Israel Diego (UMID: israeldi)
Stats 504
December 5th

Medicare Services clustered by Geographical Structure

The Russia Longitudinal Monitoring Survey (RLMS) consists of detailed survey data that monitors Russian individuals' health status, dietary intake, and personal consumption. We aim to understand the impact of various health factors on an individual's health evaluation score.

Our approach for modeling the health evaluation score is to use dimension reduction regression model which aims to show the most important directions of our data. Reducing the dimensionality of our data also makes local smoothing computation more tractable. Specifically, we apply Sliced inverse regression (SIR), where given a response variable y and predictor variable $x \in R^p$, the model is characterized by :

$$y = g(b'_1 x, \ldots, b'_k x)$$

where $p > k$, the b_j vectors are the regression coefficients, and g is an unknown link function. Thus, we effectively reduce the dimension from p to k and our focus is on estimating the regression "directions" b_j, not the link function g.

The RLMS data contains survey respondent information dating from 1994 through 2018. We exclude survey responses where the individual did not know, refused to answer, or no answer was given. Our response variable, health-evaluation score, is a categorical variable ranging from 1 to 5 (1-Very good health, 2-Good, 3-Average, 4-Bad, 5-Very bad). We consider the following predictor variables,

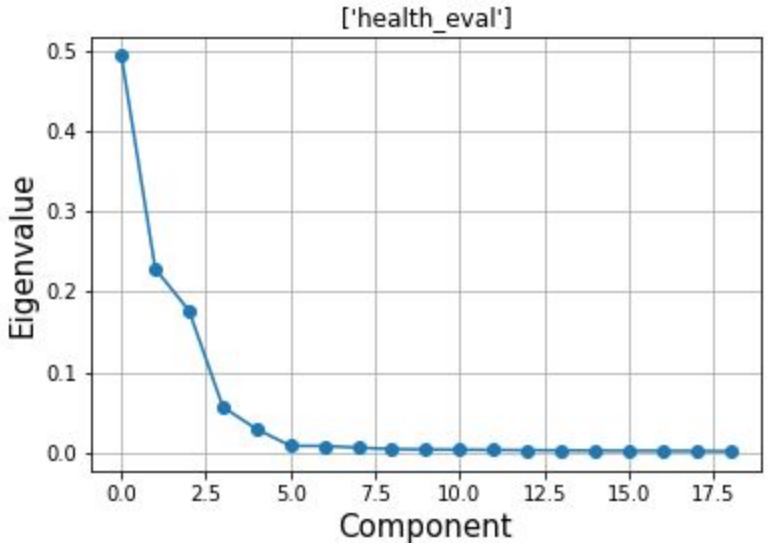
Table 1

Binary Health Variables	Other predictors

smokes: Smokes? **diabetes:** Ever diagnosed with diabetes? **chr_spinal:** Chronic spinal disease? **chr_stomach:** Chronic stomach disease? **chr_kidney:** Chronic kidney disease? **chr_liver:** Chronic liver disease? **chr_lung:** Chronic lung disease? **chr_heart:** Chronic heart disease? **heart_attack:** Ever diagnosed with heart attack?	**weight:** self-reported weight (kg) **height:** self-reported height (cm) **after_tax_wages:** after tax wages last 30 days **work_status:** equals 1 if working **marital_staus:** categorical indicating marriage status **occupation:** categorical indicating group type of occupation **hours_worked:** hours worked last 30 days **female:** female or not **status:** urban, rural, oblastnoy center, or PGT

Our aim is to understand how different health conditions contribute to an individual's health evaluation score. First, in **Figure 1**, we plot the eigenvalues of our fitted data after performing SIR on the centered data,

Figure 1



We effectively reduced the dimensionality of the data since we can explain about 98.5% of variation with 5 dimensions, 95% with 4 dimensions, and 90% with 3 dimensions. For brevity,

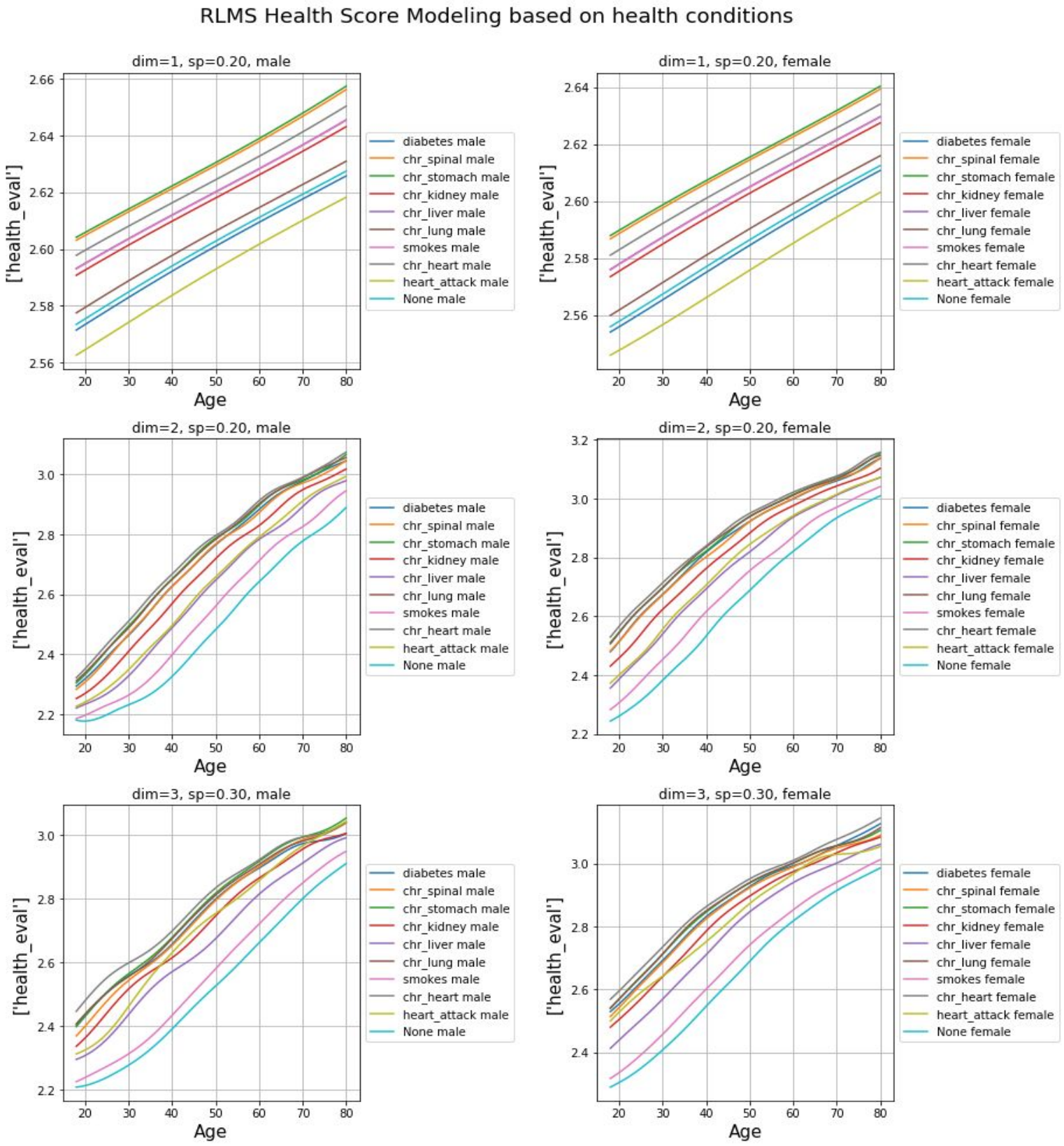
we explore the first three dimensions. To construct the predictions of the health score, we apply the Gaussian kernel function to gender and one health condition at a time, and plot the score as a function of age. All other variables were fixed at their average. Results are shown in **Figure 2** (page 4).

First we note that health evaluation scores increase with age; recall that a higher value indicates a less healthy individual. In the first dimension, plots of both men and women who have had a heart attack or have diabetes are predicted to be healthier than individuals with no condition, which is a bit counterintuitive. Individuals with chronic stomach disease or chronic spinal disease scored highest, i.e. were least healthy. However, the range of these scores is quite small, between 2.56-2.66 which is between good health and average health.

In the second dimension, results are more reasonable and closer to what we might expect. Individuals with no condition scored the lowest (good health), while individuals with diabetes or chronic diseases such as stomach, lung, heart, or spinal scored highest (average health). The range of the predicted values has also grown to between 2.1 and 3.1. We also notice that young females in their 20's start at higher health scores than men do and converge to about the same scores as men when reaching age 80. The third dimension results are similar to the second dimension, with some lines shifting in between. In both cases, smoking was the least harmful health condition, and the chronic diseases mentioned in dimension 2, including diabetes, resulted in the highest health scores.

We showed how dimension reduction regression can be applied to high-dimensional data in order to understand the relationship between predictors and the response without having to specify the mean-structure of the model. In our experiment, we focused on a health score and analyzed how different health factors predict an individuals health score, and thus gave some indication as to what health risk factors were deemed be most unhealthy.

Figure 2


RLMS Health Score Modeling based on health conditions
dim=1, sp=0.20, male
dim=1, sp=0.20, female
dim=2, sp=0.20, male
dim=2, sp=0.20, female
dim=3, sp=0.30, male
dim=3, sp=0.30, female
['health_eval']
Age
diabetes male
chr_spinal male
chr_stomach male
chr_kidney male
chr_liver male
chr_lung male
smokes male
chr_heart male
heart_attack male
None male
diabetes female
chr_spinal female
chr_stomach female
chr_kidney female
chr_liver female
chr_lung female
smokes female
chr_heart female
heart_attack female
None female