MODEL-BASED CALIBRATION OF A NON-INVASIVE BLOOD GLUCOSE MONITOR

by

Yelena Shulga

A Project Report

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Applied Statistics

January 2006

APPROVED:

_____

Dr. Jayson D. Wilbur, Advisor

_____

Dr. Bogdan M. Vernescu, Department Head

ABSTRACT

This project was dedicated to the problem of improving a non-invasive blood glucose monitor being developed by the VivaScan Corporation. The company has made some progress in the non-invasive blood glucose device development and approached WPI for a statistical assistance in the improvement of their model in order to predict the glucose level more accurately. The main goal of this project was to improve the ability of the non-invasive blood glucose monitor to predict the glucose values more precisely. The goal was achieved by finding and implementing the best regression model. The methods included ordinary least squared regression, partial least squares regression, robust regression method, weighted least squares regression, local regression, and ridge regression. VivaScan calibration data for seven patients were analyzed in this project. For each of these patients, the individual regression models were built and compared based on the two factors that evaluate the model prediction ability. It was determined that partial least squares and ridge regressions are two best methods among the others that were considered in this work. Using these two methods gave better glucose prediction. The additional problem of data reduction to minimize the data collection time was also considered in this work.

ACKNOWLEDGEMENTS

I would like to express my gratitude and appreciation to my advisor, Professor Jayson Wilbur for his support, valuable guidance, and helping me throughout my research project about the improvement of the accuracy of non-invasive blood glucose monitor with statistical modeling. It was a pleasure for me to work with Professor Wilbur on this project as well as attending the classes that he taught. I would like to thank VivaScan Corporation group: Robert Peura, Hannu Harjunmaa, and Rebecca Burrell for providing me with the unique opportunity to work on the non-invasive glucose reading device.

I would like also to thank the WPI graduate students, with whom I worked hard together during the years of my Master's study. I am also very grateful to all Professors from WPI math department whose lectures I attended, for their time, advice, and help. The study at WPI allowed me to gain invaluable knowledge in statistics and practical skills.

I am very grateful to my husband, Yuri Yudin for his support and patience.

Finally, I would like to express a special thank to my father Alex Shulga, and to my mother Galina Shulga for taking care about my infant son during my busy study time. Without them, this work could not have been possible to complete.

Contents

List of Tables

List of Figures

1. Introduction

1.1 Diabetes Problem

According to the Center for Disease Control and Prevention (CDC), the number of Americans with diabetes more than doubled from 1989 to 2002 – from 5.8 million to 13.3 million. One in three Americans born after 2000 will develop diabetes in their lifetime [1]. People with diabetes have a shortage of insulin. This is a hormone that allows glucose, or sugar, to enter and be converted to energy. If left unchecked and uncontrolled, diabetes can lead to the serious conditions including heart attack, stroke, blindness, kidney failure, and blood vessel disease.

Despite these severe health problems, diabetes can be controlled and it can be managed. A recent 10-year study showed that diabetics who kept their blood glucose under control could reduce their risk or slow down the development of health complications that can happen from diabetes by 50 percent or more [2].

Monitoring blood glucose levels is a necessary daily procedure for people with diabetes. The results from these observations show the effectiveness of medications, diet, and life-style. Diabetics should regularly test and record their blood glucose. The results of self-blood-glucose-monitoring allow people with diabetes and their health care providers to effectively adjust their diabetes plan.

1.2 Blood Glucose Monitors

Most current methods for self-blood-glucose-monitoring are invasive in that they require a blood sample for each test, usually obtained from a fingertip. Patients with diabetes must monitor their blood glucose level several times each day. The blood

sampling can be painful and cause calluses to form. It also increases the risk for warts and infections. Therefore, scientists have been trying to find new ways for people with diabetes to measure their blood sugar without needing a skin puncture to get a blood sample (i.e., non-invasive methods) [3].

1.3 Non-invasive Technologies

Non-invasive technologies are those which do not penetrate into deep layers of tissue. The advantages of such technologies over the invasive ones are less intense and/or less frequent pain, as well as the reduced risk of infection. Thus, the non-invasive devices allow the diabetics to test their blood glucose more often and therefore maintain better health. Potential non-invasive ways to determine blood glucose levels include: measuring the energy waves (infrared radiation) emitted by the body, applying radio waves to the fingertips, using ultrasound, measuring glucose levels in saliva or tears, shining a beam of light onto the skin or through body tissues

One main disadvantage of many non-invasive devices is the lack of accuracy of the measurement of the glucose relative to traditional invasive measurement devices. The reliable accuracy is important, such as adjusting the amount of insulin to take, will be based on the results of the device. Improving the accuracy of the non-invasive blood glucose monitor is a major problem for the device developers.

1.4 VivaScan Non-invasive Monitor

This project was sponsored by VivaScan Corporation (VSC). VivaScan is a company in Massachusetts that is developing a non-invasive, optical glucose sensor.

VivaScan's device shines infrared light through the earlobe. The device uses two near-infrared light beams, one tuned as a baseline to reject interfering substances and the other to register blood-sugar content according to how much light is absorbed by glucose just beneath the skin. The device gently compresses the earlobe to squeeze blood out of the tissue, and then releases the lobe to restore normal blood flow. Light-intensity readings are taken before, during, and after the squeeze [4].

1.4.1 Data Collection and Calibration

The data collection process for statistical analysis includes collection of measurements of light intensity (non-invasive device output). Data acquisitions are made every 15 minutes. Immediately after every (or every other) measurement, an invasive reference reading is taken. For the invasive reading the HemoCue glucose meter with high accuracy (±3.5%) is used. If a measurement does not have a reference glucose value, then the linear interpolated glucose value between two neighboring points is used. Thus, each measurement produces one data point.

The collected data include three main predictor variables which will be referred in this report as *X1, X2, X3,* and additional 17 variables that may be important for the prediction of the glucose level. The collected data were divided into two parts. One part is called the calibration data. These data are used to fit the ordinary least squares (OLS) regression model to establish the relation between the three predictor variables *X1*, *X2*, *X3* and a reference glucose values. Second part is called prediction (test) data and is used to validate the built regression model performance in prediction of the blood glucose level.

The VSC non-invasive devise should be calibrated for each patient to cover the blood glucose range expected on the patient. During the calibration procedure three parameters are adjusted to optimize the model. Model is considered to be optimized when the correlation of the predictor variables *X1, X2, X3* with glucose achieves the maximum possible value and the p-value of the regression model reaches 1% of statistical significance (p-value is 0.01).

The calibration process starts when seven data point are collected. As seven points are collected, the OLS regression model is fitted. With the addition of a new measurement the regression model is fitted again and parameters are adjusted every time. These calibration measurements are accumulated until the criterion of 1% statistical significance is reached. When the criterion of 1% statistical significance for the corresponding regression model is obtained the adjusted parameters are fixed and we can say that calibration process is completed. Now the test data can be collected to validate the built regression model performance.

1.4.2 Problem Statement

As it was mentioned above, the non-invasive methods are biased in the blood glucose prediction. This is mainly the engineering problem, but it could be turned to the statistical problem of the data analysis. The goal of this project was to improve the ability of the non-invasive blood glucose monitor to predict the glucose levels more correctly by finding and implementing the best regression model. The following proposed regression methods were considered and compared with the ordinary least squares regression (OLS) results in the prediction of the glucose values using the actual VSC data. These models

are partial least squares regression (PLS), robust regression (ROBUST), weighted least squares regression (WLS), local regression (LOCAL), and ridge regression (RIDGE).

1.4.3 Measures

The performance of each of the model was evaluated by two factors: Average prediction error in percentage (PAPE) and percent of acceptable points according to the "± 20% rule". Average prediction error (APE) is:

$$\text{APE}=\frac{1}{n}\sum_{i=1}^{n}|Y_i-\hat{Y}_i| \tag{1.4.1}$$

Y_i – actual glucose value for *i*th observation

$\hat{Y}_i$ – predicted glucose value for *i*th observation

n is the number of the observations in the dataset

Average prediction error in percentage (PAPE) can be written as following:

$$\text{PAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\frac{|Y_i-\hat{Y}_i|}{Y_i} \tag{1.4.2}$$

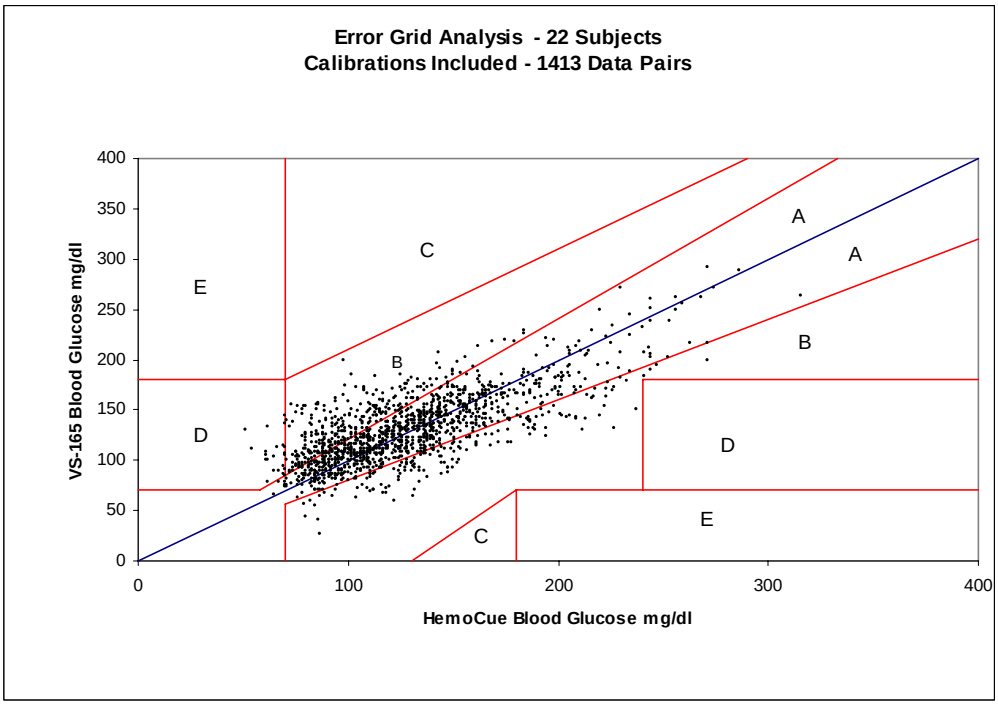
A point is accepted according to the "± 20%" rule if the prediction error is less than or equal to 20%:

$$\frac{|Y_i-\hat{Y}_i|}{Y_i}100\% \le 20\% \tag{1.4.3}$$

For a good regression model, the value of PAPE should be small, while percent of acceptable points should be large. To compare the performance of the regression models the combination of these two values was considered.

A Clarke graph can be used to demonstrate the "± 20%" criterion. This graph represents the error grid analysis and is usually used to evaluate the performance of the blood glucose monitor. On Clarke graph the reference blood glucose values (HemoCue output) are plotted against the values generated by monitoring system (VSC device output). Zone A represents glucose values that deviate from the reference by no more that 20%. If a point falls into zone A, it is acceptable according to the "± 20%" rule. Values falling within this range are considered clinically accurate [5].

Figure 1.1: Clarke Error Grid



2. Methods Description

2.1 Ordinary Least Squares Regression

We assumed that there is a linear relationship between the glucose level and the independent (predictor) variables. This relation can be expressed in the ordinary least squares regression model (OLS) form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad (2.1.1)$$

where:

Yi are the true glucose values obtained with HemoCue invasive glucose meter

$\beta_0, \beta_1, ..., \beta_{p-1}$ are regression parameters

$X_{i1}, X_{i2}, ..., X_{i,p-1}$ are variables measured by the device (non-invasive device output)

ε_i are independent $N(0, \sigma^2)$

i=1,…*n*,

p – number of regression parameters.

The response function for the regression model (2.1.1) is:

$$\mathrm{E}\{\mathrm{Y}\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} \qquad (2.1.2)$$

The OLS regression model with normal error terms implies that the observed glucose values are independent normal variables, with mean $\mathrm{E}\{Y\}$ and with constant variance σ^2.

The regression parameters $\beta_0, \beta_1, ..., \beta_{p-1}$ in (2.1.2) are unknown. In OLS regression method they are estimated using the least square criterion:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1})^2 \qquad (2.1.3)$$

The least square estimators are those values of $\beta_0, \beta_1, ..., \beta_{p-1}$ that minimize Q.

When the regression parameters estimates $b_0, b_1, ..., b_{p-1}$ are found, the OLS regression model can be used to predict the new glucose values.

When a regression model is considered for an application, we want to be sure that the model is appropriate for the use. The OLS regression method suffers from the limitations and may not be the best model. One of the limitations is that the errors should be normally distributed. Also the OLS method is sensitive to the outliers and to the non-constancy of the error variance.

If the assumption about normality of the errors is violated, the regression OLS function may not be appropriate for the glucose prediction. When outlying observations are present in the data, it can seriously distort the estimated regression function and may affect the normality of the distribution of the error terms. The VSC data have relatively small number of observations. The presence of outlying cases in a small dataset can greatly impact the fitted regression function. In the case when the error variance is not a constant, it causes variance of regression parameters estimates to be large, as well as p-value of the regression model may be affected.

2.2 Weighted Least Squares Regression

For OLS regression model, the error terms ε_i are assumed to be independent normal random variables, with mean zero and constant variance σ^2. When the error variance is not constant but varies in a systematic fashion, a direct approach is to modify the model to allow for this and use the method of weighted least squares (WLS) to obtain the estimators of the parameters.

Denote the variances of the error terms ε_i by σ_i^2 indicating that the variances of the errors are different. The errors ε_i are defined as:

$$\varepsilon_i = Y_i - E(Y_i) \tag{2.2.1}$$

The residuals are the differences between the observed values Y_i and fitted values $\hat{Y}_i$:

$$e_i = Y_i - \hat{Y}_i \tag{2.2.2}$$

So, e_i reflects the properties assumed for ε_i.

Suppose that the error variances are known, then the method of maximum likelihood can be used to obtain the regression coefficients in 2.1.2. The likelihood function from the OLS method is modified by replacing $\sigma^2 i$ with weights ω_i, where

$$\omega_i = \frac{1}{\sigma_i^2} \tag{2.2.3}$$

Now, we are minimizing a weighted sum of squares:

$$Q_w = \sum_{i=1}^{n} \omega_i (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1})^2 \tag{2.2.4}$$

In matrix notations the maximum likelihood estimators of the regression coefficients are:

$$b_w = (X^T W X)^{-1} X^T W Y \tag{2.2.5}$$

Where b_w is the vector of *p* estimated coefficients, and *W* is $n \times n$ diagonal matrix containing the weights ω_i.

When σ_i^2 are unknown, they can be estimated. Since $E(\varepsilon_i) = 0$ by assumption, $\sigma_i^2 = E\{\varepsilon_i^2\} - (E\{\varepsilon_i\})^2 = E\{\varepsilon_i^2\}$. The squared residuals e_i^2 can be used to estimate σ_i^2, and the absolute residuals $|e_i|$ can be used to estimate the standard deviation $\sigma_i = \sqrt{\sigma_i^2}$. Thus, the variances can be estimated by fitting the regression model using unweighted least squares (OLS) and then regressing the squared residuals e_i^2 against the predictor variables. The standard deviations can be estimated by fitting the regression model using OLS and then regressing the absolute residuals $|e_i|$ against the predictor variables. The fitted values of variance function and the standard deviation function are used to estimate the weights:

$$\omega_i = \frac{1}{(\hat{s}_i)^2}$$, where $\hat{s}_i$ is fitted standard deviation

$$\omega_i = \frac{1}{\hat{v}_i}$$, where $\hat{v}_i$ is fitted variance

The estimated regression coefficient can be obtained now using (2.2.5).

The residual plots where residuals plotted against the predictor variables or the fitted values $\hat{Y}_i$ are used to investigate the constancy of the error variances. These plots for VSC data indicate that the variances of errors are increasing or decreasing in a systematic manner or vary in more complex fashion related to independent variables *X*s or the predicted response E(*Y*). This fact denotes that the variances of the error terms may be not constant and using WLS method instead of OLS may be reasonable.

2.3 Partial Least Squares Regression

Sometimes predictor variables tend to be correlated among themselves. The situation when the predictor variables are correlated among themselves is called multicollinearity. When multicollinearity exists, the estimated regression coefficients in the OLS model tend to vary from one sample to the next as well as they depend on which variables are included in the model and which are left out. Partial least squares regression (PLS) method helps to overcome the problem of multicollinearity. This method is also instrumental if there are many predictor variables.

The goal of PLS method is to predict Y from X and to describe their common structure. PLS regression is a method of using both the predictor matrix (matrix of independent variables) X and the response matrix Y (matrix of dependent variable) to extract a set of factors (latent vectors) with the constraint that these factors explain as much as possible of the covariance between X and Y [6]. The number of the extracted factor is usually specified to be less that the number of predicted variables Xs. The emphasis of PLS method is made on predicting Y and not necessarily on trying to understand the relationship between the variables.

In application of the PLS regression method, both the predictor and response matrices are decomposed, such that

$$X_c = TP^T + E$$

$$Y_c = UQ^T + F$$

where P is the factor loading matrix, Q is the coefficient loading matrix, and E and F are factors in X and Y that are not described by the PLS model. In the above equations, X_c

and Y_c represent the mean centered matrices of *X* and *Y* respectively. PLS method tries to find a score vector in the column space of X_c and a score vector in the column space of Y_c such that

$$t = X_c w$$

$$u = Y_c q$$

to give the maximal squared covariance for $(u^T t)^2$. That is, the process aims to maximize $(q^T Y_c^T X_c w)^2$ subject to $|w|=|q|=1$. The solution to this equation is given by an eigenvalue problem of $Y_c^T X_c$:

$$q^T Y_c^T X_c w = \lambda w$$

where λ is the eigenvalue associated with *w*. Rather than linking measurements *X* and *Y* directly, the method tries to establish the inner relationships between latent variables *T* and *U*, derived from *X* and *Y*, respectively, i.e.:

$$U = TB + U_E$$

where *B* is a diagonal matrix that has the regression weights as the diagonal elements, and U_E is an error term. When these error terms are ignored, we can obtain the predicted value of Y_c as

$$Y_c = TBO^T$$

In PLS the factors are extracted in order of significance. For a good PLS model, a set of the first few extracted factors explains much of the covariance between *X* and *Y*. To apply PLS method to the VSC data, we assumed *X* to be a predictor matrix of three

variables *X1, X2, X3*, and *Y* to be the response vector of the glucose values. After analyzing the PLS regression results for the VSC data, it was found that the PLS model with one extracted factor give a smaller average prediction error (APE) than the PLS method with two and three factors.

2.4 Ridge Regression

Ridge regression is another method along with partial least squares regression method that can help to overcome serious multicollinearity problem. The limitation of PLS method is that it may be difficult to obtain concrete meanings of the extracted factors and explain the relationship between the variables. Ridge regression uses modified method of ordinary least squares which allows one to obtain a linear relationship between the predictor variables. The PLS is preferred to the ridge method when it is necessary to substantially reduce the number of predictors to the small number of extracted factor. Ridge regression principle is based on the fact that the biased estimator of the regression coefficient with small variance may be preferred to unbiased estimator with large variance.

Usually ridge regression is applied to the centered and scaled model. Consider the OLS model (2.1.1). The new standardized (centered and scaled) predictor variable and the response variable can be written as:

$$X_{ik}' = \frac{X_{ik} - \bar{X}_k}{s_k},\ Y_i' = \frac{Y_i - \bar{Y}}{s_Y}$$

where k=1,…,p-1, $s_k = \sqrt{\sum_i (X_{ik} - \bar{X}_k)}$, and $s_Y = \sqrt{\sum_i (Y_i - \bar{Y})}$.

Then, the OLS now in the standardized form:

$$Y'_i = \beta'_1 X'_{i1} + \beta'_2 X'_{i2} + \cdots + \beta'_{p-1} X'_{i,p-1} + \varepsilon'_i \qquad (2.4.1)$$

The solution for (2.4.1) in matrix notations is in the form:

$$\mathbf{b} = (X^T X)^{-1} X^T Y \qquad (2.4.2)$$

where $\mathbf{b}$ is $(p \times 1)$ vector of the least squares estimated regression coefficient,

$X^T X$ is $(p-1) \times (p-1)$ correlation matrix of *X* variables,

and $X^T Y$ is $(p-1) \times 1$ vector of coefficients of simple correlation between *Y* and *X*.

The ridge standardized regression estimates of $\beta'_0, \beta'_1, ..., \beta'_{p-1}$ are obtained by introducing a biasing constant $c \geq 0$ (also called shrinkage parameter) into the OLS model solution (2.4.2) in the following form:

$$\mathbf{b}^R = (X^T X + cI)^{-1} X^T Y \qquad (2.4.3)$$

where $\mathbf{b}^R$ is the $(p-1) \times 1$ vector of the standardized ridge regression coefficients and *I* is the $(p-1) \times (p-1)$ identity matrix.

The constant c reflects the amount of bias in the estimators. When *c*=0, the ridge regression coefficients in (2.4.3) reduces to (2.4.2). When *c*>0, the ridge regression estimators are biased but tend to be less variable than OLS estimators.

The problem is to choose the optimum value of *c* for which the ridge regression estimator $\mathbf{b}^R$ has a smaller mean squared error (MSE) than OLS estimator $\mathbf{b}$. A commonly used method of finding the biasing constant *c* is based on the *variance inflation factors* (VIF*)* and the *ridge trace*. The VIFs are referred to the problem of multicollinearity and widely used to detect this problem. These factors measure how much the variances of the estimated regression coefficients are inflated as compared to when predictors are not linearly related. A large VIF value (about 10) often indicates the

severe multicollinearity. The ridge trace is a plot, where the values of the $p-1$ estimated ridge regression coefficients plotted against different values of *c*. Practically, *c* value can be found by analyzing the ridge trace and VIFs. In ridge trace we choose the minimum value of *c* after which the regression coefficients are moderately stable. For VIFs we choose *c* for which this factor becomes sufficiently small. The choice of *c* is a judgmental one. The full review of proper choices of *c* is given in Draper and Van Nostrand (1979) and Hocking (1976).

2.5 Robust Regression

Statistically, an outlier is an observation that lies outside the overall pattern of a distribution. The outlying cases may be a result of a recording error, measurement error or other extraneous effects, and hence should be discarded. Outliers can create great difficulty. When outlying observations are present, use of the least squares estimators may lead to serious distortions in the estimated regression function. However, not all outliers have strong influence on the fitted regression function.

The robust regression methods have an advantage over the OLS model in damping the influence of outlying cases in an effort to provide a better fit for the majority of cases. Robust regression methods are also useful when automated regression analysis is required. For example, the VSC non-invasive devise uses a calibration procedure to adjust the parameters and to build an individual regression model for each patient. There may be no time or no possibility for identification of all outlying cases and analysis of their influence. Robust regression methods will automatically guard against undue influence of outlying cases in this situation.

Typically, three classes of problems have been addressed with robust regression techniques: problems with outliers in response surface (*Y*), problems with multivariate outliers in the predictor space (*X*), and problems with outliers in both the response surface and predictor space.

There are numerous robust regression methods. In statistical applications, the methods most commonly used today are Huber M estimation, high breakdown value estimation (LTS and S), and combinations of these two methods (MM). If an estimator can resist a large number of outliers, it is said that the estimator has a high breakdown value. M estimation is the simplest approach both computationally and theoretically. Although it is not robust with respect of the outliers in the predictor space, it is still useful in analyzing data for which the problem is mainly in the response (*Y*). Least Trimmed Squares (LTS) estimation is a high breakdown value method. S estimation is also a high breakdown value method. With the same breakdown value, it has a higher statistical efficiency than LTS estimation. MM estimation combines high breakdown value estimation and M estimation. It has both the high breakdown property and a higher statistical efficiency than S estimation.

M-estimator is the "maximum likelihood type" estimator. Instead of minimizing a sum of squares of the residuals in (2.1.3):

$$Q = \sum_{i=1}^{n} e_i^2 \text{ , where } e = Y - X\beta$$

M estimator $\hat{\beta}_M$ minimize

$$Q = \sum_{i=1}^{n} \rho\left(\frac{e_i}{\sigma}\right) \qquad (2.5.1)$$

with respect to the parameters $\beta_0, \beta_1, ..., \beta_{p-1}$. $\hat{\beta}_M$ is also a solution of *p* equations of form

$$\sum_{i=1}^{n} x_{ij} \Psi\left(\frac{e_i}{\sigma}\right) = 0 \text{ , j=1,…,p-1} \qquad (2.5.2)$$

If σ is unknown, it can be estimated using for example Huber or Tukey methods. With adding weight function, (2.5.2) becomes:

$$\sum_{i=1}^{n} x_{ij} \Psi\left(\frac{e_i}{\sigma}\right) w_{i\beta} = 0 \qquad (2.5.3)$$

where $w_{i\beta}$ is a weight function that can be chosen from a number of weight functions that are available for this method.

The least trimmed squares (LTS) estimate introduced by Rousseeuw (1984) is given by

$$\underset{\hat{\beta}}{Minimize} \sum_{i=1}^{h} e_{(i)}^2 \qquad (2.5.4)$$

where $e_{(1)}^2 \le \cdots \le e_{(n)}^2$ are the ordered squared residuals (note that the residuals are first squared and than ordered), and *h* is defined in the range $\frac{n}{2} + 1 \le h \le \frac{3n + p + 1}{4}$ (*p* is the number of predictors).

A new improved estimator was introduced recently by Yohai (1985) – MM method. Yohai's estimator is defined in three stages. In the first stage an initial high breakdown estimate $\hat{\beta}^*$ is calculated, such as LTS or S. Then, an M-estimate of scale *s*' is computed on the residuals $e_i(\hat{\beta}^*)$. Finally, find a local minimum $\hat{\beta}_{MM}$ of

$$Q_{MM}(\beta) = \sum_{i=1}^{n} \rho\left(\frac{e_i}{s'}\right) \qquad (2.5.5)$$

which satisfy $Q_{MM}(\hat{\beta}_{MM}) \leq Q_{MM}(\hat{\beta}^{*})$.

In the VSC data most of the outliers from X-space are rejected at the calibration stage. Since the problem with the outliers in the VSD data may be connected with the response direction (Y) as well as some outliers in X space may be left unattended, it is reasonable to employ the robust regression methods to the problem.

All four robust estimation procedures mentioned above were applied to the VSC data. The performance of each method was evaluated by average prediction error in percentage (PAPE). It was found that the robust regression model with MM estimation gives the lowest PAPE among the other three robust regression models. However, the robust regression MM procedure yields similar results as OLS in estimation of PAPE. It may mean that ordinary least squares method is not unduly influenced by outlying cases.

2.6 Local Regression

Robust regression requires knowledge of the regression function. When the appropriate regression function is not clear, nonparametric regression may be useful. Nonparametric regression fits are useful to obtain estimates of mean responses without specifying the nature of the response function.

The LOESS procedure implements a nonparametric method for estimating regression surfaces developed by Cleveland and Devin [8]. The LOESS method assumes that the predictor variables have already been selected, that the response function is smooth, and the error terms are approximately normally distributed with constant variance. The LOESS procedure allows great flexibility because no assumptions about the parametric form of the regression surface are needed.

Let *i*=1 to *n*, where *n* is the number of observations in the model. Then the *i*th value y_i of the response vector *Y* and the corresponding value x_i of the vector X of two predictors are related by

$$y_i = f(x_{i1}, x_{i2}) + \varepsilon_i \tag{2.6.1}$$

where *f* is the regression function that left unspecified and ε_i are independent $N(0, \sigma^2)$.

The basic idea of local regression or LOESS method is that that near $x_0 = (x_{01}, x_{02})$ the regression function f can be locally approximated by a member of a simple class of parametric functions. Such a local approximation is obtained by fitting a regression surface to the data points within a chosen neighborhood of the point x_0.

The LOESS method fits either a first-order model or a second-order model based on cases in the neighborhood. The radius of each neighborhood is chosen so that the neighborhood contains a specified percentage of the data points.

The size of the local neighborhoods is determined by the smoothing parameter value *s*. This parameter also controls the amount of smoothing being performed (small *s* – more smoothing). When $s < 1$, the local neighborhood used at a point x_0 contains the *s* fraction of the data points closest to the point x_0. When $s \geq 1$, all data points are used.

Suppose *q* denotes the number of points in the local neighborhoods and $d_1, d_2, \cdots, d_q$ denote the Euclidean distances in increasing order of the *q* points closest to x_0. The weight function used in the LOESS method is defined as follows:

$$w_i = \begin{cases} \left[1 - (d_i / d_q)^3\right]^3 & d_i < d_q \\ 0 & d_i \geq d_q \end{cases} \tag{2.6.2}$$

where *i*=1,…,*n*.

Thus, the weights are decreasing with distance. The points that are close to x_0 receive maximum weights and cases outside the neighborhood receive weight zero. The regression coefficients in a firs order or a second order model can be estimated by minimizing the locally weighted sum of squares.

In most cases, the local regression model based on VSC data cannot be used to predict all glucose values in the test VSC data. It happens because some points from the test VSC dataset are not contained in the box bounding the fitting data points.

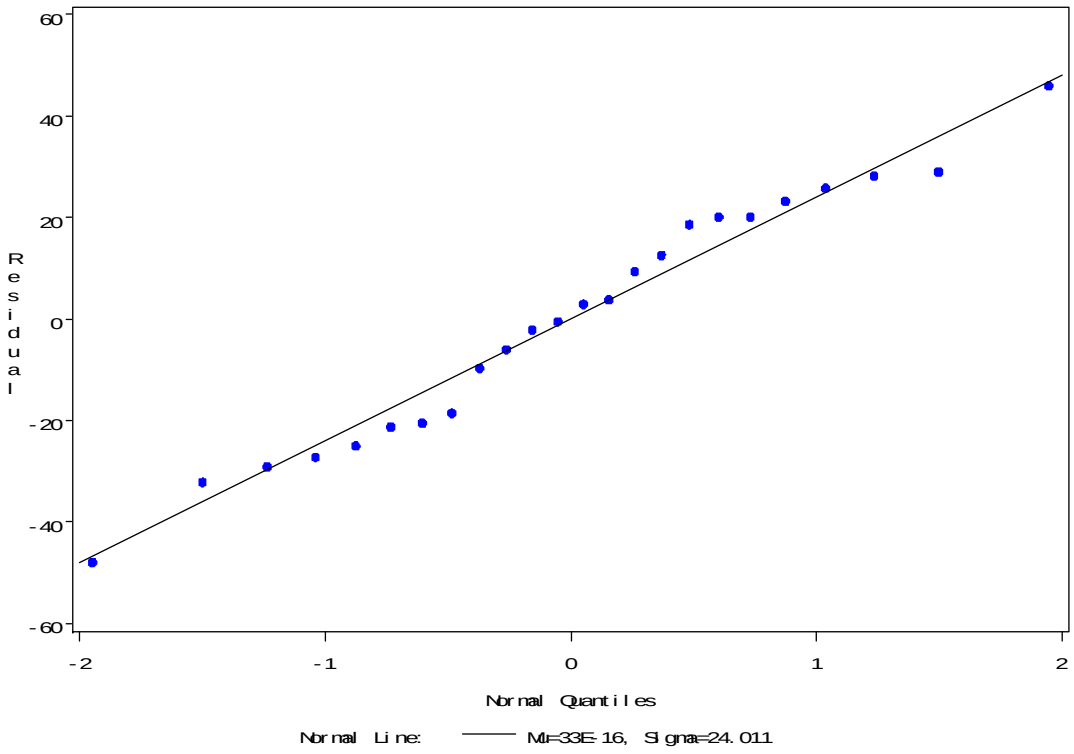
3. Data Analysis and Diagnostics

This chapter contains the analysis of VSC data. It was discussed before that the problems of non-normality and non-constancy of the regression errors, presence of the outliers and influential observations, multicollinearity may affect the regression model, and thus may make the model unfit for the accurate prediction. The methods for detecting these problems for the VSC data will be discussed in this section.

3.1 Checking for Normality of the Errors

For the OLS model as well as for many other regression models considered in this project we assume that the error terms ε_i in (2.2.1) are independently normally distributed. Since thee residuals e_i in (2.2.2) reflect the properties of ε_i, the normal probability plot of residuals is used to investigate the normality of the errors. In this plot each residual is plotted against its expected value under normality. Departure from normality is indicated by observations which do not lie close to the reference line. The normal plots of VSC data did not reveal serious departure from normality for most patients.

Figure 3.1: **Normal Plot for Patient #7**



Normal Plot of residuals for Patient #7 does not indicates serious departure from normal distribution.

3.2 Outliers Detection

Outliers and the problems that they may produce were described in Section 2.5. The outlying cases can occur both in calibration and in prediction (test) data. In calibration data that were used for model building such outlying observations may be detected and possibly discarded, while the outliers in the prediction data are usually hard or impossible to find. The presence of outlying observations in test data is very relevant when the regression model is built upon the Low/High glucose design. In this case outliers are connected with abnormally low or high predicted glucose values.

There are many different tools for detecting outliers in the calibration data. Some outlying observations may affect the fitted regression function. These cases are said to be the influential points. To measure the influence that *i*th observation has on $\hat{Y}_i$ the DFFIT value is used:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{-i}}{S_i \sqrt{h_{ii}}}$$

where $\hat{Y}_i$ – predicted value, and $\hat{Y}_{-i}$ – is the predicted value when *i*th case is omitted.

DFFIT measures the difference between the predicted value with and without the data point. A large value indicates that the observation is very influential. For small dataset like VSC data are, the absolute value of DFFIT greater than 1 indicates the influential case.

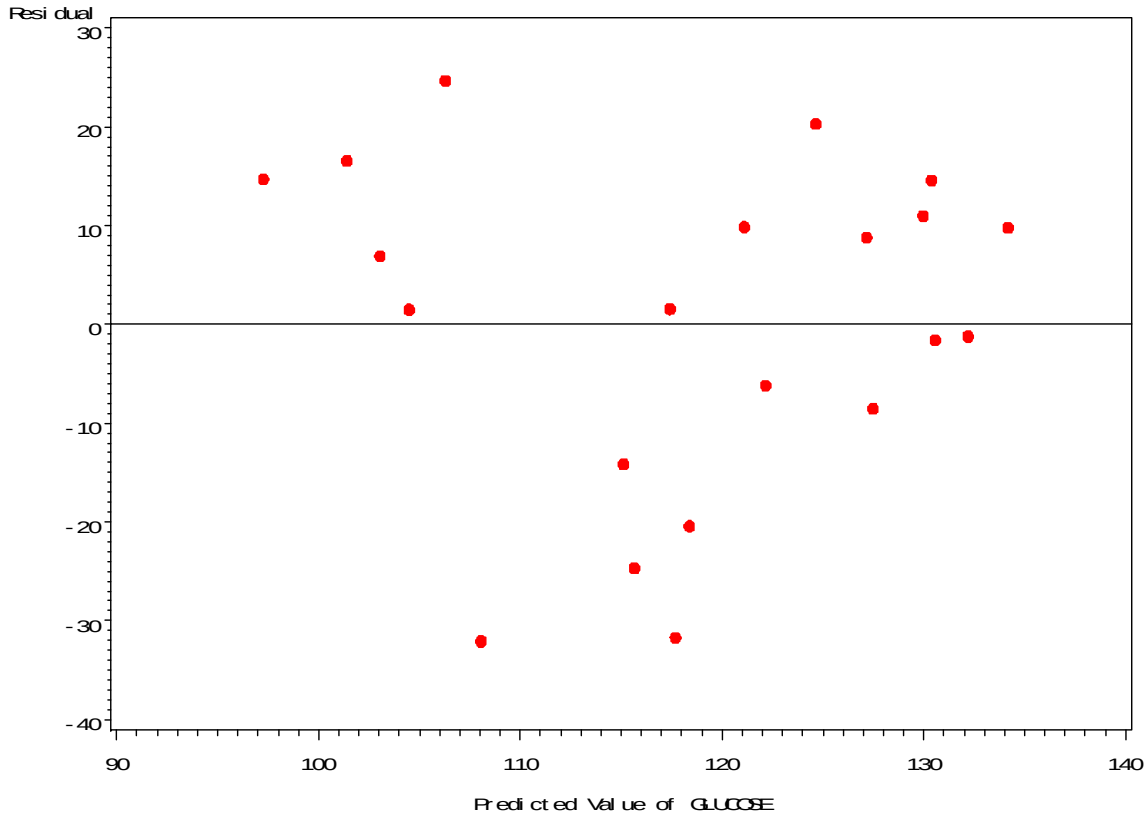
All VSC data have outlying cases with different degree of influence. These cases may affect the regression equation, and thereby the prediction of the glucose value may be not accurate. So that application of the robust regression method to the data seems reasonable here.

3.3 Detecting Heteroscedasticity

When **var(**ε_i**)** is not a constant for *i*=1,..n, this condition is called heteroscedasticity (unequal error variances). This causes variances of parameter estimates to be large and can affect tests substantially (for example the general linear hypothesis test). To examine heteroscedasticity, the residual plots can be used. In these plots residuals are plotted against the fitted values $\hat{Y}_i$ or against each of the predictor variables. Residual plots are also useful in detection of outliers in the data where outlying

observations are represented by cases that are separated from the group of points and are located far away from the reference zero line. Ascending or descending band of residuals indicates the existence of heteroscedasticity. The residual plots for patient #1, #5, and #7 of VSC data indicate the non-constancy of the error variances.

Figure 3.2: Residual Plot for Patient #5



Residual plot for patient #5 displays the band of residuals narrowing to the right showing non-constant variance.

3.4 Detecting Multicollinearity

Multicollinearity (or collinearity) may create great difficulty in estimation of regression coefficients. This problem and its effects were described in section 2.3. Variance inflation factor (VIF) is used to measure collinearity. The VIF exists for each *X* variable and measures the increase in variance compared to when the predictor variables are not linearly related. Variance inflation factors are the diagonal elements of $(X^T WX)^{-1}$ from (2.2.5), where *W* is $n \times n$ diagonal matrix containing the weights ω_i. Large VIF value among all *X* variables indicates a serious multicollinearity. The VIFs for VSC data do not reveal severe multicollinearity: VIFs are greater then 1 but less that 10.

Table 3.1: Variance Inflation Factors for Patient #7

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	VIF
Intercept X1	1	118.30672	6.11315	19.35	<.0001	0
X2	1	326.11734	70.09382	4.65	0.0002	1.17620
X3	1	8398.78327	3423.76695	2.45	0.0235	1.20403
	1	429.34856	123.86042	3.47	0.0024	1.30737

3.5 Summary

In this project data for seven patients were analyzed. The diagnostics of the data did not reveal serious departure from normality of the error terms for most patients as well as severe multicollinearity. However, the presence of influential outliers was detected in each dataset. The heteroscedasticity in some VSC data was also found.

4. Methods Comparison

This chapter contains the results of six regression methods in prediction of glucose values for seven different patients. For each patient the regression methods results were evaluated by comparison average prediction error in percentage (PAPE) and percent of acceptable points according to the "± 20% rule". Tables in this chapter contain the results for six regression models. In each cell of the table the first number is the average prediction error in percentage and the number in parentheses is the percent of acceptable points according to the "± 20%" criterion. Graphs represent the predicted plots of three regression models along with the reference glucose polyline that shown in red. The regression models in the graphs are PLS, ridge and OLS.

4.1 Data Designs

Two data design methods were considered in this project. Original design is based on using all calibration measurements to build the regression model and use the rest of the data (test data) for glucose prediction. The calibration process may take several hours or even several days. During this procedure, the patients should be in the test room. The calibration data may contain up to 40 data points. Another design was considered in this work to reduce the number of observations that are used for the model building, to minimize the waiting time and number of tests for the patients. In this approach, the observations with only low and high reference glucose values from the calibration part were used for the modeling. This design is called Low/High glucose value design and may allow one to obtain the adequate regression model with a smaller number of the

observations. Original and Low/High glucose value designs were compared in this project in prediction of glucose levels.

4.2 Patient #1

The VSC data for patient #1 contains 58 observations: 40 are from calibration data and 18 are from test data. The glucose range in the calibration part is within 49 to 244 mg/dl. The wide glucose range and the sufficient number of data points allow one to use High/Low glucose value design.

Table 4.1: Regression Methods Comparison for Patient #1

	Robust	**WLS**	**Local**	**Ridge***	**PLS**	**OLS**
Original design						
Original data (40 obs.)	**32.38 (38.89)**	**35.18 (50.00)**	**34.82 (50.00)**	**36.12 (44.44)**	**34.12 (50.00)**	**34.77 (50.00)**
w/o outliers	**25.41 (41.18)**	**25.42 (47.06)**	**26.30 (47.05)**	**27.17 (47.06)**	**25.34 (47.06)**	**25.88 (47.06)**
High/Low design**						
5 low - 5 high	**27.23 (41.18)**	**27.96 (47.06)**	**30.88 (38.46)**	**32.72 (29.41)**	**34.05 (41.18)**	**27.23 (41.18)**
* c=0.5 – biasing constant ** one outlier from pred. is deleted						

As shown in Table 4.1, PLS and OLS models give similar results. Using the High/Low glucose value design allows to reduce the number of data points in the model from 40 to 15 and obtain satisfactory prediction of the glucose. However, High/Low glucose values method produces one outlier in the prediction (Y) space.

Figure 4.1: OLS, PLS, and Ridge Models Performance for Patient #1

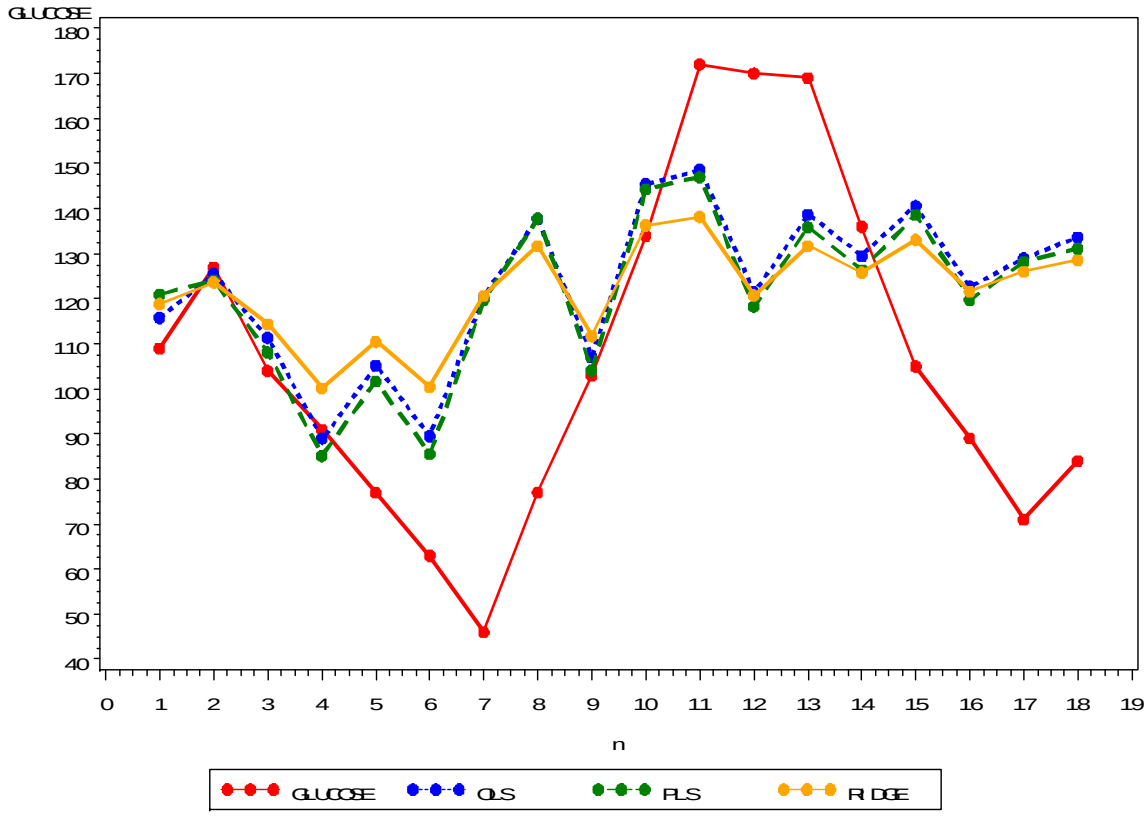


Figure 4.1 shows that the OLS and PLS are very close in the prediction of the glucose level. The models predict well for the glucose range of 90 to 150 mg/dl.

4.3 Patient #2

The VSC data for patient #2 contains 48 observations: 14 are from calibration data and 34 are from test data. The glucose range in the calibration part is small – from 68 to 96 mg/dl. High/Low glucose value design is not used for this patient, since the glucose range is small and the number of data points is not sufficient for this type of design.

Table 4.2: Regression Methods Comparison for Patient #2

	Robust	**WLS**	**Local**	**Ridge***	**PLS**	**OLS**
Original design						
original data (14 obs.)	**29.87 (44.18)**	**29.99 (47.06)**	**38.52 (16.6)**	**29.82 (41.18)**	**29.58 (41.18)**	**29.88 (44.12)**
* c=0.1 – biasing constant						

The results for robust, ridge, PLS and OLS regressions are very close, while weighted method gives the highest percent of acceptable points.

Figure 4.2: OLS, PLS, and Ridge Models Performance for Patient #2

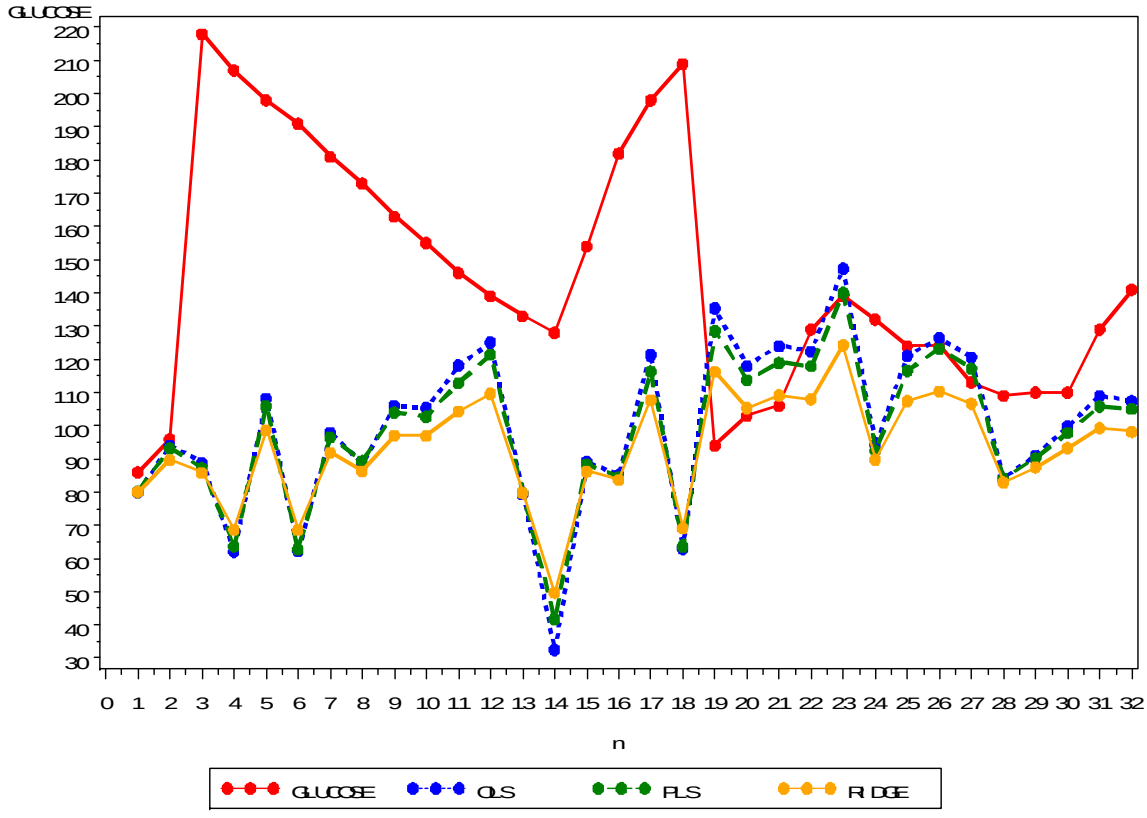


Figure 4.2 shows that three models predict well in the glucose range of 40 to 150 mg/dl. Regression models do not predict accurately for high glucose level. Probably, this is due to the regression models built upon the data with low glucose values.

4.4 Patient #3

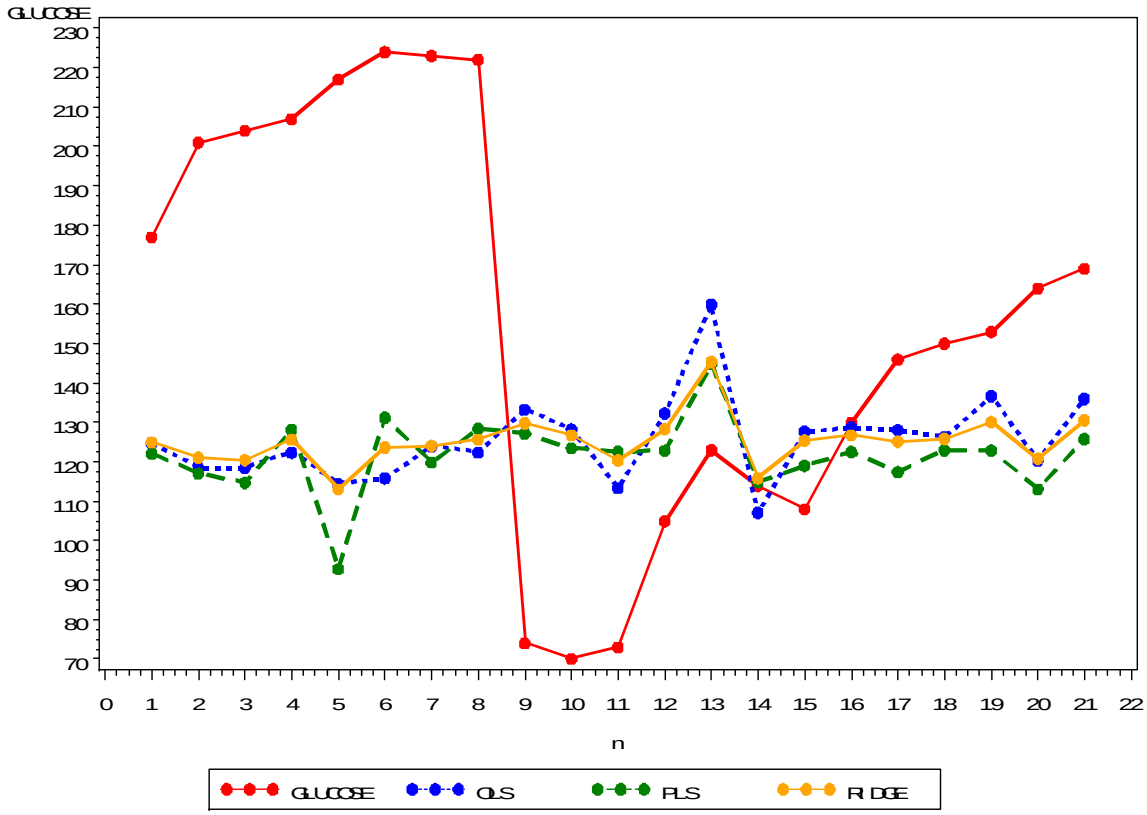
The VSC data for patient #3 contains 41 observations: 20 are from calibration data and 21 are from test data. The glucose range in the calibration part is from 73 to 220 mg/dl.

Table 4.3: Regression Methods Comparison for Patient #3

	Robust	**WLS**	**Local**	**Ridge***	**PLS**	**OLS**
Original design						
Original data (20)	**34.68 (28.57)**	**33.71 (28.57)**	**36.41 (35.71)**	**33.63 (33.33)**	**34.44 (38.09)**	**34.42 (33.33)**
* c=0.5 – biasing constant						

According to Table 4.3, PLS and ridge are two the best methods: PLS gives the highest percent of acceptable points, and ridge method gives the lowest average prediction error.

Figure 4.3: OLS, PLS, and Ridge Models Performance for Patient #3



OLS, PLS, and Ridge predict in the glucose range of 100 to 160 mg/dl.

4.5 Patient #4

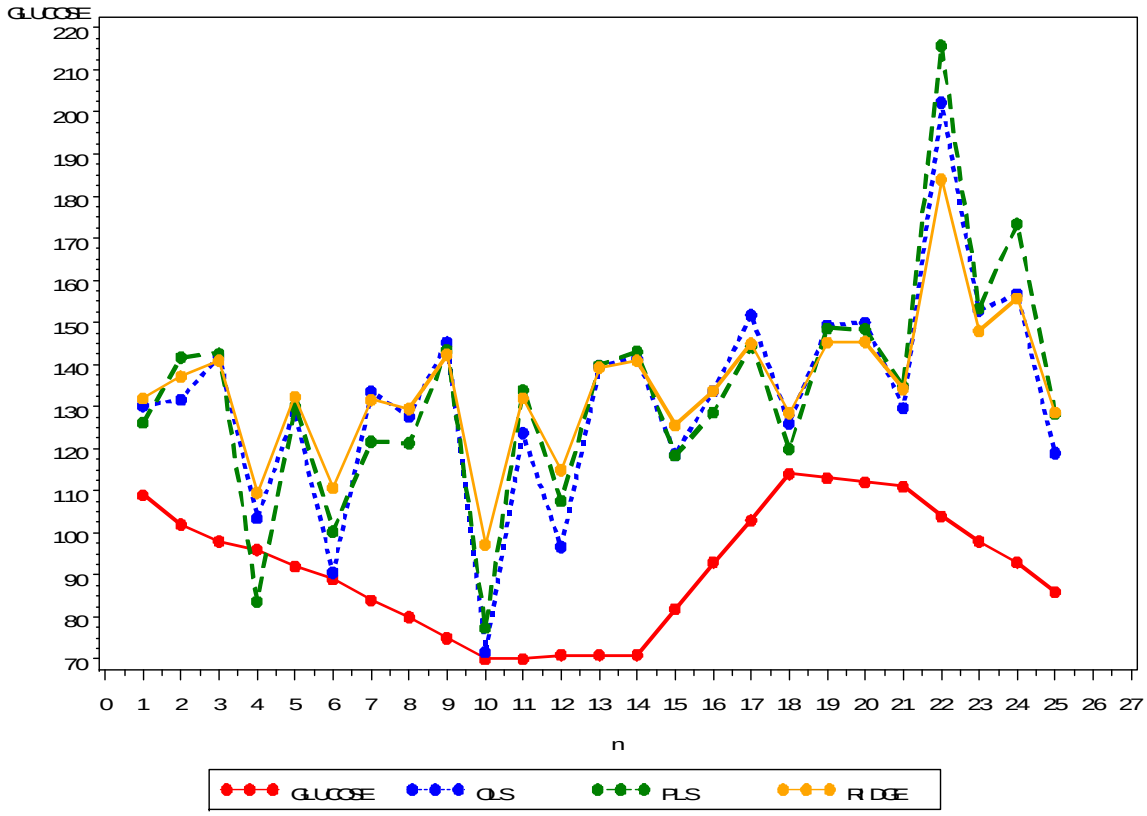
The VSC data for patient #4 contains 39 observations: 14 are from calibration data and 25 are from test data. The glucose range in the calibration part is from 115 to 158 mg/dl.

Table 4.4: Regression Methods Comparison for Patient #4

	Robust	**WLS**	**Local**	**Ridge***	**PLS**	**OLS**
Original design						
Original data (14)	**47.52 (24.00)**	**46.14 (16.00)**	**47.45 (16.66)**	**47.33 (24.00)**	**48.81 (20.00)**	**46.13 (24.00)**
* c=0.1 – biasing constant						

There is no substantial improvement over OLS model in the glucose prediction.

Figure 4.4: OLS, PLS, and Ridge Models Performance for Patient #4



Regression models tend to overestimate the true glucose levels. This happens because the regressions were built upon the data with intermediate to high glucose values, but used for predicting low glucose values.

4.6 Patient #5

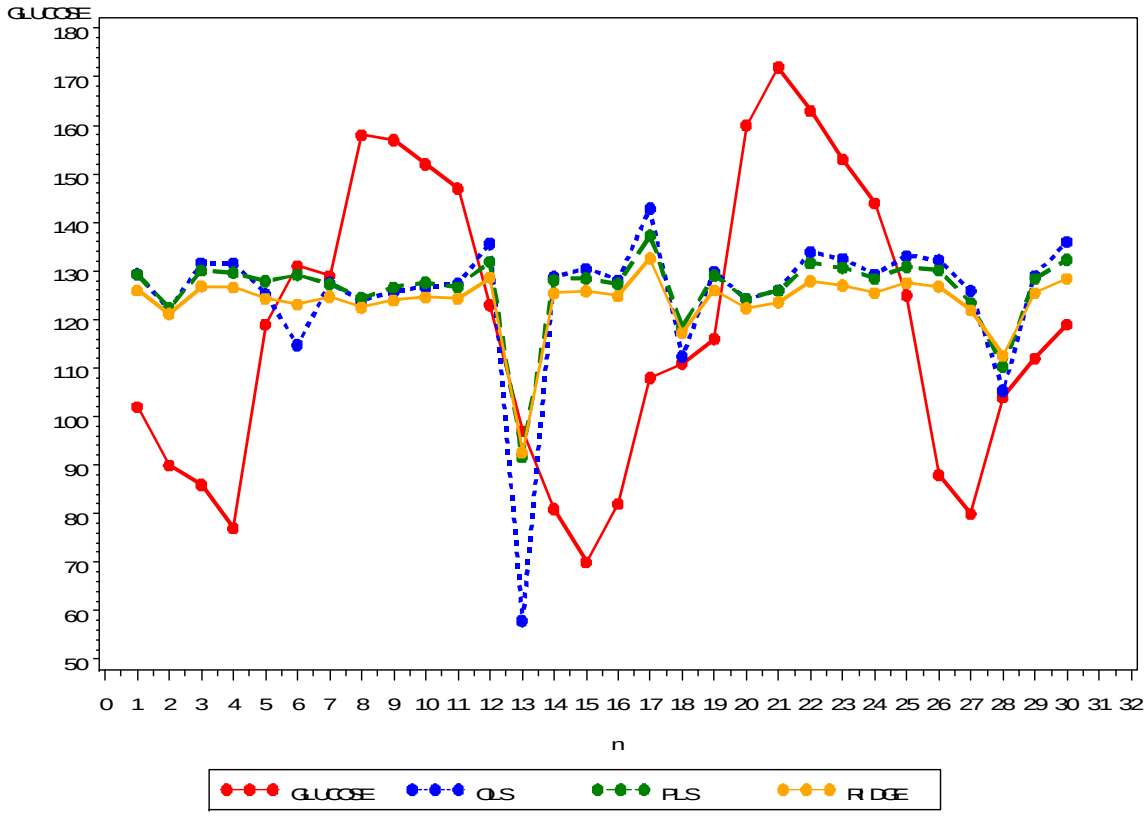
The VSC data for patient #5 contains 51 observations: 21 are from calibration data and 30 are from test data. The glucose range in the calibration data is from 76 to 145 mg/dl.

Table 4.5: Regression Methods Comparison for Patient #5

	Robust	**WLS**	**Local**	**Ridge****	**PLS**	**OLS**
Original design						
Original data (21)	**27.00 (53.33)**	**28.73 (53.33)**	**32.11 (45.45)**	**24.13 (50.00)**	**25.03 (56.66)**	**27.00 (53.33)**
High/Low design*						
6 low - 5 high	**31.47 (44.82)**	**29.96 (41.37)**	**34.72 (36.36)**	**26.01 (51.72)**	**28.58 (50.00)**	**31.47 (44.82)**
* one outlier is deleted ** c=0.8 – biasing constant						

The obtained results show that PLS and ridge are the best methods in the glucose prediction. Use of High/Low glucose value design allowed to reduce the number of observations from 21 to 11 and obtain satisfactory prediction of the glucose values.

Figure 4.5: OLS, PLS, and Ridge Models Performance for Patient #5



Models have tendency to predict mainly in the glucose range of 110 to 140 mg/dl.

4.7 Patient #6

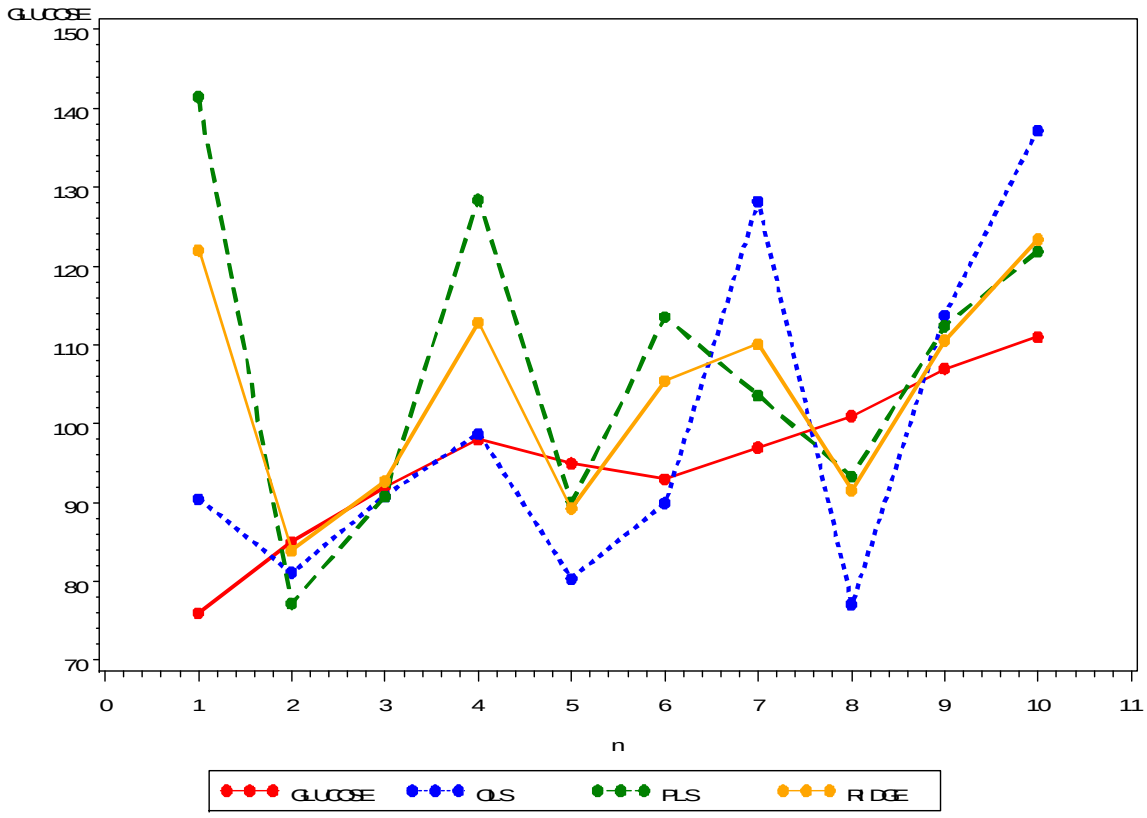
The VSC data for patient #6 contains 17 observations: 7 are from calibration data and 10 are from test data. The glucose range in the calibration data is from 67 to 171 mg/dl.

Table 4.6: Regression Methods Comparison for Patient #6

	Robust	**WLS**	**Local**	**Ridge**	**PLS**	**OLS**
Original design						
original data (7)	**14.27 (62.5)**	**13.36 (60.00)**	**12.27 (66.66)**	**13.44 (90.00)**	**18.42 (70.00)**	**13.00 (70.00)**
* c=0.5 – biasing constant						

Ridge is the best method with 90% of acceptable points

Figure 4.6: OLS, PLS, and Ridge Models Performance for Patient #6



4.8 Patient #7

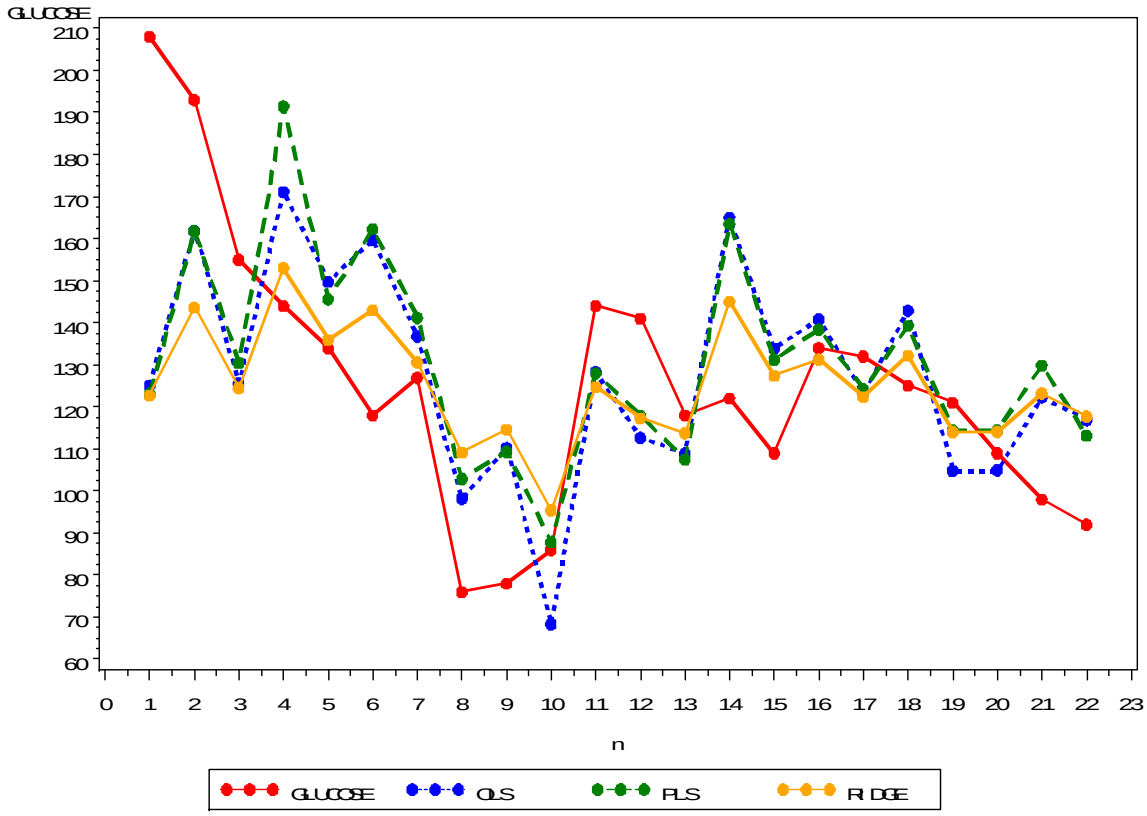
The VSC data for patient #7 contains 46 observations: 24 are from calibration data part and 22 are from test data. The glucose range in the calibration part is within 81 to 208 mg/dl. The wide glucose range and the sufficient number of data points allows one to use High/Low glucose value design.

Table 4.7: Regression Methods Comparison for Patient #7

	Robust	**WLS**	**Local**	**Ridge***	**PLS**	**OLS**
Original design						
original data (24)	**19.61 (54.50)**	**20.24 (50.00)**	**19.88 (61.90)**	**16.74 (68.18**)	**19.01 (59.09)**	**19.61 (54.45)**
w/o outliers	**19.05 (57.14)**	**19.96 (61.90)**	**18.16 (57.90)**	**16.45 (61.90)**	**17.53 (71.43)**	**19.05 (57.14)**
High/Low design**						
7 low - 7 high	**18.55 (57.14)**	**18.50 (61.90)**	**20.82 (50.00)**	**17.01 (61.90)**	**19.02 (66.67)**	**18.55 (57.14)**
w/o outliers	**18.81 (55.00)**	**18.84 (57.14)**	**19.59 (56.25)**	**16.71 (55.00)**	**17.60 (66.67)**	**18.81 (55.00)**
High/Low design**						
6 low - 6 high	**18.78 (57.14)**	**18.72 (57.14)**	**22.87 (45.45)**	**16.92 (61.90)**	**17.77 (66.67)**	**18.78 (57.14)**
w/o outliers	**18.65 (55.00)**	**18.44 (57.14)**	**20.93 (45.45)**	**16.16 (55.00)**	**18.15 (66.67)**	**18.65 (55.00)**
* c=0.5 ** one outlier is deleted						

According to the results for patient #7, ridge regression is the best method that gave the smallest PAPE and the largest percent of acceptable points. Use of High/Low glucose value design reduced the number of observations in twice and allowed to obtain good prediction of the glucose values.

Figure 4.7: OLS, PLS, and Ridge Models Performance for Patient #7



The models predict well in the wide glucose range of 70 to 180 mg/dl.

5. Conclusion

5.1 Summary

Based on the analysis of the results of six regression models for seven patients, PLS and ridge showed the best results in the glucose prediction. For five out of seven patients, PLS and/or ridge predict better than the other models analyzed in this project including OLS. Thus, use of PLS or ridge regression methods may reduce percent of average prediction error to 15%, and increase percent of acceptable points up to 22%. PLS and ridge regression models are close to OLS model, since they are just the improved modification of OLS. So, if OLS does not predict well, PLS and ridge may make the prediction more precise, but not too much.

Regression models that were considered in this project predict well in some specific glucose range. For most patients this range is within 90 to 150 mg/dl. If the regression model built upon the data with small glucose range, this makes the built regression function useless in the prediction the glucose value outside this range. Thus, for accurate glucose prediction, it is important to use the data with wide range of the glucose values for the regression model building.

Using Low/High glucose value design allowed to obtain an adequate regression model with a smaller number of the observations. The regression models built upon the data with low and high glucose values and a small number (14 or less) of the observations predicts relatively well: the average prediction errors and the percent of acceptable points are close or the same as in the original design model. One drawback of the Low/High glucose design is that it produces the outliers in the test data. These outlying cases are

connected with abnormally high (greater that 500) or low (less than 40) predicted glucose values. These abnormal predicted glucose values should not be considered as true values.

5.2 Future Work

Since there are many outliers and influential observations in the data and ridge regression method gave better results in the prediction of the glucose values, the next step of the work is to try the robust ridge regression. The robust ridge regression method is a combination of properties of the robust estimation and the ridge regression. This method allows to protect against outliers in the data and shrink the regression coefficients toward zero making the estimator variance smaller. There are several approaches to combine the properties of robust estimators with ridge estimators. For example, ridge regression based on the robust choice of shrinkage parameter c in (2.4.3.) can be used.

In this project data for seven patients were researched in the prediction of the glucose level. For most patients PLS and Ridge are the best regression methods that allow to improve the accuracy of the glucose prediction. It would be interesting to determine in the future work if these methods give the same beneficial results in the prediction of the glucose level for more patients.

References

1. *Diabetes 4-1-1: Facts, Figures, and Statistics at a Glance*, American Diabetes Association, 2004
2. *Diabetes Control and Complications Trial (DCCT)*, New England Journal of Medicine, 329(14), September 30, 1993
3. "*Noninvasive Blood Glucose Monitors*," Mini-factsheet, National Diabetes Information Clearinghouse, March 1998
4. The Whitaker foundation www.whitaker.org
5. W. L. Clarke, D. Cox, L. A. Conder-Frederick, W. Charter, and S. L. Pohl "*Evaluating Clinical Accuracy of Systems for Self-Monitoring of Blood Glucose*", Diabetes care, Vol. 10, No. 5, 1987
6. M. Lewis-Beck, A. Bryman, and T. Futing. "Partial *least squares regression (PLS-regression),*" *in Encyclopedia for Research Methods for the Social Sciences*, Thousand Oaks, CA: Sage, 2003
7. Colin Chen. "*Robust Regression and Outlier Detection with the ROBUSTREG Procedure*, SAS Institute Inc., Cary, NC 2002
8. Cleveland, W.S., Devlin, S.J., and Grosse, E. Regression By Local Fitting, (1988), *Journal of Econometrics*, 37, 87 -114
9. Robert A. Cohen. *An Introduction to PROC LOESS for Local Regression*, SAS Institute Inc. Cary, North Carolina, USA
10. J. Neter, W. Wasserman and M.H. Kutner. *Applied Linear Statistical Models*. Irwin Publishing Company, Boston, 1996
11. Clive Loader. *Local Regression and Likelihood.* New York: Springer, 1999
12. Norman R. Draper, Harry Smith. *Applied regression analysis*. Thousand Oaks, John Wiley & Sons, 1998
13. P. J. Rousseeuw , A. M. Leroy, *Robust regression and outlier detection.* John Wiley & Sons, Inc., New York, NY, 1987

Appendix

SAS codes

Data Diagnostics

```
data dat;
infile patient7;
input ID GLUCOSE X1 X2 V1-V7 V8 V9 X3 V10-V17;
run;

*Calibration data;
data model;
set dat;
if _n_ gt 24 then delete;
run;

*Test data;
data test;
set dat;
if _n_ le 24 then delete;
drop glucose;
run;

*Outliers detection;
proc reg data=model;
model GLUCOSE=X1 X2 X3/influence ;
output out=model1 r=resid p=predict;
run;

*Multicollinearity detection;
proc reg data=model;
model GLUCOSE=X1 X2 X3/vif ;
run;

*Normal Probability Plot of Residuals;
goptions reset=all;
title 'Normal probability plot of residuals';
symbol1 c=blue v=dot h=.8;
proc capability data=model1 noprint;
qqplot resid;
run;

*Constancy of the Error Variance;
goptions reset=all;
title 'Errors vs. Ppredicted';
symbol1 i=none v=dot c=red;
proc gplot data=model1;
plot resid*predict/vref=0;
run;
```

Ordinary Least Squares Regression

```
*Building the OLS regression model for the first 24 observations of the data,
prediction the glucose values for the rest;

*Obtaining
- percent of acceptable points according to the "+/-20%" rule
- average prediction error;

data dat;
infile patient7;
input ID GLUCOSE X1 X2 V1-V7 V8 V9 X3 V10-V17;
run;

*Calibration data;
data model;
set dat;
if _n_ gt 24 then delete;
run;

*Test data;
data test;
set dat;
if _n_ le 24 then delete;
run;

proc reg data=model OUTEST=out;
model GLUCOSE=X1 X2 X3;
run;

*Prediction glucose value for the test data;
proc score data=test score=out out=pred type=parms;
var X1 X2 X3;
run;

data pred;
set pred;
err=glucose-model1;
err=abs(err);
errperc=err*100/glucose;
if errperc gt 20 then p=0;
if errperc le 20 then p=1;
run;

*Computing percent of acceptable points according to the "+/-20%" rule;
proc iml;
use pred;
read all var {p} into y;
n=nrow(y);
acc=t(y)*y; *acc - number of acceptable points;
accpct=acc*100/n; *accpct - percent of acceptable points;
print acc accpct;
run;

*Computing average prediction error;
proc means data=pred;
var errperc err; *errperc - average prediction error in percentage; *err - average prediction error;
run;
```

Weighted Least Squares Regression

```
*Building the WLS regression model for the first 24 observations of the data,
prediction the glucose values for the rest;

*Obtaining
- percent of acceptable points according to the "+/-20%" rule
- average prediction error;

data dat;
infile patient7;
input ID GLUCOSE X1 X2 V1-V7 V8 V9 X3 V10-V17;
run;

*Calibration data;
data model;
set dat;
if _n_ gt 24 then delete;
run;

*Test data;
data test;
set dat;
if _n_ le 24 then delete;
run;

proc reg data=model;
model GLUCOSE=X1 X2 X3;
output out=b2 r=resid p=pred;
run;

*Computing the absolute and squared residuals;
data b2;
set b2;
absr=abs(resid);
sqrr=resid*resid;
run;

proc reg data=b2;
model absr=X1 X2 X3;
output out=b3 p=shat;

*Computing weights;
data b3;
set b3;
wt=1/(shat*shat);

*Wieghted regression;
proc reg data=b3 OUTEST=wmodel;
model GLUCOSE=X1 X2 X3;
weight wt;
run;

*Prediction glucose value for the test data;
proc score data=test score=wmodel out=pred type=parms;
var X1 X2 X3;
run;

data pred;
set pred;
```

```
err=glucose-model1;
err=abs(err);
errperc=err*100/glucose;
if errperc gt 20 then p=0;
if errperc le 20 then p=1;
run;

*Computing percent of acceptable points according to the "+/-20%" rule;
proc iml;
use pred;
read all var {p} into y;
n=nrow(y);
acc=t(y)*y; *acc - number of acceptable points;
accpct=acc*100/n; *accpct - percent of acceptable points;
print acc accpct;
run;

*Computing average prediction error;
proc means data=pred;
var errperc err; *errperc - average prediction error in percentage; *err - average prediction error;
run;
```

Partial Least Squares Regression

```
*Building the PLS regression model for the first 24 observations of the data,
prediction the glucose values for the rest;

*Obtaining
- percent of acceptable points according to the "+/-20%" rule
- average prediction error;

data dat;
infile patient7;
input ID GLUCOSE X1 X2 V1-V7 V8 V9 X3 V10-V17;
run;

*Calibration data;
data model;
set dat;
if _n_ gt 24 then delete;
run;

*Test data;
data test;
set dat;
if _n_ le 24 then delete;
drop glucose;
run;

data all;
set model test;
run;

proc pls data=all nfac=1;
model GLUCOSE=X1 X2 X3;
output out=pred p=predgluc;
run;

data pred;
set pred;
if _n_ le 24 then delete;
keep id predgluc;
run;

data glucose;
set dat;
if _n_ le 24 then delete;
keep id glucose;
run;

data predicted;
merge glucose pred;
err=glucose-predgluc;
err=abs(err);
errperc=err*100/glucose;
if errperc gt 20 then p=0;
if errperc le 20 then p=1;
run;

*Computing percent of acceptable points according to the "+/-20%" rule;
proc iml;
use predicted;
```

```
read all var {p} into y;
n=nrow(y);
acc=t(y)*y; *acc - number of acceptable points;
accpct=acc*100/n; *accpct - percent of acceptable points;
print acc accpct;
run;

*Computing average prediction error;
proc means data=predicted;
var errperc err; *errperc - average prediction error in percentage; *err - average prediction error;
run;
```

Ridge Regression

```
*Building ridge regression model for the first 24 observations of the data, prediction the glucose values for the rest;

*Obtaining
- percent of acceptable points according to the "+/-20%" rule
- average prediction error;

data dat;
infile patient7;
input ID GLUCOSE X1 X2 V1-V7 V8 V9 X3 V10-V17;
run;

*calibration data;
data model;
set dat;
if _n_ gt 24 then delete;
run;

*test data;
data test;
set dat;
if _n_ le 24 then delete;
run;

proc reg data=model OUTEST=out;
model GLUCOSE=X1 X2 X3/ridge = 0.500;
output out=model2 r=resid p=predict;
run;

data out;
set out;
if _RIDGE_ = . then delete;
run;

*prediction glucose value for the test data;
proc score data=test score=out out=pred type=ridge;
var X1 X2 X3;
run;

proc print data=predicted;
run;

data pred;
set pred;
err=glucose-model1;
err=abs(err);
errperc=err*100/glucose;
if errperc gt 20 then p=0;
if errperc le 20 then p=1;
run;

*computing percent of acceptable points according to the "+/-20%" rule;
proc iml;
use pred;
read all var {p} into y;
n=nrow(y);
acc=t(y)*y; *acc - number of acceptable points;
accpct=acc*100/n; *accpct - percent of acceptable points;
print acc accpct;
```

```
run;

*computing average prediction error;
proc means data=pred;
var errperc err; *errperc - average prediction error in percentage; *err - average prediction error;
run;
```

Robust Regression

```
*Robust regression with MM method and Yohai's optimal function for MM estimate;

*Building robust regression model for the first 24 observations of the data,
prediction the glucose values for the rest;

*Obtaining
- percent of acceptable points according to the "+/-20%" rule
- average prediction error;

data dat;
infile patient7;
input ID GLUCOSE X1 X2 V1-V7 V8 V9 X3 V10-V17;
run;

*Calibration data;
data model;
set dat;
if _n_ gt 24 then delete;
run;

*Test data;
data test;
set dat;
if _n_ le 24 then delete;
run;

*Robust regression MM method;
proc robustreg data=model method=mm(chif=yohai) outest=rmodel;
model GLUCOSE=X1 X2 X3;
run;

proc score data=test score=rmodel out=pred type=parms;
var X1 X2 X3;
run;

data pred;
set pred;
model1=_;
err=glucose-model1;
err=abs(err);
errperc=err*100/glucose;
if errperc gt 20 then p=0;
if errperc le 20 then p=1;
run;

*Computing percent of acceptable points according to the "+/-20%" rule;
proc iml;
use pred;
read all var {p} into y;
n=nrow(y);
acc=t(y)*y; *acc - number of acceptable points;
accpct=acc*100/n; *accpct - percent of acceptable points;
print acc accpct;
run;

*Computing average prediction error;
proc means data=pred;
var errperc err; *errperc - average prediction error in percentage; *err - average prediction error;
run;
```

Local Regression

```
*Building local regression model for the first 24 observations of the data,
prediction the glucose values for the rest;

*Obtaining
- percent of acceptable points according to the "+/-20%" rule
- average prediction error;

data dat;
infile patient7;
input ID GLUCOSE X1 X2 V1-V7 V8 V9 X3 V10-V17;
run;

*Calibration data;
data model;
set dat;
if _n_ gt 24 then delete;
run;

*Test data;
data test;
set dat;
if _n_ le 24 then delete;
run;

proc loess data=model;
SCORE data=test ID=(X1 X2 X3)/print;
ods output ScoreResults=pred;
model GLUCOSE=X1 X2 X3;
run;

data pred;
set pred;
err=glucose-p_glucose;
if err=. then delete;
run;

data pred;
set pred;
err=abs(err);
errperc=err*100/glucose;
if errperc gt 20 then p=0;
if errperc le 20 then p=1;
run;

*Computing percent of acceptable points according to the "+/-20%" rule;
proc iml;
use pred;
read all var {p} into y;
n=nrow(y);
acc=t(y)*y; *acc - number of acceptable points;
accpct=acc*100/n; *accpct - percent of acceptable points;
print acc accpct;
run;

*Computing average prediction error;
proc means data=pred;
var errperc err; *errperc - average prediction error in percentage; *err - average prediction error;
run;
```

P-values Plot

```
*obtaining p-values for the first 5,8,...n observations;

data dat;
infile patient7;
input ID GLUCOSE X1 X2 V1-V7 V8 V9 X3 V10-V17;
run;

*Macro to perform p-value calculating for 5,8,...n observations;

%macro pv;
data pval;
run;
%do l = 5 %to 46;

    data model&l;
    set dat;
    if _n_ gt &l then delete;
    run;

    proc reg data=model&l;
            model GLUCOSE=X1 X2 X3;
    run;

*p-value calculating;
proc iml;
            use model&l;
            read all var {X1 X2 X3} into x;
            read all var {GLUCOSE} into y;
  n=nrow(x);
            p=ncol(x);
            x=J(n,1)||x;

            xpx=t(x)*x;
            ixpx=inv(xpx);

            H=x*inv(t(x)*x)*t(x);
            betahat=ixpx*t(x)*y;
            resid=(I(n)-H)*y;

            sse=t(resid)*resid;
            mse=sse/(n-p-1);

            ssr=t(y)*(H-J(n,n,1/n))*y;
            msr=ssr/p;

            alpha=0.05;
            fstar=msr/mse;
            fcrit=finv(1-alpha,p,n-p-1);
            pvalue=1-probf(fstar,p,n-p-1);
            create pval&l var{fstar, pvalue};
            append;
  close pval&l;

run;

data pval;
set pval pval&l;
run;
```

```
%end;
data pval;
set pval;
if _n_=1 then delete;
run;
%mend;
%pv;


data pval;
set pval;
N=_n_+4;
run;

*ploting p-value vs. number of observations in the model;
goptions reset=all;
title 'p-value vs. number of observations in the model';
symbol1 i=joint v=dot c=red;
proc gplot data=pval;
plot pvalue*N/vref=0.01;
run;
```