different modeling techniques using SAS Enterprise software.
4.7. Scorecard design results

4.1. Screening design-fractional factorial

The application of the fractional factorial design assisted in successfully ranking the effects of a pregnant woman’s demographic characteristics on the risk of acquiring HIV infection. As shown in Fig. 5.1, the mother’s age ranked as the most influential factor, followed by her educational level. Interestingly, the sexually transmitted disease syphilis was observed to be of little influence. It has however, being reported in literature that there is very little direct relationship between exposure to syphilis and the risk of acquiring an HIV infection.

![Lenth’s plot of the effect of demographic characteristics on the risk of acquiring HIV infection](image)

Fig. 5.1: Lenth’s plot of the effect of demographic characteristics on the risk of acquiring HIV infection

4.2. Response surface methodology-Central composite face-centred design

The purpose of this study was to explore interaction and second-order relationships between demographic characteristics in influencing the risk of acquiring HIV infection amongst pregnant women in South Africa.
The central composite face-centred design confirmed the results obtained by fractional factorial design (screening exercise), that mother’s age was the most influential factor in determining the risk of a pregnant women acquiring HIV infection, as shown in Fig. 5.2 (Sibanda & Pretorius 2012).

![Coefficient Plot](image)

**Fig. 5.2: CCF coefficient plot**

The coefficient plot (Fig. 5.2), was drawn to represent the information provided by the 2-factor interaction response model generated by the central composite face-centered design. In general coefficient plots tend to clearly represent the relative importance of each variable on the model equation. The coefficient plot confirmed that the interaction between the mother’s age and other demographic characteristics such as father’s age, education and parity exhibited marked influence on the risk of acquiring HIV infection (Sibanda & Pretorius 2012).

The perturbation plot (Fig. 5.3) was employed to study the effects of all demographic characteristics on the risk of acquiring an HIV infection at a given point within the design space. A steep gradient for a given demographic characteristic indicates that the HIV risk is sensitive to that factor, while a flat line depicts lack of sensitivity to variations in that demographic characteristic.
The R-square value of the predictive model for the central composite face-centred design was 98% compared to that of the fractional factorial design at 33.5%. R-square refers to the fraction of variance and it is a square of the correlation coefficient between two dependent variables. It is therefore a statistical term that informs how good one variable is at predicting another. In simple terms therefore, it means the response surface model better fits the data than the screening model. Lastly, the predicted values from the central composite face-centred design compared remarkably well with the actual or observed values (Sibanda & Pretorius 2012).

4.3. Comparison of two response surface methodologies

A central composite face centred design was compared to the Box-Behnken design with respect to the capability to determine the effect of demographic characteristics on HIV risk. The
central composite face-centred and the Box-Behnken designs produced the same results. The 2-factor interaction polynomial functions for the mother’s age, father’s age, education and parity were found to be statistically more significant compared to linear main-effect models as shown in Fig. 5.5 (Sibanda & Pretorius 2013).

![Co-efficient Plot](image)

**Fig. 5.5:** Coefficient plots of CCF BBD designs

Finally, the central composite face-centred and the Box-Behnken designs confirmed the results obtained by fractional factorial design, that mother’s age had the greatest influence on the risk of acquiring HIV infection (Sibanda & Pretorius 2013).

![A plot of actual vs predicted HIV risk for the BBD design](image)

**Fig. 5.6:** A plot of actual vs predicted HIV risk for the BBD design
A plot of the predicted vs. actual HIV risk for the Box-Behnken design shows a high level of accuracy. For an example, at an HIV risk of 0.17, the model predicted a value of 0.17. The same level of accuracy was observed for the HIV risk levels of 0.28 and 0.33 that yielded predicted risk values of 0.28 and 0.33 respectively (Sibanda & Pretorius 2013).

**Fig. 5.7:** A plot of actual vs predicted HIV risk for the CCF design

The plot of predicted vs actual HIV risk values derived from the central composite face-centred design shows relatively lower level of accuracy compared to the Box-Behnken design. For example for an actual HIV risk value of 0.1, the model gives a predicted value of 0.05 (Sibanda & Pretorius 2013).

4.4. Comparison of response surface methodology with a binary logistic regression

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 (Sibanda & Pretorius 2013).

4.5. Application of multi-layer perceptron (MLP) to model HIV data

The sensitivity test showed that mother’s age and the father’s age had the greatest influence on the risk of acquiring HIV infection. Therefore, the MLP is a useful tool for prediction, function approximation and classification. The practical benefits of a modelling system that can accurately reproduce any measurable relationship are huge (Sibanda & Pretorius 2011).
4.6. Model Comparison using ROC curves using SAS Enterprise Miner TM software (Neural networks, logistic regression, decision trees and full factorial design).

![Graph showing model comparison using ROC curves](image)

**Fig. 5.8:** A cumulative percentage response chart

The cumulative percentage response chart arranges observations into deciles based on their predicted probability of response, and then plots the actual percentage of respondents. The neural network appears to be the most superior model followed by the logistic regression and the decision tree performed least.

![Graph showing noncumulative percentage response chart](image)

**Fig. 5.9:** Noncumulative percentage response chart
A noncumulative plot is more helpful and shows the effectiveness of the model for each level of the score. Fig. 5.9 reveals that the neural network model’s predictive power drops gradually after the top 30 percent of scores. By the forth decile, the neural network model is less effective than the logistic regression and full factorial models. Interestingly after the sixth decile neural network and logistic regression models are worse than the decision tree. Crossovers are observed when models are compared on a noncumulative lift charts. However, when the lines do not cross each other the upper one is considered to be the superior model. On that basis, Fig. 5.9 shows that the neural network model is the most superior model if the only twenty percent of individuals are considered. Interestingly, after the seventh decile, the decision tree is by far more superior.

On the other hand, the receiver-operating-characteristics (ROC) curve is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of positives against the fraction of false positives out of the negatives at various threshold settings. True positive rate is also known as sensitivity, and false positive rate is one minus the specificity or true negative rate. The ROC curve is therefore an appropriate tool for selecting the most superior models and is therefore a measure of the cost vs benefit analysis of diagnostic decision making.

![ROC Curve](image)

**Fig. 5.10:** Receiver Operating Characteristic curves (ROC)
On the basis of the ROC curves, the neural network was found to be marginally superior to the logistic regression model and the decision tree was the least performing model.

Fig. 5.11: Response threshold charts

Fig. 5.11 shows that the prediction accuracy of the statistical models varies across different threshold levels.

4.7. Building a scorecard using weight of evidence bins in SAS Enterprise software

The credit scoring add-on in SAS Enterprise Miner™ was used to build binary target (HIV positive, HIV negative) scorecards. The process involved grouping variables using weight of evidence (WOE) then performing logistic regression to produce predicted probabilities. The scorecard had a slightly lower fit statistics compared to a direct prediction model like neural networks. However, the scorecard performed at par with the decision tree. Based on the above results, it was concluded that even though the performance of the scorecard was not equal to that of the neural network it was acceptable for the purposes of investigating the differential effects of demographic characteristics on the risk of acquiring an HIV infection amongst pregnant women attending antenatal clinics in South Africa.
4.8. Implications of Research Findings

This research demonstrated that the age of the mother had the greatest influence on the risk of acquiring an HIV infection amongst pregnant women. The other factors that ranked highly were the father’s age, parity and the pregnant woman’s level of education. This information can be of tremendous value to policy makers involved in formulating strategies to curb the spread of the HIV/AIDS infection within the South African communities. It is important to fully understand the drivers of the HIV pandemic in order to arrest its spread. The appreciation that the mother’s age seems to be directly related to the increased chances of being HIV positive provides an opportunity to target the relevant age groups of women in order to raise awareness of the importance of safe sexual practices to prevent the spread of HIV/AIDS. There are many intervention strategies to curb the spread of the HIV/AIDS epidemic, such as awareness of the disease, promotion of safer sex and reducing multiple sexual partners.

It is important for governments to understand the dynamics of the spread of the HIV epidemic in order to spend the available resources effectively to curtail the pandemic. In addition, a full understanding of the HIV/AIDS will inevitably feed positively into the government’s National Strategic Plan (2012-2016). The National Strategic Plan is a framework that provides goals and strategies for South Africa’s response to HIV, STI’s and TB during the period 2012-2016. The objectives of the National Strategic Plan include; halving the number of new HIV infections, ensuring that at least 80% of people who are eligible for treatment for HIV are receiving it, ensuring the rights of people living with HIV are protected and halving the stigma related to HIV and TB.

4.9. What do the models mean?

The neural networks as statistical models can be used for the purposes of prediction as shown by increased classification accuracy based on ROC curves, confusion matrices and other measures of accuracy. Artificial neural networks have been considered to be black boxes. This means that neural networks are only understood in terms of the input variables and response. However, little is understood about processes within. This is the reason why this methodology was compared with more transparent techniques such as logistic regression, decision trees, design of experiments and scorecard methodology.
The DOE provide a deeper understanding of the effect of demographic characteristics on the risk of acquiring an HIV infection and is relatively easier to understand how the methodology arrives at its final output.

Logistic regression models are the ideal models for modeling binary response data. In this research, the logistic regression was used for benchmarking the outcomes from the other modeling processes.

4.10. Future Research

The National Department of Health in South Africa has been conducting the antenatal sentinel HIV and syphilis prevalence survey annually, for the past 21 years. This survey has been used as an instrument to monitor the HIV prevalence trends since 1990. HIV prevalence is currently the only parameter that is measured accurately, whilst the country depends on the use of mathematical modeling to estimate HIV incidence and HIV related mortality. With the vast amount of antenatal data collected in South Africa, the research conducted in the doctoral project can be extended to study the changes of the effects of demographic characteristics on HIV prevalence over time. The latter research will shed light on the possible changes in the effects of demographic characteristics on HIV prevalence over time. This information will indeed be an additional arsenal for policy makers with regards to dealing with the HIV/AIDS epidemic in South Africa.