Equalizing the Cost of Health Insurance

Casper Beentjes Alessandro Di Bucchianico
Christian Hamster Ajinkya Kadu Irene Man
Keith Myerscough Marta Regis Omar Richardson

Abstract

The Dutch government compensates health insurance companies when insuring individuals who are estimated to have high health care costs. This is necessary to avoid insurers not offering services to certain groups or not providing them with a high quality of service. It is, however, unknown to what extent the differences in health care expenses by different groups of people are truly due to a poorer or better health status. We explore several statistical approaches that facilitate explaining the cause of these differences.

Keywords: health insurance, risk equalisation, model selection, predictive model, explanatory model, model selection, lasso, elastic net, ridge regression, clustering.

1 Introduction

Health care costs in the Netherlands are paid for by private insurance companies, who receive their funds from two different sources. In the Netherlands each adult chooses an insurance company and pays a fixed premium per month. Insurers are free to set their premium, but it has to be the same for all insured adults. The first source of income for insurance companies is this monthly premium paid by all their customers.

The second source is a subsidy from the government. This subsidy is different for different insured individuals, based on a number of indicators that estimate the general health of the insured. The goal of this differentiation is to equalize the risk carried by insurers when insuring different people. Without such equalization, it would be profitable for insurers to target certain groups they estimate will generate larger health care costs and an incentive for insurers to offer good care and services to those in need would be lacking. This report focuses on the second source of funding, the risk equalization fund.

Determining the correct amount of funding for each insured individual is a challenging task and involves political considerations. This task is very important since health care costs in the Netherlands are increasing significantly due to an ageing population, which is a threat to the affordability of the national health care system. The particular problem we intend to address is that the current approach uses past health

care expenses as the basis for the estimation of required health care costs. These past expenses, however, are not necessarily a good indicator of the truly required health care costs. Several factors may lead to inflated expenses, such as, but not limited to: a propensity to 'consume' health care if it is readily available, the deliberate exaggeration of diagnoses by care givers to increase turnover and profit and inefficiencies in the execution of certain treatments. On the other hand, the real expenses do not see where care was required, but not consumed due to financial incentives, such as the legally imposed deductible of several hundred euros or the loss of income for self-employed people. These two deficiencies both have grave consequences for the functioning of the health care market. The first error, overestimating the truly required costs, removes the incentive for health insurers to put pressure on care providers to make their business more cost efficient. Furthermore, as there is a fixed total budget, an overestimated budget for one group directly harms another. The cases where the insured persons would benefit from more care but are for some reason unable to obtain this are now largely ignored by the system. This could be detrimental to their long term health, and is certainly an ethically questionable situation.

2 The current model

In this section we will outline the current procedure used by the government to determine the funding for health insurance companies. We will particularly focus on how the funding for risk equalization is computed. The risk equalization model is calculated for the total funding and then adjusted for a set premium by VWS. After the real premium collection by insurers, an insurance company either gets money from the risk equalization fund or contributes to the risk equalization fund based on the outcome of the risk equalization model. The basis of the risk equalization is a linear regression model that aims to predict the health care costs for each individual based on a number of personal variables that are deemed a good indicator of their general 'health status'. We will detail which variables are used — and to some extent why — in Section 2.1. Due to the time required for processing all health care providers' accounts of realized costs, there is a three year lag in the cost prediction. This implies the risk equalization funds are determined for 2017 using a regression model based on costs from 2014 and the characteristics (FKG/DKG) from 2013.

2.1 Model parameters

The variables used for risk equalization are intended to be variables that indicate how healthy a person is likely to be. Originally, this included only the age and gender of individuals, but the number of variables included in the model has vastly increased since then.

We have summarized the used variables in Table 1 and noted a few remarkable properties of the data below. The categories `s_fkg`, `s_dkg` and `s_hkg` all indicate specific use of products and are therefore strongly linked to specific health issues.

Table 1: A list of model parameters used in the government's risk equalization scheme

Variable prefix	Explanation
`normbedrag_somatisch`	Gross compensation of health insurer for somatic costs in the postcode based on risk equalization model
`s_iedereen`	Total number of people in the postcode, given in 'insured years' to account for people who are insured for the full year
`s_kost`	Total costs for various types of somatic care
`s_totale_kosten`	Grand total of somatic costs
`s_lgnw`	Age and gender categories
`s_fkg`	Pharmaceutical cost groups
`s_ape`	Postcode 'region'
`s_dkg`	Diagnosis cluster groups
`s_mhk`	History of medical expenses (top percentiles over past two or three years)
`s_hkg`	Medical devices-based cost group
`s_avi`	Source of income
`s_ses`	Social economic status
`s_FGG`	Physiotherapy use
`s_VGG`	Nursing and caregiving (at home) costs
`s_GGG`	Geriatric rehabilitation care
`s_gsm`	Comorbidity

The (ten) postcode 'regions' indicated by the `s_ape` variables will be discussed in greater detail in Section 2.2. Some of the categories (e.g. `s_avi` and `s_ses` or `s_fkg` and `sdkg`) are strongly correlated, resulting in a strong multicollinearity, which we will discuss in Section 4.3. It should also be noted that the category `s_avi` contains a wealth of different types of income, such as benefit schemes and regular employment, but also contains elements for students and the self-employed.

2.2 Linear regression procedure

The regression procedure used by the government consists of three steps. The first step is a linear regression that fits the grand total cost (variable `s_totale_kosten`) based on all the predictive variables (variables `s_lngw`–`s_gsm`) except for the postcode regions (`s_ape`). This first regression is performed using two constraints: (i) those coefficients associated to age and gender must result in a total that matches the true total when ignoring the other variables, and (ii) the other groups of coefficients must all result in a zero sum. These constraints result in an easier interpretation of regression parameters and facilitate the comparison of the parameters obtained from different models or different years, but in our view do not affect the result of

the regression. An exeception are pharmaceutical costs, since each individual can have more than one `s_fkg`, while for example the `s_dkg` of each individual can be categorized in one group only.

After the first step, the model is further refined by looking at 'regional' variations. The residuals from the first step are aggregated to a postcode level, and the postcodes are then clustered into ten 'regions' based on their aggregated residuals. This clustering is performed by considering ten deciles of these residuals, and thus the 'regions' do not necessarily have any geographical cohesion. A final regression is then performed using the postcode region (`s_ape`) for each individual, again constrained to a zero-sum correction.

3 Our goal

We would like to immediately point out an important distinction between the *prediction* and the *explanation* of health care costs. The current model is intended for prediction to equalize the risk for healthcare insurers. The question(s) posed relate to both explanation and prediction, to understand how the various parameters in play affect the realized healthcare costs and to what extent these are down to the actual health status of individuals. It is known in statistics that models that perform well for prediction, may perform poorly for explanation. We combine these questions in a single research question:

> What are appropriate ways to find and explain geographical differences in healthcare costs?

The answer to this question will be given by a number of different data analysis tools. We exemplify – and to some extent justify – these methods by discussing the results of their application to the aggregated data set that has been made available to us.

4 Results

The proposed models can be divided into three categories. First, we study ordinary linear regression models similar to the model used by the government, but we focus on the selection of the most significant variables. Secondly, we look at more sophisticated linear regression models in order to obtain models that are good for either explanation or prediction. Finally, we perform clustering of postcodes (or other aggregation levels) to investigate analogies and differences. This approach is significantly different from the other two and is aimed specifically at explanation rather than prediction.

We would like to stress that we are aiming solely at providing *tools* that may be used in assessing the proper choice of model and variables. Any such choice impacts the funding of health insurers and consequently their behaviour on both the healthcare market (buying healthcare from providers) and the consumer market (selling insurance to individuals). The choices made should be justifiable to the public or at least to their elected representatives. It is our contention that *after* using tools that might

seem opaque to assess the behaviour of certain models, it is possible to obtain better models that balance transparency on the one side with predictor accuracy on the other. Note, however, that an accurate predictor of realized costs might in fact be predicting something different to the necessary costs of health care.

Beside model fitting, finding which elements in a data set are outliers with respect to a given regression provides useful information on the quality of the regression. If outliers share traits that are not captured by the regression model, it may be useful to include a quantifier for these traits in the model. The outliers will also clearly demonstrate the regional differences, potentially providing an even stronger motivation for our current exercise. We discuss the results of outlier detection (aggregated at the level municipalities for privacy reasons) in Section 4.2.

We also perform recursive variable selection to get an understanding of which variables in the model are most relevant. Subsequently ranking the different models based on how well they model the variation in the data relative to the required number of variables — quantified by Mallows' C_p, see Mallows (1973) — provides valuable insight into the optimal number of variables to include. Particularly, this will indicate if the current model is over- or underfit. This is important because overfitting may lead to spurious explanations of differences. Section 4.3 contains the results of this part of the analysis.

To avoid issues arising due to the collinearity between many of the predictor variables (in particular, the danger of important variables being erroneously declared non-significant since significance is divided over several similar predictor variables), we consider adding regularization to the regression. Such regularization can facilitate the interpretation of different resulting parameters for the regression by removing some ambiguity from the system. It does, however, introduce some new (meta)parameters that need to be computed a priori. We discuss some preliminary results from this approach in Section 4.4.

The final approach we elaborate upon does not involve a regression technique, but instead attempts to find clusters of similar postcodes. Within these clusters, it may be easier to identify individuals, postcodes or groups at a different level of aggregation that are remarkably different from others in the cluster. A clustering based on data at the postcode level, the finest at our disposal, is presented in Section 4.5.

Before detailing the results of these approaches, we briefly outline some of the work that was done in preparing the data.

4.1 Data and preparation

Currently the government collects a large number of 'health status' variables on an individual person basis as input for their risk equalization calculations. These indicator variables are, amongst others, information on location of residence, age, gender, social economic status, source of income, healthcare costs in the previous three years and the morbidity of the individual. The morbidity is split in somatic morbidity and mental morbidity. For somatic morbidity the government includes 30 (classes of) diseases each with its own list of specific medicine use and medical treatments (Zorginstituut

Nederland, 2017). The use of these determines whether an individual is registered for that specific disease in the government database. For mental morbidities a similar approach is used, although with fewer diseases. The resulting dataset then contains a total of 225 variables per individual of which 26 are mental health care costs specific and 17 are used solely in the model for regional variation at the postcode level.

The original dataset held by the government is too confidential to work with as it contains information on the health of individual citizens. Therefore, we only had access to a data set aggregated at the four digit postcode level [1]. This results in 3838 postcodes with the combined 225 health status variables of the people living in those postcodes together with an extra variable depicting the total number of people in that postcode.

Due to this aggregation a few peculiarities creep into the dataset. Firstly the dataset contains several postcodes that consist solely or partially of PO boxes[2]. These do in fact not correspond to a physical location in the Netherlands where people are registered to live. Multiple scenarios exist why people can be registered under a PO box, such as when someone does not have a fixed address or lives abroad. In that case the health insurers will often register the costs of the insured person on the postcode of the insurer, which can be a PO box. For the purpose of explaining geographical differences in the Netherlands we exclude these particular set of postcodes as any geographical information on the people in these groups is lacking[3]. However, when making predictions for the health care costs for the next year the individuals in these postcodes have to be included, since they do after all contribute to the total costs and need to be included in the risk equalization calculations. This reiterates our earlier point that there should be a difference between the explanation and prediction approaches.

Aside from the PO box issue we now have the issue that postcodes can widely vary in the number of residents registered. As a result we observe a vast range of different scales of many of the indicator variables. We therefore normalise all the variables to the number of registered residents. The indicator variables then represent the average values for a registered insured person at the various postcodes.

[1] Dutch postcodes follow a four digits plus two letter format, e.g. 1234 AB. The aggregation puts together all the residents of 1234 AB and 1234 CD into the same category 1234.

[2] In major cities like Amsterdam and Rotterdam such postcodes are the ones ending at 00 or 01, but this varies from city to city. E.g., PO boxes in Nijmegen are postcodes ending on 00, 01, 03, 04 and 31 (see e.g. `http://postcodebijadres.nl/postbus+Nijmegen`.

[3] There remain postcodes that are partially PO boxes and partially residential addresses and we make the decision to omit these from the data used for the explanation as well whenever we can locate these postcodes.

4.2 Outlier identification

Summary

Goal Identify regions with exceptional costs or characteristics. These identified regions can then be investigated more thoroughly to determine if the risk adjustment model is appropriate and sufficient for these more extraordinary groups of insured. In addition exclusion of extreme values can improve the estimation of the expected costs of a group of insured.
Method Studentized residual analysis after linear regression on a regional basis for the costs and age/gender.
Main result There are a few regions with exceptional costs and characteristics. For example, Urk, Lelystad, Pekela, Koggenland, Weesp and Oegstgeest have an extraodanary age/gender-profile. Vlist, Onderbanken, Pekela, Vlieland, Menseradiel, Son en Breugel and Oud-Beijerland have exceptional costs. Pekela seems extreme with low age and high costs.
Recommendation Further research of characteristics (next to age or gender) of inhabitants in the extraordinary regions is recommended. This can lead to new characteristics that can be included in the risk equalization model. Compare the current risk equalization model with a model where extreme values are excluded to get a feeling how strongly the average results are biased by the outliers.

As mentioned in Section 2, the current model used has three steps, of which the first is a straightforward linear regression and the second and third steps aim to correct for regional variation. In the interest of simplicity, we will only consider the first step of the procedure. Due to restrictions on what we are allowed to publish, the data is first aggregated to the level of municipalities before studying outliers. We perform two linear regressions using a different subset of the available variables. The first uses only the age and gender distribution of each municipality and the second uses all predictive variables that are used in the government model.

Figure 1 displays the regression fit and the studentized residuals of that fit for each municipality in the data set. The left panel shows the results using only age and gender in the model, the right uses all predictive variables. The studentized residuals represent a rescaling of the residuals (i.e, the differences between the observations and the fitted values from the model) such that it is comparable to a standard normal distribution under the assumption that the error of the linear regression is truly Gaussian. In this way the residuals are scale free, i.e. their values are independent of the unit used for the response variable, so that it is possible to have a universal threshold to detect outliers. Red lines in Figure 1 indicate a threshold for outliers chosen at 2.5 times the standard deviation from the zero mean. The factor 2.5 is a rule-of-thumb to decide on suspect observations, based on the approximate standard normal distribution of the scaled residuals.

When using only age and gender in the model, there are three outliers either side of the 2.5 standard deviation threshold. This is not surprising in itself, but the very low fit for the municipality of Urk is exemplary of a broader trend that 'cheaper' municipalities are underestimated. It should be noted that Urk is a fairly unique location; it is a former island that still retains a somewhat isolated character. When using all predictive variables, there is a marked skew in the outliers with many more

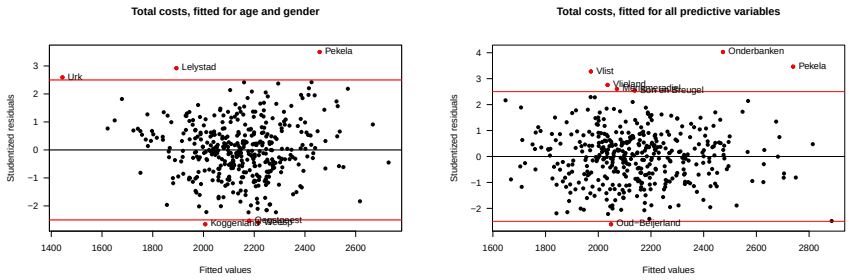

Figure 1: The studentized residuals after linear regression on a municipal level, based on only age and gender (left) or on all predictive variables (right). On the x-axis are the predicted values of the models. These studentized residuals put deviations from predictions on a universal scale. The usual threshold is 2.5. The municipalities that are highlighted in red have total costs that are not predicted well by the respective models

municipalities having a severe, positive residual. As a final note, we wish to point out the municipality of Pekela retains a strong, positive residual when using all variables.

4.3 Variable selection

Summary

Goal In the current risk equalization model there is a risk of over-fitting due to the large number of included variables. There are also some problems with multicollinearity that make results difficult to interpret. In this part of the study we identify the variables with the highest impact on the results and which variables can be left out with little impact on the result.
Method We use a stepwise forward model selection and a stepwise backward model selection procedure to find the variables with highest and lowest impact.
Main result The forward and backward selection procedure both lead to the conclusion that variables need to be excluded in order to avoid overfitting. The two methods have partially overlapping results concerning variables that can be excluded. However, it is difficult to determine which variables should be left out due to multicollinearity.
Recommendation Use this method to determine which variables potentially can be deleted. For the variables with multicollinearity problems, determine politically which variables should be left out. Estimate the regression model without these variables to determine the impact on the regular descriptives of the model. Use a measure for multicollinearity in addition to the current descriptives to judge the performance and validity of the model.

There are several different ways to study which variables are important so that they should be included in a regression model. Note that this is not the same as ranking the variables in a model according to their importance (see Grömping (2007) for an excellent discussion of relative importance in regression analysis). In this section, we

focus on forward and backward regression, which are elementary, heuristic ways for iterative model selection.

Stepwise *forward* model selection starts with a (trivial) linear regression model with no variables and then at each the variable that results in the lowest R^2 error term is added to the regression. As such, an increasingly complex regression is constructed. *Backward* model selection starts using *all* model variables and then iteratively removes those variables that have the smallest impact on the error. Both methods result in a hierarchy of models and a list of variables for each model. The models are then ranked by a score that balances the complexity of the model with the accuracy of the fit. This is intended to counter the overfitting that would occur if only the R^2 error is used as a norm – in that case the more variables the better, which may lead to overfitting. By studying which variables are used most, or at least used by the best models, provides useful information on which (categories of) variables have the most predictive power.

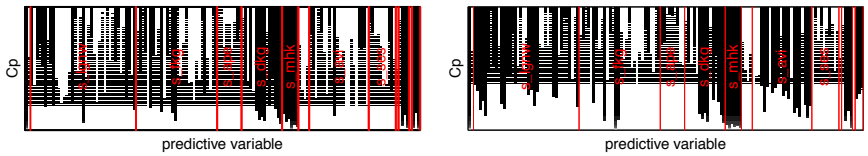

Figure 2: Ranking of models based on Mallows' C_p. Each row of either chart represents a linear regression using different variables, indicated by the dark cells in that row. The left-hand panel uses forward selection, the right hand panel uses backward selection. Note that the vertical axes are ranked, but correspond to slightly different Mallows' C_p values (the lower the better), both ranging from 136 up to roughly 4900.

Figure 2 demonstrates the results of forward (left panel) and backward (right) model selection. The models are ordered vertically by Mallows' C_p (see Mallows (1973) and Gilmour (1996), an alternative to R^2 that penalizes for having too many variables), a lower score indicates a better model. Note that there is no linear scale on the axis. It is well-known in the statistical literature that R^2 is not a model selection criterion (see e.g. Kvålseth (1985)). Dark pixels indicate variables included in the model, the darkness of the colour scales linearly with Mallows' C_p. The individual variables have not been labelled on the abscissa, since this would be too dense to read (but they can be easily read off a tabular output). Instead the different categories have been indicated.

Variables that are included in the top rows in the diagrams provide the most useful information in predicting health care costs. We observe in both forward and backward selection that the best scoring models are those with roughly half the number of

variables included. Models that include more variables rank low, indicating that using all these variables is overfitting the data. This conclusion is particularly apparent from the forward selection, but also holds in the backward case.

The difference in the variables selected by the forward and backward procedures is minimal. We therefore first focus on aspects that are visible in both, a few differences will be pointed out later. The most striking feature is the ubiquitous inclusion of the history of health care expenditure (`s_mhk`) the list of variables. These categories reflect top quantile health care use in previous years, and probably is particularly useful in predicting the high costs for chronic patients. Similarly, a history of high costs for nursing and caregiving is also a strong indicator of realized health costs. Comorbidity (`s_gsm`, far right) is also included in all models with a high rank. However, this is likely to cause multicollinearity in models that also include `s_fkg`, `s_mhk` and `s_hkg`. The *diagnostic* cluster groups are mostly good indicators, with the notable exception of cluster groups 1, 3, 5 and 10. As we have no information on the meaning of these clusters, we can draw no further conclusions from this.

The forward selection selects substantially fewer variables from the age and gender (`s_lgnw` and 'source of income' (`s_avi`) categories. It appears that this is compensated for by the inclusion of more socio-economic status (`s_ses`) variables. This is possible in part due to the multicollinearity embedded into the variables by making an explicit division by age in the variables from the socio-economic status and source of income categories. Besides this deliberate multicollinearity, we suspect there is a strong correlation between source of income and socio-economic status, leading to ambiguity in the choice of variables.

While this variable selection procedure is of limited sophistication, it does point to two suggestions. First, the current model appears to be overfitting the data, as suggested by the improved Mallows' C_p for models with fewer variables. Second, the multicollinearity embedded (in part deliberately) into the model stands in the way of interpreting the individual significance of certain variables.

4.4 Advanced regression techniques

Summary

Goal Solving the problem of multicollinearity with other regression techniques.
Method Alternative regression techniques: LASSO regression, Ridge regression and Elastic Net.
Main result The Elastic Net outperforms both the Ridge regression and LASSO regression by achieving a much lower Mean Squared Error, while controlling the number of variables.
Recommendation Explore the impact of the new methods on the regression model (on individual level). Compare the current descriptives and compare the coefficients of the regression. Describe these models for non-mathematicians in order to let them comprehend these models and interpret the results.

The standard linear regression approach fails to find good explanatory models when the data contains strong multicollinearity. It is evident from Section 4.3 that the dataset has many collinear variables, so that one may miss significant explanatory variables. To avoid this issue, we propose the elastic-net approach from Zou and Hastie

(2005). This approach can be thought of as a combination of the established ridge (Hoerl and Kennard (1970)) and the modern LASSO (Least Absolute Shrinkage and Selection Operator, see Tibshirani (1996)) regression techniques. These approaches have in common that they add L_1 regularizations to the L_2 (= least squares) criterion in the regression procedure, and thereby handle the multicollinearity in the data. We refer to Hesterberg et al. (2008) for an accessible review where these methods are put in perspective, while extensive treatments can be found in the monographs Efron and Hastie (2016) and Hastie et al. (2015). However, ridge regression has the disadvantage that it does not lead a parsimonious model (it does shrink unimportant variables, but they do not shrink to zero). The LASSO does shrink unimportant variables to zero, but the LASSO selects at most n variables, where n is the number of observations. It also tends to select only one variable from a group of correlated variables, ignoring the others. To overcome these limitations, the elastic net adds a quadratic part to the penalty ($\|\beta\|_2^2$), where $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$ denotes the 2-norm of a vector $\mathbf{x}$. Mathematically, the regression problem is now written as

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - X\beta\|_2^2 + \lambda \left((1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \right) \right\}, \tag{1}$$

where $\mathbf{Y}$ are the actual costs, β is a vector of weights for variables, X represents the values of these variables for various postcodes, and $\alpha \in [0,1]$. The $\|\mathbf{x}\|_1 = \sum_i |x_i|$ denotes the 1-norm. In this way the elastic net combines the advantages of these methods while minimising their disadvantages. Sometimes a factor $1/2$ is put in front of the $\|\beta\|_2$ term for mathematical convenience. Note that the elastic net includes ridge regression and the LASSO as special cases through the choices $\alpha = 0$ and $\alpha = 1$, respectively.

The main drawback of the elastic net is that it requires to select appropriate values for the regularisation parameter λ and the elastic net parameterα. In principle, one can use cross validation techniques or a grid search to find optimal values for these parameters. For correct values of λ and α, we get a vector $\widehat{\beta}$, which highlights the most important variables in the regression.

To test the elastic net approach, we fit our model on two thirds of the data and test on the remaining data. We consider 11 different values for α from 0 to 1. In Formula (1), $\mathbf{Y}$ represents the total Somatic costs, while X represents the various variables corresponding to somatic costs except the postcode clusters, that is, all variables from Table 1 except `s_ape`.

Table 2 presented the mean squared error (MSE) for a few different tested values of the parameter α. The extreme values $\alpha = 0$ and $\alpha = 1$ correspond to ridge regression and LASSO, respectively. The minimal value is found for the parameter $\alpha = 0.7$

To present a little more insight into the behaviour of the regression techniques corresponding to different values of α we illustrate the regression result with Figures 3–5. Figure 3 shows, for the ridge regression ($\alpha = 0$), the coefficients $\widehat{\beta}$ against λ on the left and the mean squared error (MSE), measured as $\|Y - X\beta\|_2^2$, on the right-hand side. Similarly, Figures 3 and 4 show the variation for elastic net (with the optimal $\alpha = 0.7$, see Table 2) and LASSO ($\alpha = 1$) respectively.

α	MSE
0 (Ridge)	42960.8
0.2	41876.1
0.5	39314.6
0.7	**37012.1**
0.9	40353.4
1.0 (LASSO)	44236.1

Table 2: Mean-squared error for various values of the elastic net parameter α (on test data). Smaller values indicate better fit.

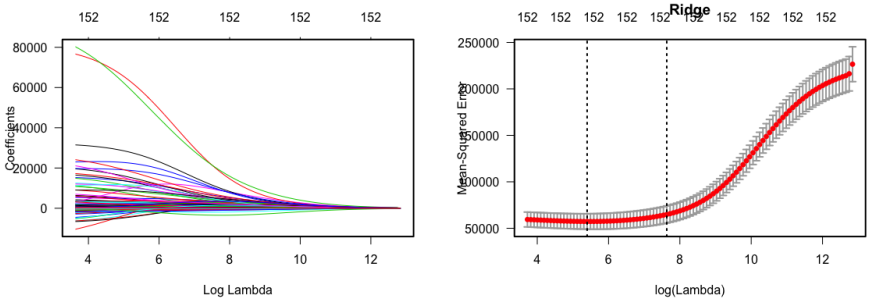

Figure 3: Ridge Regression ($\alpha = 0$). The number of coefficients (left) and the respective mean square error (right) variation with regularization parameter λ. This picture helps to find the right balance to a small number of coefficients (parsimony) while at the same time controlling the mean squared error (measure for model fit).

From each of the left hand-panels, we observe that only relatively few variables contribute strongly to the linear model. In the case of the LASSO, this is explained by multicollinearity. In the right-hand panels, vertical dashed lines indicate the region in which the MSE attains its minimal value.

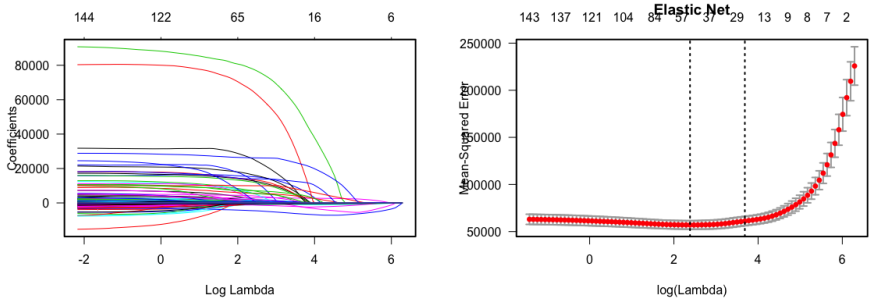

Figure 4: Elastic Net (for $\alpha = 0.7$). The coefficients (left) and the respective mean square error (right) variation with regularization parameter λ. This picture helps to find the right balance to a small number of coefficients (parsimony) while at the same time controlling the mean squared error (measure for model fit).

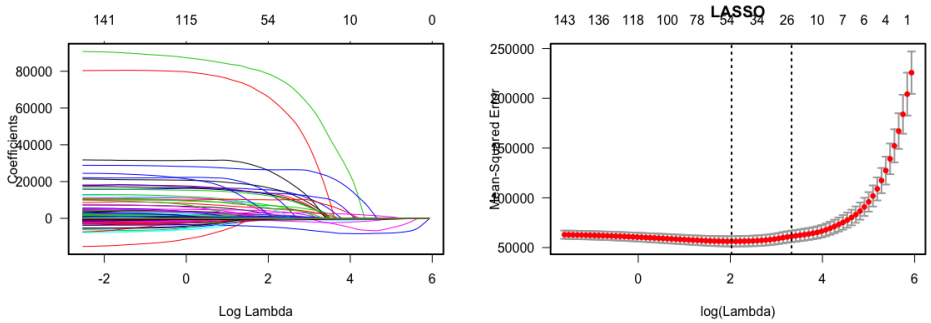

Figure 5: Lasso Regression ($\alpha = 1$). The coefficients (left) and the respective mean square error (right) variation with regularization parameter λ. This picture helps to find the right balance to a small number of coefficients (parsimony) while at the same time controlling the mean squared error (measure for model fit).

4.5 Clustering approaches

Summary

Goal Cluster postcode entries based on similarity of features, which constitute a proxy for the health profile. Within a cluster of postcode entries with similar feature values, a postcode entry with exceptional costs might indicate inefficiency.
Method We used a k-mean clustering method using a standard distance metric (Euclidean distance) on a transformed (e.g. splitting age and gender and collapsing age groups within AVI) features.
Main result There is a clear distinction between urban regions and suburbs, when clustered based on age as well as based on AVI. For the age clustering, the suburban postcode entries more frequently have low total health costs.
Recommendation This method can be extended by clustering based on different or larger sets of features. Moreover, we suggest to use this method as a new approach to discover regions with inefficient care, which is to search for outliers with high health care costs within clusters with the similar feature values.

In this section we explore the possibilities and benefits of cluster analysis on the provided dataset. We aim to use these techniques to find a natural structure present in the postcode entries (*instances*) based on the similarity in the explanatory variables (*features*). Our objective is twofold: we hope to gain intuition of the meaning of different combinations of explanatory variables, and we may find unknown patterns and relations unexplained by the regression models currently in use.

There exist many clustering techniques, but almost all of them follow similar steps:

1. Define a distance metric on the set of instances.
2. Formulate a decision rule that determines whether an instance belongs to a cluster.
3. Iteratively separate *or* group instances until every one is classified.

Among commonly techniques are hierarchical clustering, distribution-based clustering, and centroid-based clustering.

To provide some context: hierarchical clustering algorithms often evaluate the 'distance' between two observations, i.e. some quantitative notion representation of the difference between their properties. Nearby observations are linked to form a cluster, and nearby clusters are merged to larger clusters. This is a convenient strategy for exploratory clustering approaches, but expensive to apply to large data sets and difficult to interpret for high-dimensional data.

Distribution-based clustering algorithms try to find a set of clusters by choosing from a family of distributions that matches the observations.

While this collection of algorithms generally has no difficulty clustering all observations, these methods are prone to overfitting data. In addition, for many observation properties (especially in our dataset) it is difficult to find an underlying family of distributions.

Centroid-based clustering is based on the assumption that each cluster has a central observation: a centroid. Observations are classified based on their distance to the

nearest centroid. The benefits of this type of approach is that the created clusters have an intuitive and well-defined meaning: they can be interpreted by their centroid observations. The downside is that a priori the number of clusters must be known, and that the method allows only for the creation of clusters with a specific shape: convex clusters.

In our analysis, we choose a centroid-based clustering algorithm: k-means clustering. This choice is based on the fact that the data is high-dimensional and that before we start our analysis, we lack knowledge of how the data is structured or how the clustering could be interpreted. It is worth the effort to investigate if other algorithms are more suitable.

4.5.1 Distance metric

In our implementation, we used the function `kmean` of the statistical software `R`, which performed the k-mean clustering using standard distance metrics such as the Euclidean distance. However, Euclidean distance has no valid meaning for the variables in their current state. Since the spread in money-related variables is much higher than in age-related variables, a transformation is required to ensure that the Euclidean distance is normalized. Furthermore, from an information entropy perspective, some variables in the data may be redundant. These are also transformed to a more compact form. In the following section, we explain in what way the data is non-uniform and redundant, and how we transform it.

4.5.2 Feature transformation

The variables as given in the data are of different types. Some count the number of individual belonging to a category of a binary attribute (e.g. the number of individual using certain drug) or to a category of a categorical attribute (e.g. the number of individuals with source of income high). As, for each attribute, each category is represented by a variable, there are as many variables as there are categories. In fact, some variables represent the count of a category from a certain age group, which is a finer resolution. As a result, an attribute that has been broken down in many categories and in age groups comprises many variables.

The exact distribution in age group for each category seems redundant, therefore we apply the following transformation:

- We collapse the age-gender categories into gender and add a feature with the gender-ratio per instance.
- We collapse the age-'source of income' categories into age.
- We convert the binary variables to ratios.

Although we have reduced the number of variables of some attributes, some attributes still have many variables. To uniformly distribute the contribution of each

attribute to the distance, we weight each with the reciprocal of the number of categories such that the total weight of all categories sums to one.

Finally, we weigh each of the variables with the reciprocal of its observed variance. This follows from our assumption to let all features be equally important.

Our transformation creates a uniform information representation, and combine some of them into new features. Although it loses some accuracy in the exact age distribution of some attributes, this greatly reduces the dimensionality of the data, which is a welcome benefit for the quality of the cluster analysis.

It should be noted that we do not have to include all features when we compute a cluster. Different subsets of features may yield different clusterings. For this reason, our scripts allow for an arbitrary subset of features.

4.5.3 Finding the optimal cluster

The k-means clustering algorithm is a probabilistic algorithm that finds clusters in the following way. First, it randomly chooses k centres in the feature space. Each of these centres represents an initial cluster. Then, for each cluster it repeatedly executes the following steps.

1. Find the closest instance and add it to the cluster.
2. Compute the new centre of all instances in the cluster.

The clustering is finished when all the instances belong to one of the k clusters. The quality of a set of clusters is determined by computing the sum of squares (CSS) of each of the instances with the centre of its cluster. The lower the CSS, the better the clustering. This procedure is repeated an arbitrary number of times, each time choosing randomly new initial locations, and finally choosing the clustering with the minimal CSS. The k-means clustering algorithm can find clusterings for any number of clusters. A heuristic way to determine the optimal number of clusters is to find the elbow in the plot of the CSS against the number of clusters.

4.5.4 Results

In our exploratory analysis, we used two different groups of features to generate two clustering. The first clustering uses the age categories as features and the second uses the source of income (`s_avi`). The elbow method determines the optimal number of clusters for both clusterings to be 3. A detail of the geographic distribution of the clusters using age categories is depicted in Figure 6. In both maps, clusters clearly distinguish between urban (Amsterdam, Utrecht, Almere, Amersfoort) and suburban areas. We now discuss the findings for the two clusters separately.
Figure 7 shows the characteristic profiles for the clusters based on age. The left-hand panel shows the distribution of people over the age groups per cluster, the right-hand panel shows the distribution of health care costs. We see that clusters 1 (black) and 3 (red) have relatively young age profiles, of which cluster 1 has the most centralized distribution of health care costs. One would expect young clusters (i) to be associated

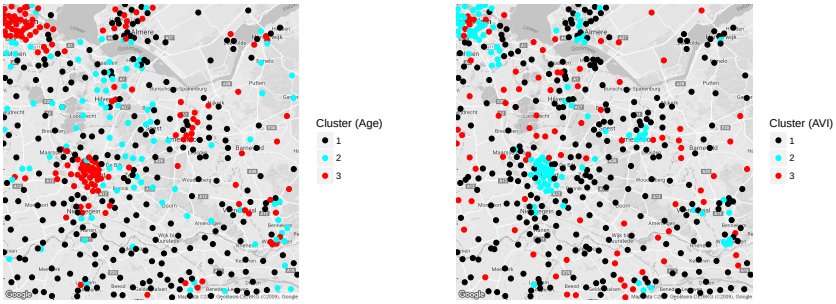

Figure 6: Detail of maps showing the result of clustering postcodes based on age (left) and based on AVI categories (right).

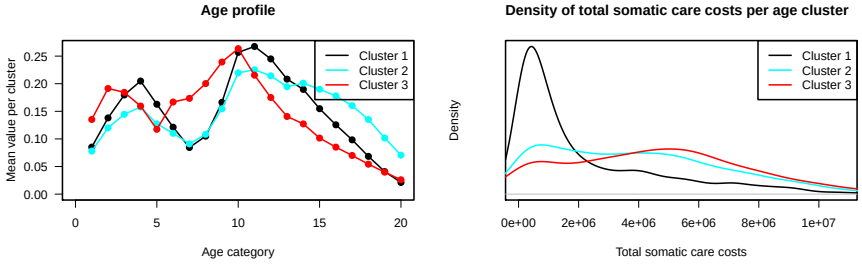

Figure 7: Cluster profiles when clustering based on age. Left: distribution of people over different age groups per cluster. Right: distribution of health care costs per cluster.

with better health and therefore low health care costs and (ii) to cluster around the urban area, which are somewhat confirmed by in our results.

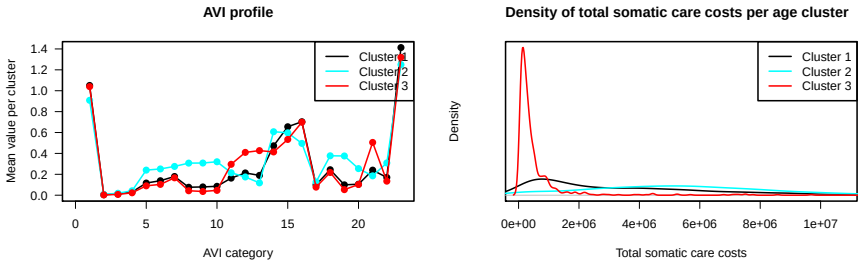

Figure 8: Cluster profiles when clustering based on source of income (`s_avi`). Left: distribution of people over different source of income categories. Right: distribution of health care costs per cluster.

Figure 8 shows the characteristic profiles for the clusters based on source of income (`s_avi`). The left-hand panel shows the distribution of people over the source of income categories per cluster, the right hand panel shows the distribution of health care costs. Cluster 1 (black) has the lowest distribution of health care costs. It would be interesting to see whether it agrees with the actual meaning of the avi-categories.

Due to time constraints, we were only able to analyse clustering based on age and avi-categories. For further analysis, it would be interesting to explore the clustering based on different sets of features. For instance, a larger set of features may better describe the health profile of regions.

Moreover, we suggest a new approach to discover municipalities with inefficient care, which is to search for outliers with high health care costs within clusters. Conceptionally, this approach is similar to outlier identification as discussed in Section 4.2, since it also searches for outliers after adjustment for explanatory variables.

5 Conclusions

The Dutch government compensates health insurance companies when insuring individuals who are estimated to require more or more expensive health care. This is necessary to avoid insurers avoiding certain groups or not providing them with a high quality of service. To estimate the required costs, the government uses a number of personal characteristics that are deemed good indicators of the general health status. The basis for this estimator is a linear regression model that fits the real health care costs based on the chosen parameters. It is, however, unclear whether the realized costs are due to a difference in health status, or due to other reasons that affect the health care expenses made. The model employed for the Dutch government is made for prediction rather than explanation.

We used several different techniques that investigate these differences, providing a first step towards understanding if these differences are preferably compensated for or not. The conclusions are somewhat diverse — in part due to how the research was carried out. Studying the outliers in the data using a linear regression model revealed no surprising results. We did, however, find that certain variables in the model are of much greater importance than others. Comparing Mallow's C_p for models using different subsets of the variables suggests the current model might suffer from overfitting. This method can help in simplifying the risk equalization model together with the clustering approach. Work needs to be done on the impact of these methods on the usual criteria for evaluating the risk equalization model. Elastic net regression provides for regularization of the model parameters (and thus avoids overfitting), at the cost of introducing a single metaparameters. With this regularization in place, it is easier to explain the impact of various parameters on quality of the regression. By applying clustering techniques we exposed a remarkable difference in the health expenditure between clusters based on only a few of the prognostic variables. Differences *within* these clusters may provide valuable information on possible causes for health care expenditure differences.

The techniques presented in this work all contribute to a greater understanding of the factors influencing the health care costs. One main conclusion is that we can leave out variables with no or limited loss of the quality of the model. The methods we used can help in selecting variables that contribute and variables that have no contribution. It can also help in selecting and combining variables in the current model in order to be capable of interpreting the results of the regressions and solve the issues with multicollinearity in the model.

6 Recommendations

Besides the methods presented above, we also have a number of suggestions for techniques that may improve the prediction, or improve the understanding of the factors at play. In particular we briefly discuss two approaches and their merits.

Since the data show significant differences across The Netherlands such as dependencies on the region, on the municipality, on the geographical location, and on the presence of academic hospitals just to name a few, it is strongly advisable to take into account this heterogeneity when fitting a unique model on all the available data. An option is the inclusion of random effects in a simple linear regression model, the so-called linear mixed model of Laird and Ware (1982). The approach is similar to linear regression, but now part of the observed effect is supposed to be due to some random effects. Consequently, the estimate is a sum of two terms: the product of a design matrix (i.e. a matrix of the covariates) with a vector of coefficients, the fixed effects, and the product of another design matrix (containing the same or other covariates) with a vector of random coefficients drawn from a particular distribution. Including these random effects shrinks the estimates of the first deterministic part towards the mean. Since the second term is a random sample drawn from a wider

population, this allows taking into account the effects of some variables that are only partially observed, such as the overall health status. Furthermore, this method is advantageous in terms of estimation. In fact, fitting a traditional regression model including a fixed effect for each unit can become cumbersome when the chosen unit is small, and thus the number of units and corresponding coefficients is large. The bigger advantage of random effects model with respect to fixed effects model is the reduction in the number of parameters. That is, if in a regression model we include a variable for each region (province, zipcode or any other cluster) and we have n regions, then we will have to estimate n coefficients (one for each region). Instead, if we include the random effect for that same variable (region, province, zipcode or any other cluster), then the number of parameters to be estimated reduces to 2 (mean and variance) in case of assumption of normality for the random effects. I.e. the n coefficients (one for each region) are samples from a normal distribution with mean μ and variance $sigma^2$.

Introducing the random effects reduces to one the number of parameters to be estimated to account for the heterogeneity among units, namely the variance of the random effects. For further details and implementation, see for example Verbeke and Molenberghs (2009), Verbeke et al. (2010), and Fitzmaurice et al. (2008).

Another approach is given by the mixture model, which fit different distributions (or the same distribution with different parameters) for each region (province, zipcode or any other cluster). If all the data can be modelled by the same distribution, then there is no need for mixture model - the data is homogeneous. On the other hand, if there is heterogeneity in the data, it can be captured by this flexible model fitting different distributions of the data on different regions.

This method can be used not only to fit a completely new model, but also to check whether the distribution of the data of a certain supposed subpopulation is indeed different from the others. Since it is possible to include and merge almost any desired distribution, it results in an extremely flexible and thus powerful tool in capturing and explaining heterogeneity.

Many statistical softwares have built-in procedures to fit both linear mixed models and mixture models, see R, SAS, STATA, SPSS and Matlab among others. As highlighted earlier, heterogeneity is visible at different levels, thus when fitting these models one might want to explore the effect size of various subpopulations at different scales, such as postcode, municipality, province, and so on.

References

B. Efron and T. Hastie. *Computer Age Statistical Inference.* Cambridge University Press, 2016.

G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs. *Longitudinal Data Analysis.* CRC Press, 2008.

S. Gilmour. The interpretation of mallows's C_p-statistic. *The Statistician*, 45(1): 49–56, 1996.

U. Grömping. Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2):139–147, 2007.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity*. CRC Press, Boca Raton, Florida, 2015.

T. Hesterberg, N. Choi, L. Meier, and C. Fraley. Least angle and ℓ_1 penalized regression: A review. *Statistical Surveys*, 2:61–93, 2008.

A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

T. Kvålseth. Cautionary note about R^2. *The American Statistician*, 39(4):279–285, 1985.

N. Laird and J. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

C. Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Series B (Methodological)*, pages 267–288, 1996.

G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer Science & Business Media, 2009.

G. Verbeke, G. Molenberghs, and D. Rizopoulos. Random effects models for longitudinal data. In *Longitudinal Research with Latent Variables*, pages 37–96. Springer, 2010.

Zorginstituut Nederland. Zvw 2017, 2017. URL `https://www.zorginstituutnederland.nl/financiering/risicoverevening-zvw/zvw-2017`.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.