Exercise

Preparing diabetes data for modelling

Julian Flowers

2024-04-22

Contents

Introduction

In this exercise we will be using data from the English Public Health Outcomes Framework (PHOF). We will be downloading data from source and the exercise is to test the relationship between diabetes care processes and outcome.

To do this we will be performing supervised analysis of diabetes data from the PHOF - constructing linear models with diabetes outcome as the dependent variable and care processes and the independent (predictor) variables. To allow for diabetes frequency we will also include diabetes prevalence in our models, as well as a summary measure of socio-economic status (SES). In England, the Index of Multiple Deprivation 2019 (IMD) is widely used as a summary SES index.

We will be analysing the data for sub-national health administrative units called sub-ICBs (SICB). England is subdivided into 106 SICBs - these are the units of health care planning and performance.

Thesae data can be obtained directly from the

Get started

First we need to load the R packages we need to extract data and for analysis.

```
needs(tidyverse, pak, tidymodels, mgcv, glmnet, corrplot)
```

Get the data

Now we load the diabetes data. In the PHOF, these data have a ProfileID of 139 and SICBs have an AreaTypeID of 66. The code segment below shows how data is loaded.

```
diabetes_model <- read_csv("~/spha/data/dm_model_data.csv", show_col_types = FALSE)
diabetes_codebook <- read_csv("~/spha/data/dmmod_codebook.csv", show_col_types = FALSE)
```

The resulting dataset has 104 records, and consists of 81 diabetes metrics.

Explore the data

The first step is explore relationships between variables. A common way to do this is to construct a correlation matrix or a correlogram. In R, the `corrplot` package is widely used for this.