Risk Adjustment Network Conference (Netherlands, 18-20 September 2014)

Communicating Vessels vs. Regression Approach

Lennart Pirktl | Federal Office of Public Health (FOPH) | Switzerland

September 20, 2014

Summary

In the context of risk adjustment there is an easy way to compute R^2 without performing an actual OLS regression and without using large individual data sets.

This method is highly recommended whenever regression through the origin is chosen.

If the number of risk classes needs to be reduced and fusion candidates of interest are those with minimal loss of R^2, then both Ward's and k-means algorithm fulfill this goal.

Refinement of Swiss risk equalization in 2017 leads to a high number of classes which can efficiently be reduced from 3120 ($R^2 = 29.4$) to 1612 ($R^2 = 28.3$).

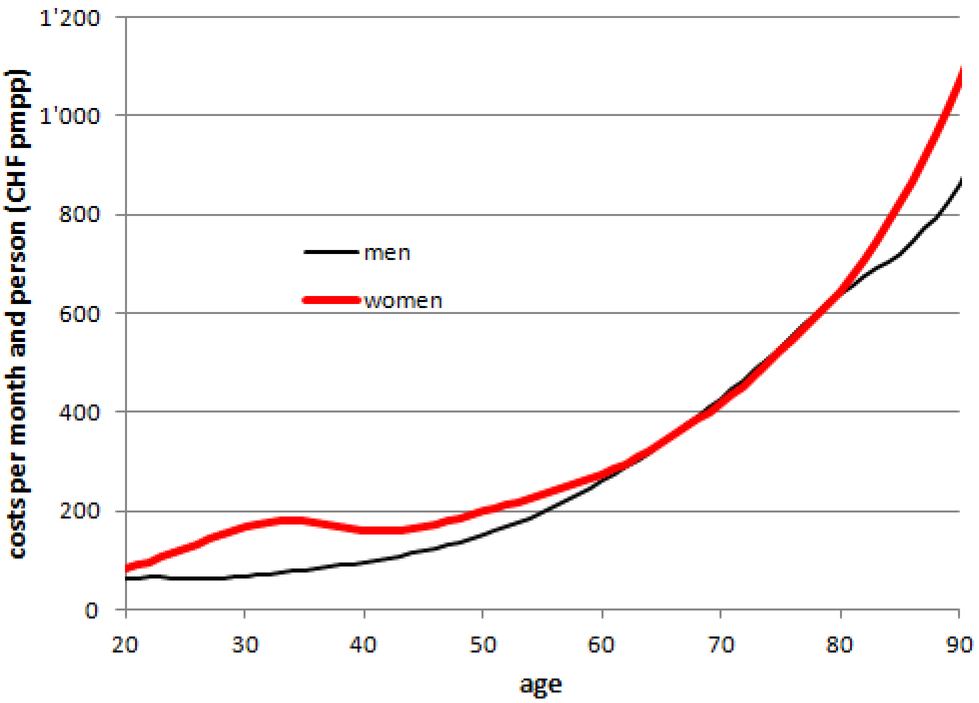
For a given set of predictor variables $X_1, \dots, X_p$ the effectiveness of risk equalization (RE) depends on the prediction qualities of the chosen regression model $L = f(X_1, \dots, X_p)$, where the response L denotes the estimated costs of an individual, based on his demographic and health characteristics $X_1, \dots, X_p$. If the model is additively separable $\theta(L) = f\left(X_1, \dots, X_p\right) = \sum_{j=1}^{p} \varphi_j(X_j)$ then optimal transformations $\varphi_1^*, \dots, \varphi_p^*$ are determined by the ACE algorithm (Breiman and Friedman 1985).

The Swiss model started with the demographic indicators AGE, SEX and CANT (Switzerland being divided into 26 cantons) without considering pharmaceutical cost groups (MEDI) or hospitalization categories (HOSP). Risk equalization is calculated individually for each canton.

In 2012 a simple dichotomous HOSP-variable was introduced: “Patient stayed more than three nights in hospital during previous year”.

In 2017 a simple dichotomous MEDI-variable will be introduced: “Previous year value of consumed drugs by patient exceeds CHF 5,000”, before introducing more explicit PCG types in future years.

Since women's and men's cost curves have quite different shapes,



additive separability does not hold and it is preferable to keep the model in its most general form of $L = f(X_1, \dots, X_p)$, where all combinations of the discrete variables AGE(15), SEX(2), HOSP(2), CANT(26) and soon MEDI(2) are considered.

In order to optimize the threshold value for drug consumption MEDI, we analyzed individual records for 4.7 million adults, containing AGE, SEX, CANT, indicator HOSP for people who stayed more than three nights in hospital during 2010, and gross drug expenditures in 2010 and net health care expenditures (L_i) in 2011 for mandatory basic coverage.

Since the number of records is fairly large, we adopted Beck's view and notation of the Swiss RE method: Beck (2013, p.407f) shows that the "practical method" used to determine the rules of risk adjustment is equivalent to the regression on the $m + 1$ dummy variables $x_{i,j}$, where there are n insured persons $(1 \leq i \leq n)$ with individual expenses L_i belonging to exactely one of the $m + 1$ classes $1 \leq j \leq m + 1$ containing d_j persons each.

Discarding class $m + 1$ to prevent multicollinerarity, Beck shows that the solutions of

$$L_i = a_0 + \sum_{j=1}^{m} a_j x_{i,j} + \varepsilon_i$$

are

$$a_0 = \bar{L}_{m+1}$$

$$a_j = \bar{L}_j - \bar{L}_{m+1},$$

where

$$\bar{L}_j = \frac{1}{d_j} \sum_{i=1}^{n} x_{i,j}\, L_i \;= \frac{L_j}{d_j}$$

denotes the average cost in class j. Therefore the solution is simply

$$\hat{L}_i = a_0 + \sum_{j=1}^{m} a_j x_{i,j} \;= \bar{L}_{m+1} + \sum_{j=1}^{m} (\bar{L}_j - \bar{L}_{m+1}) \cdot x_{i,j} \;= \bar{L}_{k_i},$$

where individual i belongs to class k_i. This shows that the solution $(\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{Y}$ of normal equations $(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$ can be obtained directly by calculating one-dimensional means $\bar{L}_j$ and thereby avoiding the inversion of $\boldsymbol{X}'\boldsymbol{X}$.

Of course one has the choice between two alternatives: omitting one arbitrarily chosen variable from the regression

$$L_i = a_0 + \sum_{j=1}^{m} a_j x_{i,j} \;+ \varepsilon_i$$

or performing regression through the origin (RTO):

$$L_i = \sum_{j=1}^{m+1} a_j x_{i,j} \;+ \varepsilon_i$$

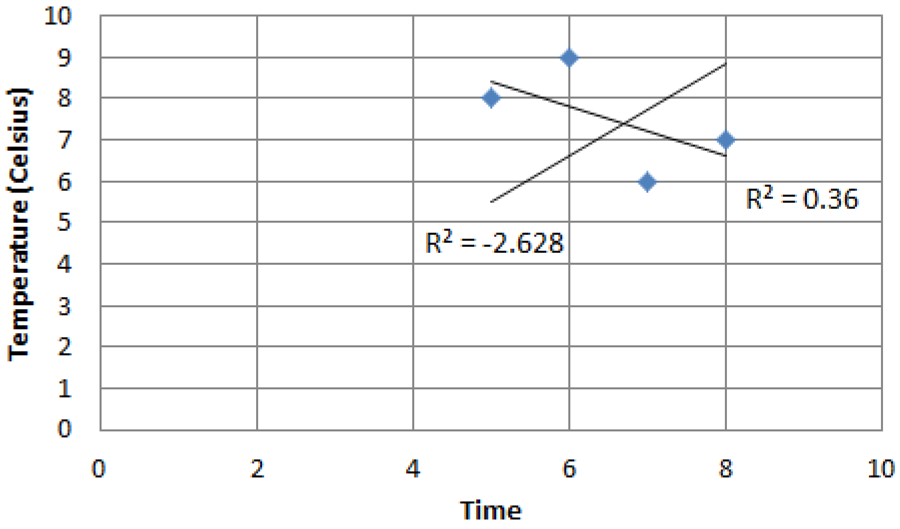
Although both alternatives produce identical results $\hat{L}_i = \bar{L}_{k_i}$, certain statistical programs fail to list meaningful R^2-values when RTO is chosen. Using a simple example, Pirktl and Square (1985) illustrated that SAS, SPSS and BMDP increase R^2 in an absurd way when $a_0 = 0$ is forced. In that example R^2 increased from 0.36 to 0.92. When expressing temperature in °K instead of °C R^2 increased even more to 0.97.

Using the same data set (called MIRACLE) 30 years later with EXCEL, one still gets misleadingly high values of R^2. At least the graphical output obtained by choosing

> "*highlight chart, add trendline, select linear trend/regression type,*
> *set intercept=0, display R-squared value on chart*":

provides the user with the following complex but rather intuitive result:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(\hat{Y}_i - Y_i)^2}{\sum(Y_i - \bar{Y})^2} = -2.628$$



$R^2 < 0$ means that the residual sum of squares $\sum(\hat{Y}_i - Y_i)^2$even exceeds the original total sum of squares $\sum(Y_i - \bar{Y})^2$ for RTO.

The misleading high values of R^2 in the printed output (.92 and .97) result from the following definition

$$R^2 = \frac{\sum \hat{Y}_i^2}{\sum Y_i^2}$$

Therefore it advisable to calculate R^2 "by hand" whenever RTO is chosen, in order to avoid the above mentioned trap of not knowing whether a correct formula was used for R^2 or not. The appropriate formula is:

$$R^2 = \frac{1}{ns_L^2} \sum_{j=1}^{m+1} d_j(\bar{L}_j - \bar{L})^2$$

Proof:

The most general and meaningful definition of this coefficient of determination is

$$R^2 = 1 - \frac{RSS}{TSS},$$

where the total sum of squares

$$TSS = SS_{tot} = \sum_{i=1}^{n} (L_i - \bar{L})^2$$

can be partitioned into the explained sum of squares (ESS) and residual sum of squares (RSS) as follows:

$$ESS = SS_{reg} = \sum_{i=1}^{n} (\hat{L}_i - \bar{L})^2 = \sum_{j=1}^{m+1} d_j (\bar{L}_j - \bar{L})^2$$

$$RSS = SS_{res} = \sum_{i=1}^{n} \hat{\varepsilon}_i{}^2 = \sum_{i=1}^{n} (L_i - \hat{L}_i)^2 = \sum_{j=1}^{m+1} d_j (L_i - \bar{L}_j)^2$$

Therefore, R^2 can be obtained without knowledge of individual data as the variance of predicted expenses divided by the variance of observed expenses:

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n} \left(\hat{L}_i - \bar{L}\right)^2}{\sum_{i=1}^{n} (L_i - \bar{L})^2} = \frac{\sum_{j=1}^{m+1} d_j (\bar{L}_j - \bar{L})^2}{\sum_{i=1}^{n} (L_i - \bar{L})^2} = \frac{1}{n s_L^2} \sum_{j=1}^{m+1} d_j (\bar{L}_j - \bar{L})^2$$

The easy way of calculation is one of the reasons, why R^2 is the most commonly used evaluation method for prediction models. But there is a large field of such measures. A comprehensive overview can be found in van Veen et al. (2013).

Conclusions

1) The risk equalization regression problem

$$L_i = a_0 + \sum_{j=1}^{m} a_j x_{i,j} + \varepsilon_i$$

may be solved without individual data. The solutions are

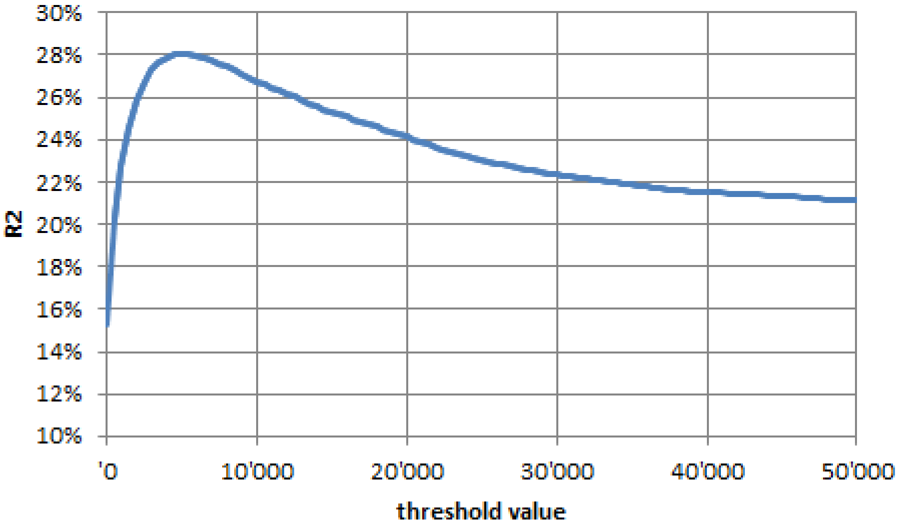
$$a_0 = \bar{L}_{m+1} \,,\; a_j = \bar{L}_j - \bar{L}_{m+1} \quad \text{with} \quad R^2 = \frac{1}{n s_L^2} \sum_{j=1}^{m+1} d_j (\bar{L}_j - \bar{L})^2 .$$

2) There is no need for individual data as long as insurers supplement their data delivery $\{d_j; L_j\}$ with the sum of squared yearly expenses $\{d_j; L_j; Q_j\}$:

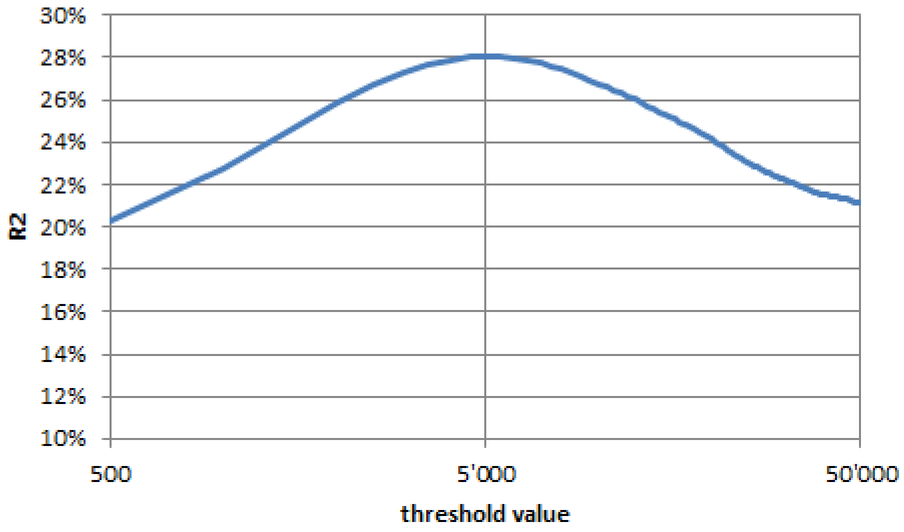
$$R^2 = \frac{1}{n s_L^2} \sum_{j=1}^{m+1} d_j (\bar{L}_j - \bar{L})^2 = \frac{\left(\frac{1}{n} \sum d_j \bar{L}_j{}^2\right) - \bar{L}^2}{\left(\frac{1}{n} \sum Q_j\right) - \bar{L}^2} = \frac{\left(\sum d_j \bar{L}_j{}^2\right) - n\bar{L}^2}{\left(\sum Q_j\right) - n\bar{L}^2} \quad .$$

3) Furthermore, with $\bar{L}$, s_L^2 and n being constants, R^2 is maximized when the variance of the $\bar{L}_j$ is maximized; which may even be determined without knowledge of Q_j.

Using this "shortcut", R^2 was easily evaluated for more than 1000 threshold values determining whether an individual falls into MEDI category "high drug consumption in year *t-1":*



Obviously 5,000 CHF is a good choice for the MEDI threshold value:
(same curve with logarithmic scale)



Reduction of the number of risk groups

Not only is there a simple way to compute R^2 without performing the actual OLS regression, but there is also a simple rule how to reduce the number of classes with minimal loss of R^2 whenever the number of classes becomes too high:

For any pair of candidate classes p and q, their fusion reduces

$$R_{before}^2 = \frac{1}{n s_L^2}\left[\left(\sum_{j=1}^{m+1} d_j\, \bar{L}_j^2\right) - n\bar{L}^2\right]$$

to

$$R_{after}^2 = \frac{1}{n s_L^2}\left[\left(\sum_{j\notin\{p,q\}} d_j \bar{L}_j^2\right) + (d_p + d_q)\left(\frac{d_p\bar{L}_p + d_q\bar{L}_q}{d_p + d_q}\right)^2 - n\bar{L}^2\right].$$

Calculating the “loss” of $ns_L^2R^2$ as the immediate result of a potential fusion of classes p and q we get:

$$\begin{aligned}
(R^2_{before}{-}R^2_{after})\cdot ns_L^2 &= d_p\bar{L}_p^2 + d_q\bar{L}_q^2 - n\bar{L}^2 - (d_p + d_q)\left(\frac{d_p\bar{L}_p + d_q\bar{L}_q}{d_p + d_q}\right)^2 + n\bar{L}^2 \\
&= d_p\bar{L}_p^2 + d_q\bar{L}_q^2 - (d_p + d_q)\left(\frac{d_p\bar{L}_p + d_q\bar{L}_q}{d_p + d_q}\right)^2 = d_p\bar{L}_p^2 + d_q\bar{L}_q^2 - \frac{\left(d_p\bar{L}_p + d_q\bar{L}_q\right)^2}{d_p + d_q} \\
&= \frac{(d_p + d_q)(d_p\bar{L}_p^2 + d_q\bar{L}_q^2) - \left(d_p\bar{L}_p + d_q\bar{L}_q\right)^2}{d_p + d_q} \\
&= \frac{d_p^2\bar{L}_p^2 + d_pd_q\left(\bar{L}_p^2 + \bar{L}_q^2\right) + d_q^2\bar{L}_q^2 - \left(d_p\bar{L}_p + d_q\bar{L}_q\right)^2}{d_p + d_q} \\
&= \frac{d_p^2\bar{L}_p^2 + d_pd_q\left(\bar{L}_p^2 + \bar{L}_q^2\right) + d_q^2\bar{L}_q^2 - d_p^2\bar{L}_p^2 - 2d_pd_q\bar{L}_p\bar{L}_q - d_q^2\bar{L}_q^2}{d_p + d_q} \\
&= \frac{d_pd_q\left(\bar{L}_p^2 + \bar{L}_q^2\right) - 2d_pd_q\bar{L}_p\bar{L}_q}{d_p + d_q} = \frac{d_pd_q\left(\bar{L}_p^2 - 2\bar{L}_p\bar{L}_q + \bar{L}_q^2\right)}{d_p + d_q} = \frac{d_pd_q\left(\bar{L}_p - \bar{L}_q\right)^2}{d_p + d_q} \\
&= \frac{d_pd_q}{d_p + d_q}\cdot\left(\bar{L}_p - \bar{L}_q\right)^2 \geq 0\,,
\end{aligned}$$

which is the distance concept used in Ward’s hierarchical clustering procedure:

$$D_{p,q} = \frac{d_pd_q}{d_p + d_q}\cdot\left\|\bar{L}_p - \bar{L}_q\right\|^2.$$

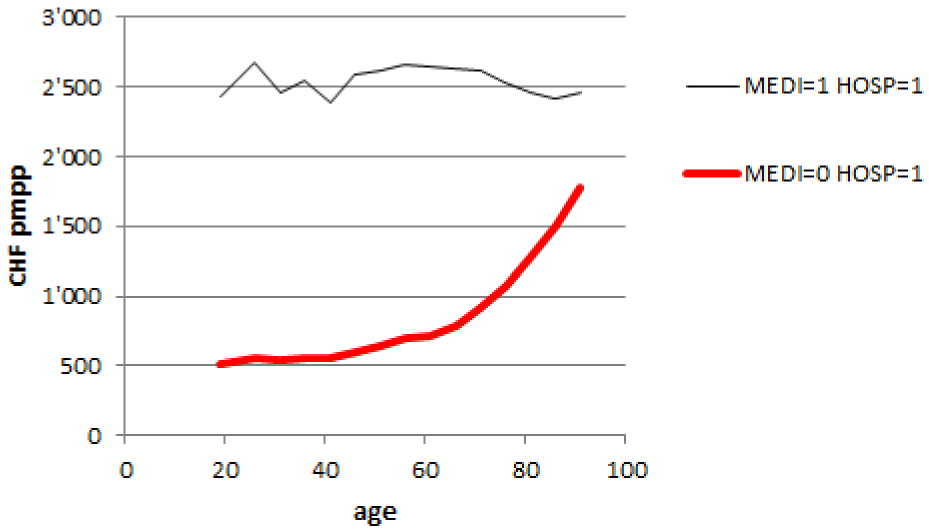
For fusion candidates (p, q) and (p', q') with equal distance $\left\|\bar{L}_p - \bar{L}_q\right\|^2 = \left\|\bar{L}_{p'} - \bar{L}_{q'}\right\|^2$ those with larger size difference $\left|d_p - d_q\right| > \left|d_{p'} - d_{q'}\right|$ will be merged first.

The result of Ward’s algorithm consists of a division of the $m + 1$ initial risk classes into $k \leq m$ groups. Therefore, one might consider Ward’s classification as an initial guess subject to refinement by a k-means algorithm:

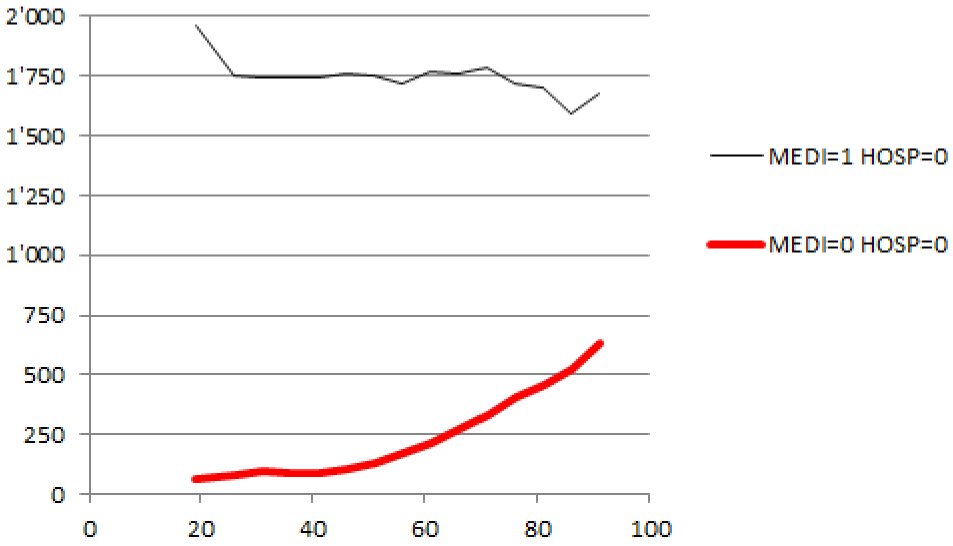
This method iteratively considers switching risk class $1 \leq j \leq m + 1$ from its actual group k_1 to any of the $k - 1$ other groups, let’s say to k_2 , whenever $\bar{L}_j$ of the corresponding risk class is closer to $\bar{L}_{k_2}$ than to $\bar{L}_{k_1}$ in the sense of squared Euclidean distance.

In the case at hand, the introduction of an additional dichotomous variable MEDI doubles the number of cells from 1560 to 3120 and brings the system to its limits as sparse populated cells get increasingly influenced by random effects and for the first time there are even a few empty cells, which is disturbing in the case of ex-ante (prospective) risk equalization.

Interestingly future costs are not substantially influenced by age among persons in the group of previous high drug expenses (MEDI=1). This is true within the group of persons with previous hospitalization HOSP=1:



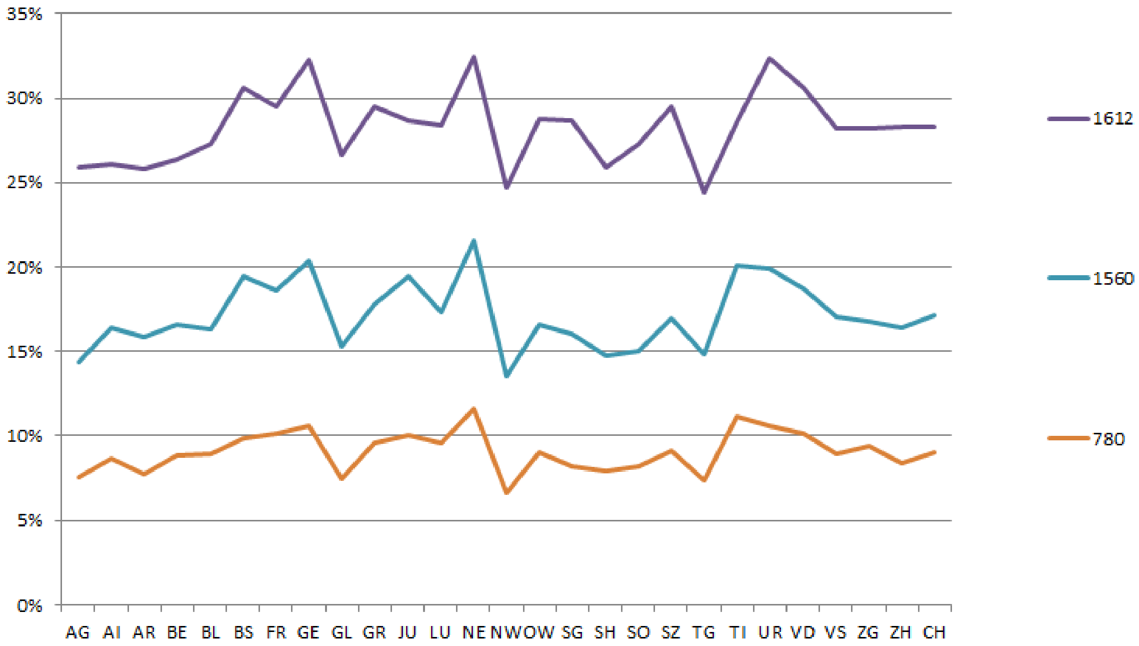
as well as for the larger group HOSP=0:



This led to the solution of merging all age and sex groups within MEDI=1 without losing much R^2:

Number of risk classes	R^2	
3120	29.4	full model
2340	28.9	dropping SEX within MEDI=1
1612	28.3	dropping AGE within MEDI=1
1560	17.2	dropping MEDI (=situation 2012-2016)
780	9.1	dropping HOSP (=situation 1993-2011)

Of course these values differ by cantons and will also differ over time:



Introduction of HOSP Introduction of MEDI

$R^2 \approx 9\% \Longrightarrow R^2 \approx 17\% \Longrightarrow R^2 \approx 28\%$

Notation

$x_{i,j}\ = 1$ if individual i belongs to class j ($= 0$ if not)

$$d_j = \sum_{i=1}^{n} x_{i,j} = \text{number of individuals in class } j$$

$$n\ = \sum_{j=1}^{m+1} d_j$$

$L_i =$ expenses of individual i during year t

$$L_j = \sum_{i=1}^{n} L_i\, x_{i,j} = \text{sum of expenses of all individuals in class } j$$

$$\bar{L}_j = {}^{L_j}\!/_{d_j} = \text{mean cost in class } j$$

$$\bar{L} = \frac{\sum L_j}{\sum d_j} = \frac{\sum L_j}{n} = \text{overall mean cost}$$

$$Q_j = \sum_{i=1}^{n} L_i^2\, x_{i,j}$$

$$s_j^2 = \frac{1}{d_j} Q_j - \bar{L}_j{}^2 = \text{variance within class } j$$

$$s^2 = \frac{1}{n}\sum Q_j - \bar{L}^2 = \text{overall variance}$$

$$R^2 = \frac{\left(\sum d_j \bar{L}_j{}^2\right) - n\bar{L}^2}{\left(\sum Q_j\right) - n\bar{L}^2}$$

$$\Delta R^2 = R^2_{before} - R^2_{after} = \frac{1}{ns^2} \cdot \frac{d_p d_q}{d_p + d_q} \cdot \left(\bar{L}_p - \bar{L}_q\right)^2 \geq 0$$

Literature

Beck K. et al. (2013), Risiko Krankenversicherung – Risikomanagement in einem regulierten Krankenversicherungsmarkt, Haupt, Bern

Breiman, L. and J.H. Friedman (1985), Estimating Optimal Transformations for Multiple Regression and Correlation, Journal of the American Statistical Association, 80, 580-598

Pirktl L. and Max R. Square (1985), Neue Perspektiven in der empirischen Forschung, RZU aktuell, 54/11

van Veen, S.H.C.M., R.C. van Kleef, W.P.M.M. van de Ven and R.C.J.A. van Vliet, (2013), A Systematic Review of Measures for evaluating Prediction Models in Risk Equalization, Working paper, Institute Health Policy & Management, Erasmus University Rotterdam, Presented at the Risk Adjustment Network Conference 2013, Tel Aviv, Israel

Ward, J.H., Jr. (1963), Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association, 58, 236–244