Statistical techniques for prediction using semi-quantitative data from micro-arrays

Categorical regression compared to Random Forests

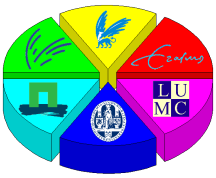
Annemariek Driessen

Master thesis defended on September 14^{th}, 2012

Advisors:
Dr. Romke Bontekoe
Dr. Anita J. van der Kooij
Prof. Dr. Jacqueline J. Meulman

Specialization: Statistical Science for the Life and Behavioural Sciences

Universiteit Leiden



Mathematisch Instituut Leiden

Abstract

The data used for this thesis are data about Bacterial Vaginosis (BV) and they have some special characteristics. The numerical values are semi-quantitative, the response is categorical (BV negative, intermediate and BV positive) and the data are high-dimensional.

Categorical regression (CATREG) is a method that can be used to analyze these data. To determine how CATREG performs in predicting future outcomes from these data it will be compared to Random Forests, one of the golden standards in statistical learning.

The dataset was randomly divided in a training and test set. The training set was used for variable selection and determining the values of the regularization parameters, and the test set was used for estimating the prediction accuracy. Based on the training set a Random Forests model and a CATREG model were chosen and used for prediction.

Random Forests and CATREG both classify 68% of the outcomes correctly, but the models are not able to distinguish well between intermediate and BV positive women. When the intermediate and BV positive women are taken together, the percentages of correctly classified women increases to 95% and 97% for Random Forest and CATREG, respectively.

Overall this analysis showed that CATREG performs as well as Random Forests in the prediction and therefore it can be considered as a worthwhile alternative.

Contents

1 Introduction **1**

2 Methods **3**
2.1 Random Forests . . . 3
2.2 CATREG . . . 4
2.3 Ridge and Lasso . . . 6
2.4 The Lasso within CATREG . . . 7
2.5 Comparing the methods . . . 7

3 Predicting Bacterial Vaginosis **8**
3.1 Description . . . 8
3.2 Analysis . . . 9
3.2.1 Data . . . 9
3.3 Random Forests . . . 10
3.3.1 Model building . . . 10
3.3.2 Testing . . . 11
3.4 CATREG . . . 12
3.4.1 Model building . . . 12
3.4.2 Testing . . . 14
3.5 Comparison results . . . 17
3.6 Dichotomization of the data . . . 18

4 Discussion and Conclusion **20**

A Bacteria names **22**

B Transformation plots **23**

1 Introduction

This project was done at the "Leids Cytologisch en Pathologisch Laboratorium" (LCPL) in Leiden, which performs cytological and pathological diagnostics for patients of general practitioners. The data used for this thesis are data with some special characteristics. The total dataset represents 111 women and 456 predictor variables (probes). For each probe a numerical value is available, but it represents a semi-quantitative value. Besides, the response variable is categorical (3 groups) and the data are high-dimensional. The aim was to build a prediction model with only a small number of probes as predictor variables.

Regression, which is a standard method to use for prediction when the data are numerical, cannot be applied to these data, because the number of predictors (p) is larger than the number of subjects (n) (p>n). A standard method that is used for analyzing high-dimensional data is the Lasso [1]. The Lasso uses continuous data as input, which is not preferable in this case, because it represents semi-quantitative data. Therefore other methods that could be used to analyze these data were investigated.

A standard approach that could be used is Random Forests, which was developed by Leo Breiman [2]. Random Forests is a widely accepted ensemble method based on building many decision trees. It is able to deal with high-dimensional data, the categorical outcome, and the semi-quantitative data. However, Random Forests does not result in a single tree, which makes it often hard to interpret the results, and to perform variable selection automatically. Therefore, another method called Categorical Regression (CATREG) was investigated as well [3].

CATREG was developed as a method for regression with categorical variables using optimal scaling. By use of optimal scaling numerical values are calculated to replace the categorical values, while at the same time regression coefficients are estimated. (CATREG can also be applied to continuous data, a continuous variable is than seen as a categorical variable with as many categories as distinct values.) CATREG has been implemented in SPSS since 1999 [4], and regularization options such as Ridge regression, the Lasso and the Elastic Net were added to CATREG in 2006 that allows CATREG to perform variable selection, and to improve prediction accuracy based on Ridge regression [3].

An advantage of CATREG over Random Forests is that within CATREG it is possible to determine a variable specific cut-off value that can be used for the dichotomization of the data [3].

Another advantage of CATREG is that the interpretation is easier, because variable selection is performed automatically whereafter a model is fitted and coefficients for the predictors are estimated, similar to linear regression.

Until now there are not many published examples of analyses that are based on regularized CATREG. To determine how CATREG performs in prediction, it will be compared to Random Forests, one of the golden standards in statistical learning.

2 Methods

2.1 Random Forests

Decision trees are fast and easy to understand models that are invariant under monotonic transformations of the predictors, and which can be used to perform variable selection [5]. Decision trees often have a low bias and a high variance. This means that a decision tree fits the data often very well (low bias), but a small change in the data can give a totally different tree due to the hierarchical structure of tree, which results in a high variance, and unsuccessful prediction.

Bootstrap averaging (bagging) was developed to reduce this high variance [5]. With bagging a number of bootstrap samples are drawn, which is done by sampling from the data with replacement. All these bootstrap samples are then used to fit a model. The results of all these models are averaged and this often reduces the variance (and only slightly increases the bias).

Random Forests is based on bagging, it starts by taking a bootstrap sample of the data, the data not in the bootstrap sample are called the out-of-bag (oob) data [5]. For each node in the tree a random selection of a few predictor variables is taken and used to calculate the best split at that node. This is continued until the tree is fully grown, whereafter the grown tree is tested on the oob data. This procedure is then repeated many times until the model converges. Just as in bagging the results of all the trees are averaged. The variance of this average is:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2, \tag{1}$$

with ρ the correlation between the trees, σ^2 the variance and B the number of bootstrap samples. When the number of bootstrap samples increases, the second term in the equation diminishes. The variance of the average is then only dependent on the correlation of the bagged trees. Random Forests improves on bagging by reducing the correlation between the trees without increasing the variance too much. This is done by the random selection of the input variable at each node.

The prediction for classification with Random Forests is calculated based on majority voting [2]. Each grown tree results in a class prediction for each input sample. When all trees are grown the majority class of each input sample is determined and this is used as predicted outcome.

For all variables two different importance measures can be calculated, viz. the mean decrease in accuracy and the mean decrease in Gini index [6, 7]. The mean decrease in accuracy is the normalized averaged difference

over all trees between the oob prediction error and the oob prediction error after permuting each variable. The Gini index is a measure for how often a randomly chosen sample would be classified wrongly when the labeling was done randomly according to the distribution of the labels in the dataset. The mean decrease in the value of the Gini index is the total decrease in node impurities from splitting on that variable averaged over all trees, where the node impurities are measured by the Gini index.

2.2 CATREG

CATREG is a method that can be used to perform regression with categorical variables without relying on the assumption that there is a linear relationship between the predictors and the response [3, 8]. It applies optimal scaling to quantify the categorical variables and the response variable. The quantifications (transformations) are optimal in the sense that the multiple R squared (R^2) between the transformed response and a linear combination of transformed predictors is optimized. CATREG uses alternating least squares to estimate both the regression model, and the quantifications. It alternates between the estimation of the transformation of the predictor variables and the response variables.

The CATREG model is defined as follows [8]:

$$\varphi_r(\mathbf{y}) = \sum_{j=1}^{J} \beta_j \varphi_j(\mathbf{x}_j) + \mathbf{e}, \tag{2}$$

and the following loss function is minimized:

$$L(\varphi_r; \varphi_1, ..., \varphi_j; \beta_1, ...\beta_j) = \left\| \varphi_r(\mathbf{y}) - \sum_{j=1}^{J} \beta_j \varphi_j(\mathbf{x}_j) \right\|^2, \tag{3}$$

with φ_r the transformation of the response variable $\mathbf{y}$, J the number of predictor variables $\mathbf{x}_j$, β_j the regression coefficient, φ_j the transformation of the j^{th} predictor variable and $\mathbf{e}$ the residuals vector.

Before the algorithm starts with updating the quantifications of the response variable, $\mathbf{v}_r$, the quantifications of the predictor variables $\mathbf{v}_j$, and the regression coefficients have to be initialized [8]. This can be done randomly or numerical within CATREG. The random initialization uses standardized random values for the quantifications of the predictor and response variable. The initial regression coefficients are the zero order correlations of the quantified response variable with the quantified predictor variables. For the

numerical initialization the initial values are determined by an analysis with numerical scaling levels for all variables.

When the initialization is performed the algorithm starts with updating the quantifications of the response variable. There are different transformations possible for the response and predictor variables, and depending on which restriction is applied, the quantifications are updated differently. With numerical scaling of the response, the quantifications are the centered and normalized observed variables, when non-numerical scaling is applied the quantifications are updated by:

$$\tilde{\mathbf{v}}_r = \mathbf{D}_r^{-1}\mathbf{G}_r^{'}\sum_{j=1}^{J}\beta_j\mathbf{G}_j\mathbf{v}_j, \tag{4}$$

with $\mathbf{G}_r$ the indicator matrix for the response variable, $\mathbf{G}_j$ the indicator matrix for the j^{th} predictor, $\mathbf{D}_r = \mathbf{G}_r^{'}\mathbf{G}_r$, a diagonal matrix with the marginal frequencies of the outcome variable on the main diagonal, and $\tilde{\mathbf{v}}_r$ the unstandardized quantifications for the nominal scaling level. For the other quantification options, we can choose between ordinal, and monotonic or nonmonotonic spline scaling levels, the $\tilde{\mathbf{v}}_r$ is restricted according to the scaling level, which results in $\tilde{\mathbf{v}}_r^*$. For the ordinal and spline scaling levels $\tilde{\mathbf{v}}_r^*$ is standardized:

$$\mathbf{v}_r^+ = N^{\frac{1}{2}}\mathbf{v}_r^*(\mathbf{v}_r^{*'}\mathbf{D}_r\mathbf{v}_r^*)^{-\frac{1}{2}} \tag{5}$$

When the quantifications of the response vector has been updated, the algorithm continues to the second step, updating the quantifications of the predictor variable and estimating the regression coefficients. The response vector is held fixed and the quantifications of the predictor variables (with non-numerical scaling level), $\mathbf{v}_j$, and the regression coefficients are estimated by use of backfitting: the transformation and the regression coefficients are estimated separately for each variable [9, 10]. This is done by computing the N-vector of predicted values, named $\mathbf{z}$, by:

$$\mathbf{z} = \sum_{j=1}^{J}\beta_j\mathbf{G}_j\mathbf{v}_j, \tag{6}$$

whereafter the contribution of variable j is subtracted from the predicted values, which is needed to update the quantification of variable j: $\mathbf{z}_j = \mathbf{z} - \beta_j\mathbf{G}_j\mathbf{v}_j$. Than $\tilde{\mathbf{v}}_j$ can be updated:

$$\tilde{\mathbf{v}}_j = \mathrm{sign}(\beta_\mathrm{j})\mathbf{D}_\mathrm{j}^{-1}\mathbf{G}_\mathrm{j}^{'}(\mathbf{G}_\mathrm{r}\mathbf{v}_\mathrm{r}^{+} - \mathbf{z}_\mathrm{j}). \tag{7}$$

The regression coefficient β_j is updated by:

$$\beta_j^{+} = N^{-1}\tilde{\mathbf{v}}_j^{'}\mathbf{D}_j\mathbf{v}_j^{+}, \tag{8}$$

whereafter the updated contribution of variable j is added to the prediction $\mathbf{z}_j$:

$$\mathbf{z} = \mathbf{z}_j + \beta_j^+ \mathbf{G}_j \mathbf{v}_j^+. \tag{9}$$

When this procedure is repeated for all the other predictor variables, the value of the loss function can be computed by $\|\mathbf{G}_r \mathbf{v}_r^+ - \mathbf{z}\|^2$, and the result is compared to the loss of the previous round. When the difference between the two losses is smaller than a certain convergence criterion the algorithm stops.

2.3 Ridge and Lasso

When fitting a regression model, two things are important: the interpretability of the model and the prediction accuracy [1, 5]. We wish to interpret the model in terms of the regression coefficients, and the model should be able to predict independent future data satisfactorily. Often the ordinary least squares (OLS) estimates are not performing well on both of these criteria. This has been remedied by new methods, notably subset selection and Ridge regression.

Subset selection is a discrete process that can be used to get a more sparse and interpretable model. Ridge regression was developed to improve the prediction accuracy. This is done by adding a penalty term (L_2-norm) to the residual sum of squares, which regularizes the squared values of the regression coefficients. The loss function of Ridge regression is:

$$L^{Ridge}(\beta_1, ..., \beta_P) = \left\|\mathbf{y} - \sum_{j=1}^{P} \beta_j \mathbf{x}_j\right\|^2, \text{subject to} \sum_{\mathrm{j=1}}^{\mathrm{P}} \beta_{\mathrm{j}}^2 \leq \mathrm{t}_2, \tag{10}$$

where the value of the penalty parameter t_2 should be determined. Both methods, subset selection and Ridge regression, improve the model on one aspect, but not on both aspects at the same time. Therefore the Least Absolute Shrinkage and Selector Operator (Lasso) was developed. The Lasso improves both the interpretability of the model and the prediction accuracy. This is done by adding a L_1 norm to the residual sum of squares, which penalizes the absolute values of the regression coefficients. The loss function becomes:

$$L^{lasso}(\beta_1, ..., \beta_P) = \left\|\mathbf{y} - \sum_{j=1}^{P} \beta_j \mathbf{x}_j\right\|^2, \text{subject to} \sum_{\mathrm{j=1}}^{\mathrm{P}} |\beta_{\mathrm{j}}| \leq \mathrm{t}_1. \tag{11}$$

By penalizing the absolute values of the regression coefficients, some of the regression coefficients become exactly zero, which results in a more sparse

and more interpretable model with a better prediction accuracy than a OLS model for smaller values of t_1 [1, 5].

2.4 The Lasso within CATREG

As shown in the above section, the transformation of each variable in CATREG is separated from all other terms in the loss function. Therefore it is easy to apply regularization at the same time. The CATREG version of the Lasso regression for variable j can be written as [3]:

$$L^{lasso}(\beta_j) = \left\|\mathbf{y} - \sum_{l \neq j} \beta_l \mathbf{x}_l - \beta_j \mathbf{x}_j\right\|^2 + \lambda_1 w_j \beta_j + \lambda_1 \sum_{l \neq j} w_l \beta_l . \quad (12)$$

The w_l and w_j are +1 or -1 depending on the sign of the corresponding regression coefficient β_l and β_l. The regression coefficient of the j^{th} predictor can then simply be updated as:

$$\begin{aligned} \beta_j^{+lasso} &= \left(\beta_j^+ - \frac{\lambda_1}{2} w_j\right)_+ \\ &= \left(\beta_j^+ - \frac{\lambda_1}{2}\right)_+ if\ \beta_j^+ > 0 \\ &= \left(\beta_j^+ + \frac{\lambda_1}{2}\right)_+ if\ \beta_j^+ < 0. \end{aligned}$$

When all predictors are updated the transformation of the response variable can be obtained whereafter the R^2 is calculated and compared with the R^2 from the previous round [3].

2.5 Comparing the methods

Within both methods, CATREG and Random Forests variable selection will be applied, whereafter this variable selection is used to test how well the prediction is, when based on a model with only the selected variables. The dataset is randomly divided in a training and a test set (ratio 2:1). The variable selection is performed on the training set and the prediction is performed on the test set. The prediction error and the variable selection are used to compare both models and to investigate how well CATREG performs in prediction compared to Random Forests.

3 Predicting Bacterial Vaginosis

3.1 Description

The data used were collected at the "Leids Cytologisch en Pathologisch Laboratorium" (LCPL) in Leiden, which performs cytological and pathological diagnostics for patients of general practitioners. In addition, the LCPL performs diagnostics on cervical samples collected for population based screening of cervical cancer. Besides these screening smears, indication based vaginal smears are also analyzed. These indication smears are taken in the context of clinical complaints and follow-up of abnormal screening smears [11]. A major group of indication smears concern women with the clinical signs and symptoms of Bacterial Vaginosis (BV).

BV has a prevalence of about 8-23% in the female population and is associated with reproductive health morbidity, pelvic inflammatory disease, postoperative wound infections and pregnancy complications [11, 12].

The vagina is colonized by many different bacteria, which in healthy women protect the vaginal environment. A bacterial imbalance can cause complaints such as milky creamy discharge and an amine or fishy odor [11, 12]. This imbalance is microscopically reflected by a decrease in protective bacteria such as the Lactobacilli and an increase in potentially pathogenic bacteria such as Gardnerella, Atopobium and Dialister [13].

All women with BV show an imbalance in their bacterial flora, but until now it is poorly characterized [13]. Literature describes BV as a complex disease that is not caused by one specific etiologic agent but by a group of genital micro-organisms [11]. The diagnosis of BV is mainly based on microscopic judgment of fixed cytological samples, which might be difficult due to the presence of inflammatory cells in the sample [12].

At the LCPL the diagnosis of BV is based on the scoring of two morphotypes of bacteria, the long and in chains arranged protective Lactobacilli and the roundish coccoid pathogenic bacteria, named Gardnerella. This is different from the standard medical diagnosis that is based on the so-called Nugent score [14].

The LCPL diagnosis is set by estimating the mean presence over 10 microscopic fields on a single slide of both bacteria (Lactobacilli and Gardnerella) by an observer. A score from 0 to 3 is given for both morphotypes, representing, 0%, 1-5%, 6-20% or >20% presence, respectively. The combination of both scores determines whether a patients is given the diagnosis BV negative, intermediate or BV positive [13]. Table 1 shows which combinations are connected to the different diagnoses.

Table 1: Overview diagnoses BV

	L0(0%)	L1(1 − 5%)	L2(6 − 20%)	L3(> 20%)
G0(0%)	Neg	Neg	Neg	Neg
G1(1 − 5%)	Int	Int	Neg	Neg
G2(6 − 20%)	Pos	Int	Int	Neg
G3(> 20%)	Pos	Pos	Int	Neg

Neg: BV negative, Int: intermediate, Pos: BV positive
The G0 - G3 scores are representing the pathogenic bacteria and the L0 - L3 scores are representing the protective bacteria.

Different bacterial communities co-exists in humans. These are necessary for survival. The total of all microbes coexisting in humans are called the "microbiome" and the human site-specific bacterial communities are called "microbiota"[2]. The problem with investigating which bacteria are present at specific human regions is that many bacteria cannot be cultivated. This was discovered following the ability to amplify specific bacterial genes (based on ribosomal DNA). Using this technique one was able to identify which bacteria were present at specific human regions resulting in the first accurate description of the human microbiome.

The LCPL developed in collaboration with TNO (a Dutch innovative research company) a micro-array that is able to detect the presence and absence of specific bacteria, parasites, and fungi based on their specific ribosomal genes. With this array it is possible to investigate which bacteria are present in a vaginal sample of BV positive, intermediate and negative women and this could give more insight into the pathogenesis of BV, and could be used in the future to set the diagnosis of BV.

3.2 Analysis

3.2.1 Data

The cytological diagnoses are used as outcome labels and the results of the micro-array are used as input data. In this study vaginal samples of 111 Dutch women, diagnosed as BV positive, intermediate or BV negative, were analyzed. DNA was isolated, multiplied by Polymerase Chain Reaction (PCR) and converted to single stranded DNA, whereafter it was hybridized on the micro-array. The micro-array was covered with 456 unique short oligonucleotide sequences (probes) representing the ribosomal genes of the different bacteria, parasites and fungi.

The micro-array has been designed to detect matching bacteria with the specific probes. The micro-array is scanned by a laser and for each probe a signal over background ratio is calculated.

These ratios are semi-quantitative, a low value represents the absence of a specific bacteria and a high value represents the presence of a bacteria, but the amount does not give any extra information. TNO suggested to dichotomize the data and use a ratio of 5 signal over background as a cut-off value for all probes. We decided to use the continuous data as input and investigate whether the value of 5 is reasonable.

Of these 111 women, 60 were diagnosed as BV negative, 26 as intermediate, and 25 women as BV positive. The mean age of the BV negative women was 45, the intermediate women had a mean age of 34 and the mean age of the BV positive women was 30. The difference was statistically significant (p-value of 4.029e-08, Kruskal-Wallis Rank sum test, R package Stats). Because age is confounded with diagnosis it was not used as a predictor variable.

When inspecting the data, it turned out that there were some variables that showed an extremely skewed distribution. These variables will influence the stability of the estimated regression coefficients. Besides it is unlikely that these variables will have a large influence on the prediction because almost all values are in the same range, except for a few extreme values. Therefore 24 probes were removed that all showed 1-5 extreme observations only; in total 432 probes were left over to use for the analysis.

The data were randomly divided in a training set and a test set (ratio 2:1). The training set contained 74 women, of which 41 were diagnosed as BV negative, 20 as intermediate and 13 as BV positive. The test set contained 19 BV negative, 6 intermediate and 12 BV positive women.

3.3 Random Forests

3.3.1 Model building

A Random Forests model (R package RandomForest) was fit on the training data, where all probes (432) were used as predictors and the numerical values 0, 1, 2 for the response variable, representing BV negative, intermediate and BV positive, respectively. Default settings were used for the number of predictors randomly selected at each node, viz. the square root of the number of predictors, and the number of trees to grow (500). The convergence of the model was checked by inspecting the plot of the estimated prediction error and showed convergence after 200 fitted trees (plot not shown).

An importance plot was produced for the fitted model, which is shown in Figure 1. The importance plot shows the 20 probes with the highest

value for the mean decrease in accuracy. The mean decrease in accuracy is the averaged decrease in accuracy over all trees, between the estimated prediction error and the estimated prediction accuracy after permuting each predictor variable. A higher value indicates a more important variable [15].

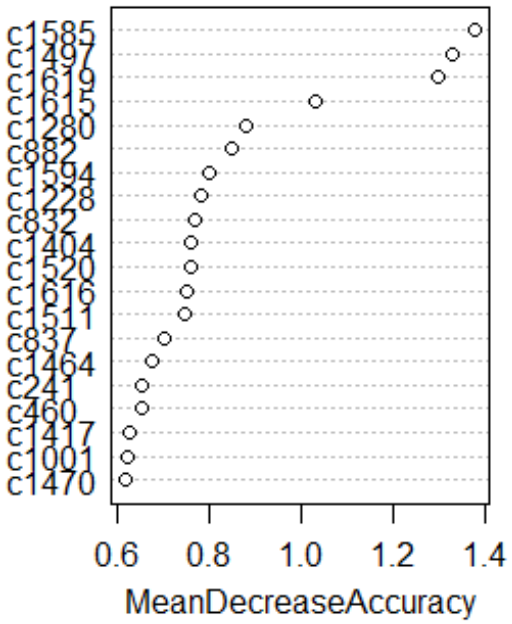


Figure 1: Importance plot. Top 20 most important variables based on the Random Forests model.

The variable selection was performed based on the importance plot, using the mean decrease in accuracy as importance measure, because this measure is a more reliable measure than the mean decrease in Gini index [6]. Based on the importance plot, three variables were selected, viz. c1585, c1619 and c1497. They were selected by inspecting the importance plot and searching for the largest break between variables. The variable names are code names for the different probes. The corresponding bacteria names are given in appendix A.

3.3.2 Testing

Based on the model with the three selected variables all women in the test set were classified. The results are shown in Table 2.
The outcome of 25 of the 37 women (68%) are predicted correctly according to their microscopic diagnosis. The outcome of 17 of the 19 BV negative women

Table 2: Test results Random Forests

	Outcome		
Predicted	BV negative	Intermediate	BV positive
BV negative	17	0	0
Intermediate	1	4	8
BV positive	1	2	4

is predicted correctly while only 8 outcomes of the 18 women diagnosed as intermediate or BV positive are predicted correctly.

3.4 CATREG

3.4.1 Model building

CATREG (SPSS version 17) was applied in combination with Lasso regularization to perform variable selection. The three categories of women (BV negative, intermediate and BV positive) must be coded in CATREG by numeric's. Although it can be assumed that there is an order in the severity of BV, this was not taken into account in the analysis. Thus the response variable was set at nominal, meaning that the order of the response categories did not needed to be preserved, and giving the algorithm the freedom to find an optimal ordering.

The predictors are continuous variables, and were discretized by recoding all values according to their rank within the predictor variable (discretization option ranking). This can be done without losing any generality, because the results of CATREG are invariant under monotonic transformations of the data. The optimal scaling level for the variables was set at ordinal spline with degree 2 and 2 interior knots; ordinal to preserve the order of the observations, giving monotonic transformations, and splines were used to obtain smooth transformations.

In CATREG there are two options to estimate the expected prediction error and the size of the penalty parameter, viz. .632 bootstrap and k-fold cross validation. The .632 bootstrap was chosen because it usually performs better than k-fold cross validation [3]. The maximum number of iterations for convergence was set to 500 and the convergence criterion was set at 10E-6 to avoid highly biased estimates of the expected prediction error due to non-converging solutions in the bootstrap.

Within CATREG the regularization options are Ridge regression, Lasso and Elastic Net. Ridge regression was rejected because it does not lead to variable selection. Elastic Net selects groups of highly correlated variables,

which was not desirable in this case. With the Lasso it is possible to select only a few variables, which can then be used for prediction, and this is exactly what was needed.

A series of Lasso penalties between 0 and 1 increasing by 0.1 was used to select a model. The Regularized Models Table with all penalties is shown in Table 3. Based on Table 3, model 4 with a Lasso penalty of 0.4 was selected:

Table 3: Overview Lasso penalties

Model number	Penalty	# predictors	APE	EPE	SE EPE
1	0.1	127	0.033	0.377	0.191
2	0.2	23	.044	.150	.041
3	0.3	13	.080	.179	.052
4	0.4	9	.113	.185	.030
5	0.5	6	.148	.267	.058
6	0.6	5	.188	.288	.045
7	0.7	4	.231	.334	.070
8	0.8	3	.275	.414	.075
9	0.9	3	.321	.424	.066
10	1.0	3	.373	.468	.054

APE: Apparent prediction error, EPE: expected prediction error, SE EPE: standard error of the expected prediction error.

the optimal model (model 2), has an EPE of .150 and a SE of 0.041; applying the one-standard error rule, the most parsimonious model with an EPE in the interval .109-.191 is selected (model 4). Model 4 contains 9 predictors, namely, c1585, c1619, c1615, c882, c1280, c1001, c678, c241 and c1497. The corresponding probe names are given in appendix A.

A model with only these probes as predictor variables and a Lasso penalty of 0.4 was fitted on the data again. The model summary and the coefficients of this model are shown in Table 4 and Table 5, respectively. The adjusted (R^2) is 0.91, the APE is 0.113 and the EPE is 0.143 (SE 0.022). The EPE of this fitted model was somewhat increased compared to the EPE of this model in Table 3. This is due to the fact that the bootstrap sampling is now only performed on the nine predictor variables instead of the 432.

Table 4: Model summary 1

Model	# predictors	APE	EPE	SE EPE	Adjusted R^2	Penalty
Lasso	9	0.113	0.143	0.022	0.91	0.4

APE: Apparent prediction error, EPE: expected prediction error, SE EPE: standard error of the expected prediction error.

Table 5: Probe summary 1

Predictor	Estimate	SE	Sig
c1497	-.103	.081	.205
c1585	-.262	.090	.001
c1619	-.288	.118	.005
c1615	.042	.061	.623
c882	.011	.033	.952
c1280	.028	.057	.869
c1001	.008	.033	.980
c687	.002	.028	.996
c241	.083	.053	.076

SE: standard error, Sig: Significance

3.4.2 Testing

The model with the 9 selected predictors was applied to the test set (which is possible in CATREG by specifying the test cases as supplementary; supplementary cases are not included in the analysis. The analysis is done on the non-supplementary (training) cases and the results (transformations and model parameters) are applied to the supplementary cases) [3]. The model summaries are shown in Table 6.

Table 6: Model summary test set

Model	#Predictors	EPE	SE	Penalty
Lasso	9	0.179	0.055	0.4

EPE: expected prediction error based on the test set, SE: standard error of the expected prediction error.

The EPE of the test set for the model with 9 predictors is 0.179. The transformation plot for the outcome of this model is shown in Figure 2. This transformation plot shows the values of the quantifications for the response variable (y-ax) and the original values of the response variable (x-as). The transformed outcome of the intermediate women (category 1) is slightly

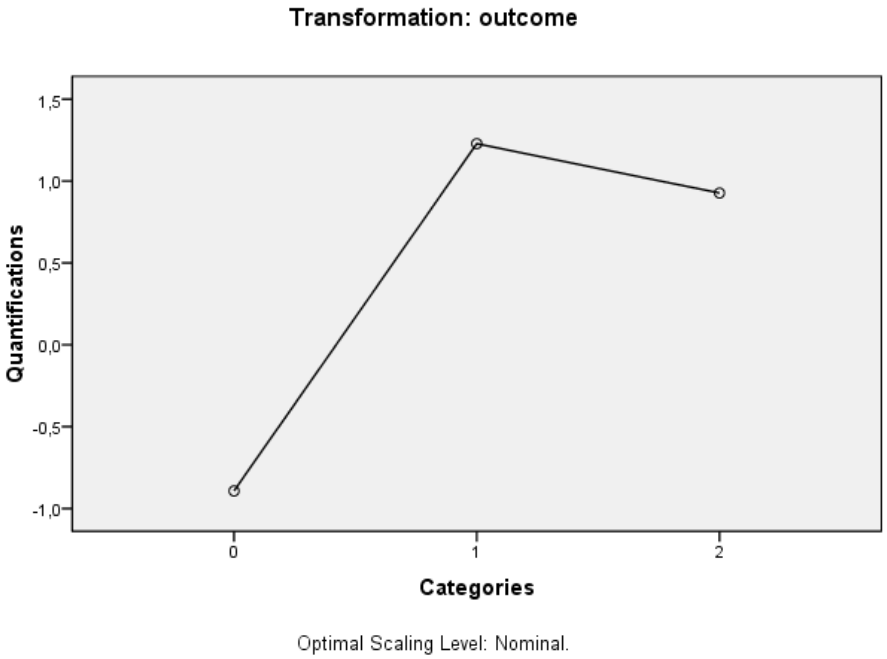


Figure 2: Transformation plot, response variable. The categories 0-2 represent, BV negative, intermediate and BV positive, respectively.

higher than the transformed outcome of the BV positive women. The difference between the BV positive and intermediate women is quite small. The difference between the BV negative women and the two other categories is quite large.

For all subjects (training and test set), the predicted values were calculated based on the selected model with 9 predictors. For five of the test cases it was not possible to calculate a predicted value because the test case had a value that was lower or higher than the lowest or highest value in the training set. For CATREG with ordinal splines it is possible to apply interpolation, but it is not possible to apply extrapolation, because this is likely to result in very unstable estimates. Therefore the quantified value of the lowest or highest value of the variable in the training set was imputed in the test set. When all the values in the test set have been replaced by optimal quantifications (obtained for the training set), we are able to use the predictors to obtain an predicted outcome.

The predicted values were used subsequent for the classification. Based on the predicted values of the training set, two different cut-off values, one

for the classification between BV negative and BV positive women, and one for the classification between intermediate and BV positive women, were calculated by use of an Reciever Operating Curve (ROC) (R package pROC). The predicted values of the BV positive women are in between the predicted values of the BV negative women and the intermediate women.

The range of the predicted values for the BV negative women in the training set is -0.88 – -0.30. The range for the intermediate women is 0.45 – 1.18 and the range for the BV positive women is 0.25 – 1.05. The predicted values of the BV negative women are lower than the predicted values of the intermediate and BV positive women. There is considerable overlap in the predicted values of the intermediate and BV positive women.

The first ROC for the BV negative and the BV positive women is shown in Figure 3. The second ROC for the intermediate and the BV positive women is shown in Figure 4. The plots show the chosen cut-off value and the corresponding specificity and sensitivity.

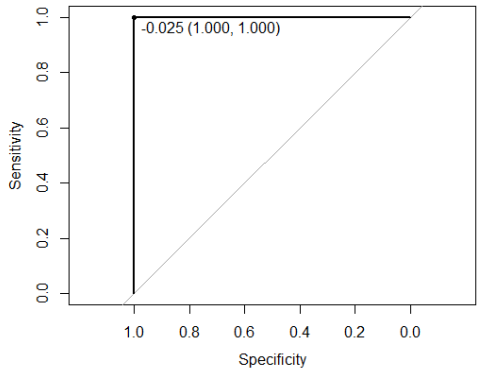


Figure 3: ROC curve 1

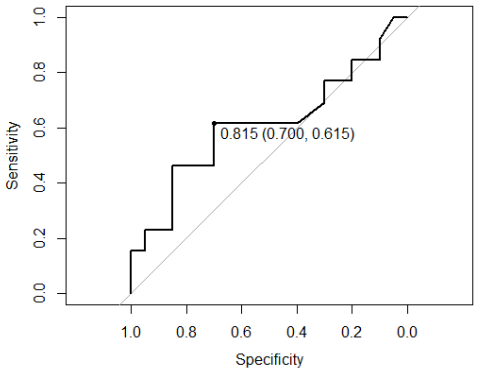


Figure 4: ROC curve 2

The cut-off value for the classification between BV positive and BV negative women is -0.025 and the AUC value of the ROC is 1. The cut-off for the second ROC curve between the intermediate women and the BV positive women is 0.815 and the corresponding AUC value is 0.61. Based on the two cut-off values of -0.025 and 0.815, the predicted values of the test set were classified. All predicted values lower or equal to -0.025 were classified as BV negative, all values between -0.025 and 0.815 were classified as BV positive and all values higher than 0.815 were classified as intermediate. The results of the classifications for the test set are shown in Table 7.

Table 7: Test results CATREG

	Outcome		
Predicted	BV negative	Intermediate	BV positive
BV negative	18	0	0
Intermediate	0	1	6
BV positive	1	5	6

The outcome of 25 of the 37 women (68%) is predicted correctly. Almost all BV negative women (18/19) are classified correctly. Almost none of the intermediate women are classified correctly, and 6 of the 12 BV positive women are not classified correctly according to the diagnosis based on the microscopic inspection.

3.5 Comparison results

The selected probes by both models, the CATREG model with Lasso regularization and the Random Forests model, are overlapping. The three selected probes by Random Forests are also selected by the CATREG model. Four of the other six selected probes used in the CATREG analysis, are present in the top twenty of the most important probes, according the Random Forests model. Both models, Random Forest and CATREG, classify 68% of the women correctly.

Both models are not performing well in predicting which women are intermediate or BV positive, while they both perform well in predicting the outcome of the BV negative women. These results show that CATREG is performing as well as the Random Forests model. Because it is not possible to distinguish well between the BV intermediate and BV positive women the results of these women are taken together and shown in Table 8.

With the Random Forests model 35 of the 37 women are classified correctly (95%) when the intermediate and BV positive women are taken together. With CATREG 36 of the 37 women are classified correctly (97%). Both models predict the BV positive women correctly, but misclassify one or two BV negative women.

Table 8: Test results

	Outcome			
	Random Forests		CATREG	
Predicted	Neg	Int + Pos	Neg	Int + Pos
Neg	17	0	18	0
Int + Pos	2	18	1	18

Neg: BV negative, Int: intermediate, Pos: BV positive

3.6 Dichotomization of the data

Because CATREG performed as well as the Random Forests model, the analysis based on the CATREG model was continued by inspecting the transformation plots (obtained for the training and the test data) of the nine selected variables. The transformation plots are shown in appendix B.

A transformation plot shows the quantifications versus the original values/categories. The transformation of a predictor reflects the relation of the predictor with the response, accounting for the influence of the other predictors. So, the transformation of a predictor gives an indication of the best cut-off value for dichotomization in predicting the response [3].

By use of the transformation plots the data was dichotomized. Because there was no clear break visible in the transformation plots, the intersection of the 0 quantification (y-axis) with the original data (ranking of the values) was used to dichotomize the data. Table 9 shows the used cut-off values for the dichotomization and the corresponding values of the signal over background ratios.

Table 9: cut-off values dichotomization

Probes	Cut-off value	Real values
c1497	49	1.96
c1585	52	3.09
c1619	49	1.71
c1615	57	1.34
c882	55	1.45
c1280	49	1.51
c1001	67	3.56
c687	67	1.38
c241	67	2.16

Three of the probes had a negative regression coefficient, viz. c1497, c1585, c1619, the other six probes had a positive regression coefficient. The

dichotomization of the data was performed as follows: for the three probes with a negative regression coefficient a rank lower or equal to the cut-off was coded as 1, when the rank was higher than the cut-off it was coded as 0.

For the six probes with a positive regression coefficient a rank higher than the cut-off was coded as 1, and a rank lower or equal to the cut-off was coded as 0.

Based on the dichotomization, the sum of the nine values was calculated for each women to investigate whether it is possible to distinguish between the different outcomes based on the dichotomized data. The results are shown in Table 10. The BV negative women all show a sum between 0 and 5. The intermediate and BV positive women show a value between 5 and 9. There is considerable overlap in the sum values between the intermediate and BV positive women, but the sum values of the BV positive women are slightly higher.

Table 10: Sum values dichotomized data

Sum	BV negative	BV intermediate	BV positive
0	24	0	0
1	13	0	0
2	15	0	0
3	6	0	0
4	1	0	0
5	1	1	0
6	0	5	3
7	0	10	7
8	0	5	9
9	0	5	6

4 Discussion and Conclusion

The CATREG model performed as well as the Random Forests model with respect to the classification of the women, both predicted 68% of the outcomes correct. Both models have difficulties with classifying the intermediate and BV positive women correctly, while the BV negative women are classified correctly. When the intermediate and BV positive women are taken together, the percentage of correctly classified women increases to 95% for the Random Forests model and to 97% for the CATREG model.

A model with only the intermediate and BV positive women was run as well, but based on these data it was not possible to select variables that are able to distinguish between intermediate and BV positive women. This could be due to the small number of samples, or because the measured probes are not representing probes that are able to distinguish between the intermediate and BV positive women.

The selected probes by both models, the CATREG model with Lasso regularization and the Random Forests model, are overlapping. The three selected probes by Random Forests are also selected by the CATREG model. Five of the other six selected probes used in the CATREG analysis, are present in the top twenty of most important probes, based on the Random Forests model. The other probe selected by the CATREG model is in the top 50 of most important variables based on the Random Forest model.

The overlap in selected probes between the two models suggest that these probes are indeed the probes that are important to make the distinction between the BV negative on the one hand and the intermediate and BV positive women on the other hand, and that CATREG selects the same probes as Random Forests (in this data set).

The dichotomization of the data resulted in the same results as the models based on the transformed values. It was not possible to distinguish well between the intermediate and BV positive women based on the sum of the dichotomized probes, but the distinction between the BV negative and the intermediate/BV positive women was very clear.

The signal to background ratios which correspond to the ranks used for the dichotomization are lower than the suggested value of 5 by TNO. Besides, the values differ for the different probes, which is not taken into account when a general cut-off level is used.

Lasso regularization is applied to the probes and they are optimally monotonically transformed within the CATREG model, while the Random Forests model is applied on the continuous data directly. This, however, does not influence the comparison between the probes selections because the Random Forests model is invariant to monotonic transformations as well [5].

The ROC curve used for the classification of the predicted values based on the CATREG model only gives a cut-off value, other values are possible as well, the consequence of a false positives and a false negative are now treated equally, but it is possible to put more weight on one of the two.

The used outcome option within the CATREG model was set at nominal; the order of the response value did not need to be preserved. This resulted in a slightly higher transformed response value for the intermediate women compared to the BV positive women. A model with ordinal outcome was also fitted on the data; this resulted in fourteen selected predictor variables, the nine also selected by the model with the nominal response option, and five extra probes, viz. c333, c249, c998, c1574, c815. These five extra probes were not in the top 20 of most important probes of the Random Forest model.

Overall the results suggest that the intermediate women are more similar to the BV positive women, but is does not confirm the idea that the intermediate women are in between the BV negative and BV positive women. Further research should be performed to investigate what the differences between the intermediate women and the BV positive women are, and whether it is possible to better distinguish between them.

Not included in this study is the effect of age, due to significant differences between the groups; when performing further investigations one also has to think about the effect of age, because it is possible that the bacterial composition changes with age.

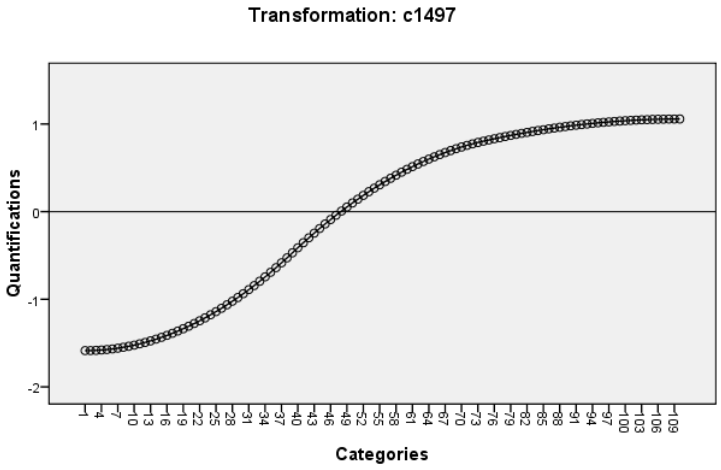
Overall the research showed that CATREG performs as well as Random Forests in the prediction, which is a very favorable result.

A Bacteria names

Table 11: Bacteria names per probe

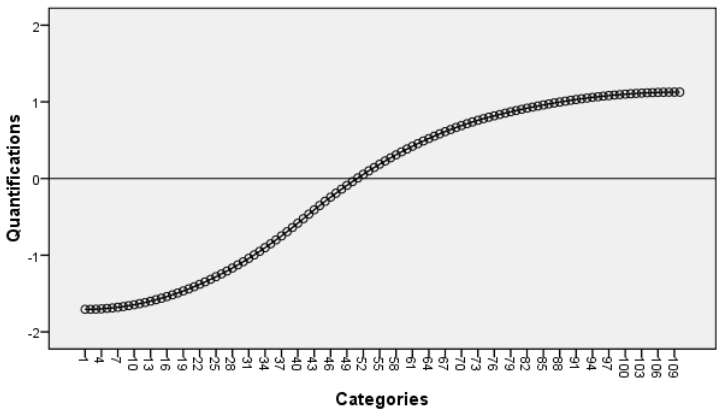
Probe	Family	Species
C1585	Comamonadaceae	Variovorax group
C1619	Comamonadaceae	Variovorax group
C1615	Enterobacteriaceae	Pantoea sp.
C882	Corynebacteriaceae	Corynebacterium urealyticum
C1280	Enterobacteriaceae	Pantoea sp.
C1001	Pasteurellaceae	Heamophilus influenzae
C687	Porphyromonadaceae	Phorphyromonas sp
C241	Prevotellaceae	Prevotella group 7
C1497	Comamonadaceae	Acidocorax group
C333	Enterobacteriaceae	Unclassified Enterobacteriaceae
C249	Pseudomonadaceae	Pseudomonas group 1
C998	Hyphomicrobiaceae	Hyphomicrobium facilis
C1574	Lachnospiraceae	Moryella uncultured
C815	Moraxellaceae	Acinetobacter baumanii

B Transformation plots



Optimal Scaling Level: Spline Ordinal (degree 2, interior knots 2).

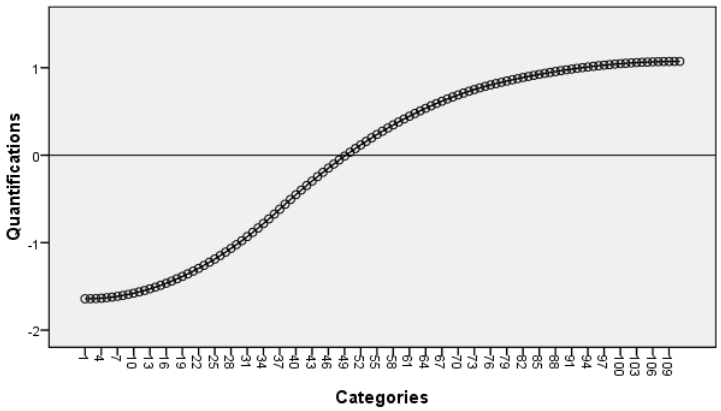
Beta: -,121.



Optimal Scaling Level: Spline Ordinal (degree 2, interior knots 2).

Beta: -,244.

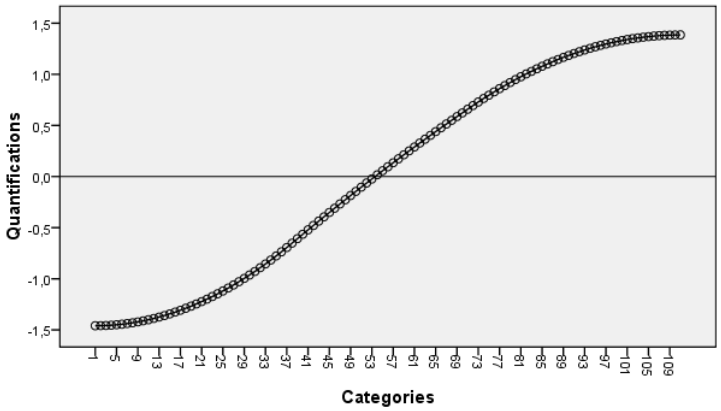
Transformation: c1619



Optimal Scaling Level: Spline Ordinal (degree 2, interior knots 2).

Beta: -,290.

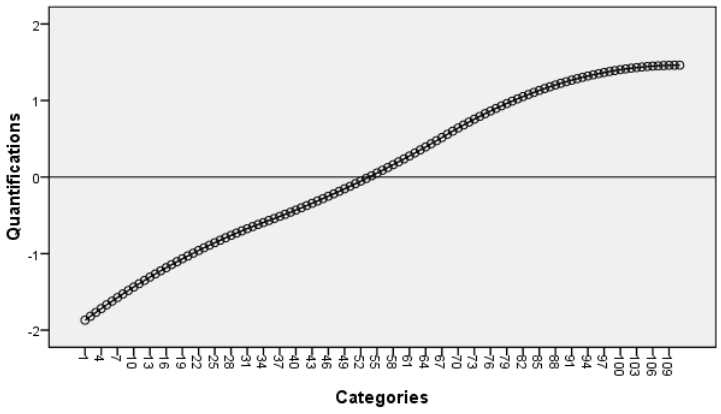
Transformation: c1615



Optimal Scaling Level: Spline Ordinal (degree 2, interior knots 2).

Beta: ,049.

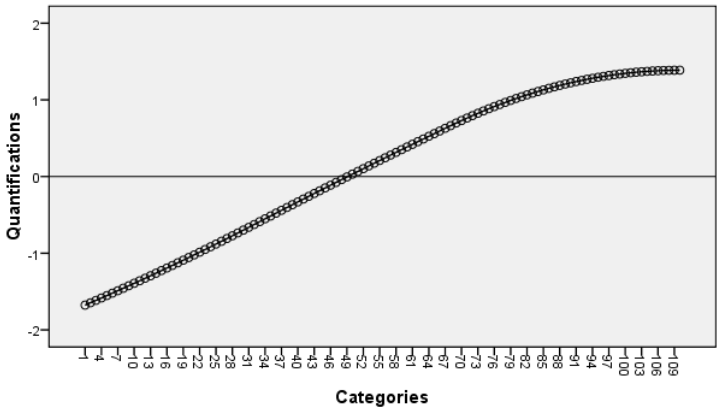
Transformation: c882



Optimal Scaling Level: Spline Ordinal (degree 2, interior knots 2).

Beta: ,045.

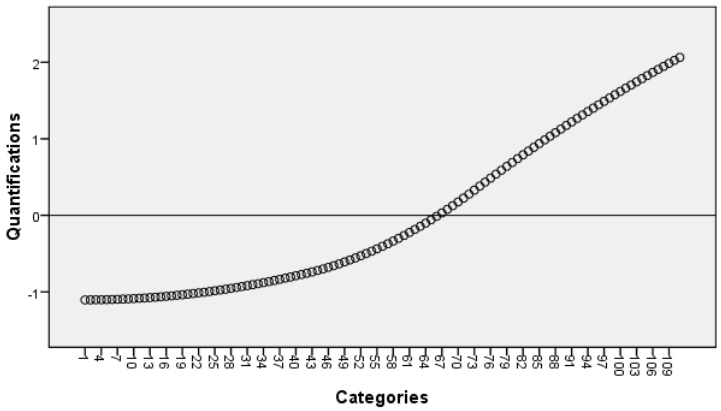
Transformation: c1280



Optimal Scaling Level: Spline Ordinal (degree 2, interior knots 2).

Beta: ,043.

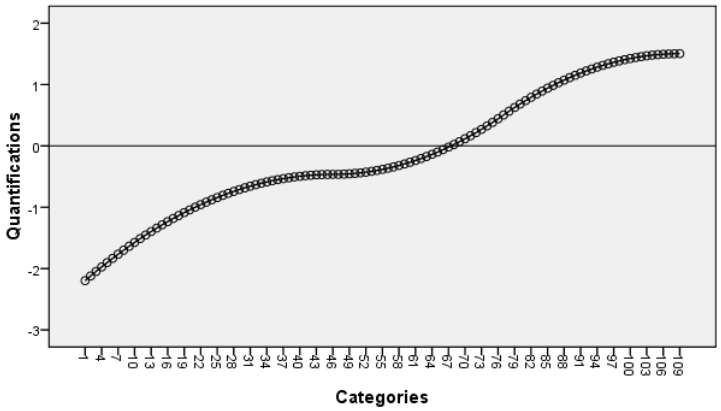
Transformation: c1001



Optimal Scaling Level: Spline Ordinal (degree 2, interior knots 2).

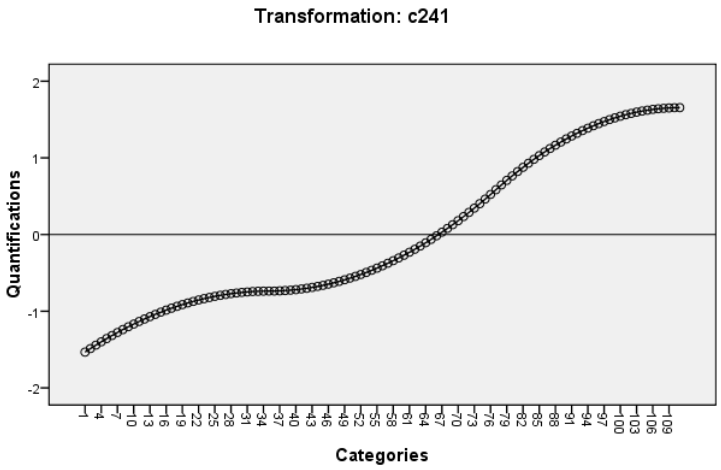
Beta: ,061.

Transformation: c687



Optimal Scaling Level: Spline Ordinal (degree 2, interior knots 2).

Beta: ,047.



Optimal Scaling Level: Spline Ordinal (degree 2, interior knots 2).

Beta: ,105.

References

[1] R.J. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

[2] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[3] A. J. van der Kooij. *Prediction accuracy and stability of regression with optimal scaling transformations.* PhD thesis, Leiden University, 2007.

[4] J.J. Meulman, W.J. Heiser, and SPSS Inc. *SPSS Categories 10.0.* Chicago: SPSS Inc., 1999.

[5] T. Hastie, R.J. Tibshirani, and J.H. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer series in Statistics, 2001.

[6] A. Liaw and M. Wiener. *Classification and Regression by randomForest.* R News, 2002.

[7] Leo Breiman. *Manual on setting up, using, and understanding Random Forests V4.0.*, 2003.

[8] A. J. van der Kooij, J. J. Meulman, and W.J. Heiser. Local minima in regression with optimal scaling transformations. *Computational Statistics and Data Analysis*, 50:446–462, 2006.

[9] J.H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823, 1981.

[10] A. Buja, T.J. Hastie, and R.J. Tibshirani. Linear smoothers and additive models (with discussion). *Annals of Statistics*, 17:453–510, 1989.

[11] J.M. Marrazzo. Interpreting the epidemiology and natural history of bacterial vaginosis: are we still confused? *Anaerobe*, 17:186–190, 2011.

[12] D.H. Martin. The microbiota of the vagina and its influence on women‘s health and disease. *American Journal of Medical Science*, 343:2–9, 2012.

[13] J.A. Dols, P.W. Smit, R. Kort, G. Reid, F. H. J. Schuren, H. Tempelman, Tj. R. Bontekoe, H. Korporaal, and M. E. Boon. Microarray-based identification of clinically relevant vaginal bacteria in relation to bacterial vaginosis. *American Journal of Obstetrics and Gynecology*, 204:1–7, 2011.

[14] R.P. Nugent, M.A. Krohn, and S.L. Hillier. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *Journal of Clinical Microbiology*, 29:297–301, 1991.

[15] D. R. Cutler, T. C. Edwards, J.R. Karen, H. Beard, A. Cutler, K.T. Hess, J. Gibson, and J.J. Lawler. Random forestsfor classification in ecology. *Ecology*, 88:2783–2792, 2007.