Household Index of Socioeconomic Status

Australian Government
Department of Health

Australian National University

Motivation

- There is ongoing interest in understanding the relationship between socioeconomic status (SES) and health care outcomes to guide policy and program decisions.
- A socioeconomic indicator at the household level holds opportunity for a more fine-grain understanding of SES.
- Existing measures of SES are based on census data, updated every 5 years.
- Some of the Socio-Economic Indexes for Areas (SEIFA) include health outcomes which complicate how they can be used for analysing health outcomes.

Partnerships

- This project is being done in partnership with ABS and ANU. We have a project steering committee that provides advice about variables, methods and interpretation.
- The results presented today are preliminary and as such are not for further distribution.
- If you are interested in learning more about this project please contact either myself (richard.hurley@health.gov.au) or Dr Allison Clarke (allison.clarke@health.gov.au).

Australian Government
Department of Health

Australian National University

Project Aims

1. Identify which variables from MADIP can be used as a proxy for socio-economic status at household level.
2. Determine from which of the available data sources in MADIP each variable should be sourced from.
3. Construct a new index of socio-economic status from the selected variables and tailored to the study healthcare outcomes.
4. Validate the performance of the new index.

Data

- MADIP data from 2016 was used to create the index.

- The variables selected are related to:
  - Personal Income Tax (PIT)
  - Social Services' Social Security and Related Information (SSRI).

Variable Selection

- The choice of initial variables that are used is crucial, since this will determine what the index is actually measuring.
- Based on discussion with our steering committee of experts we included these variables.

Variable	Type	Source
Income	Numeric	PIT
Rent	Numeric	SSRI
Total Welfare Received	Numeric	SSRI
Occupation	Categorical	PIT
Type of Accommodation	Categorical	PIT
Number of Children	Numeric	SSRI
Marital Status	Categorical	SSRI
Private Health Insurance Cover	Binary	PIT
Duration on Income Support	Numeric	SSRI

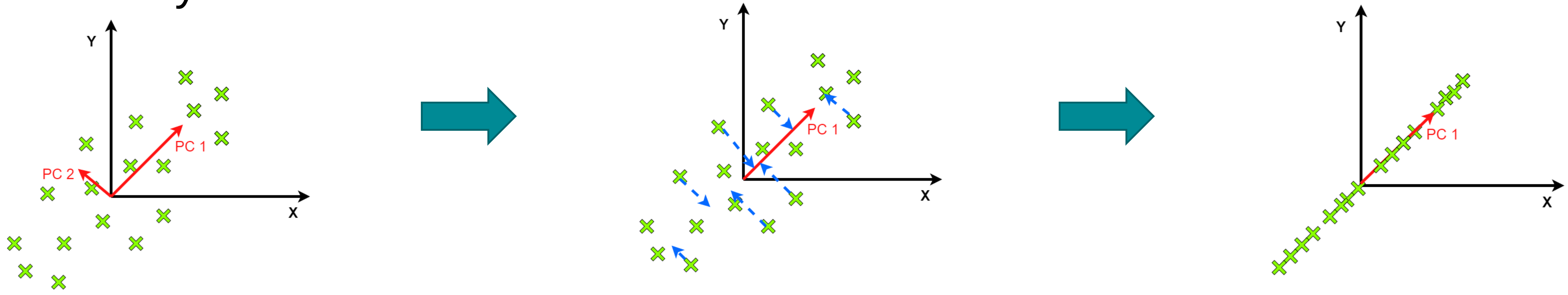
Methodology

- We followed the general methodology of previous indexes of SES released by ABS (SEIFA) which used Principal Component Analysis (PCA).

- PCA takes a dataset of individuals described by many characteristics and attempts to describe each individual with just one number.

Principal Component Analysis

- PCA compresses a set of given variables into just **one variable** (the index).
- PCA works by identifying the direction of greatest variation in the data.
- Variables can then be projected onto this direction to create a one-number summary for each individual.



Original input data

Reconstructed data

Validation

Conditions
Diabetes (Type 2)
High Cholesterol
High Sugar Levels
Heart Attack
Depression
Alcohol and Drug Problems

- To evaluate the index, we use data reconstructed from the index to **predict various healthcare conditions** (as reported on the National Health Survey) and see how this performs compared to the original input data.

Australian Government
Department of Health

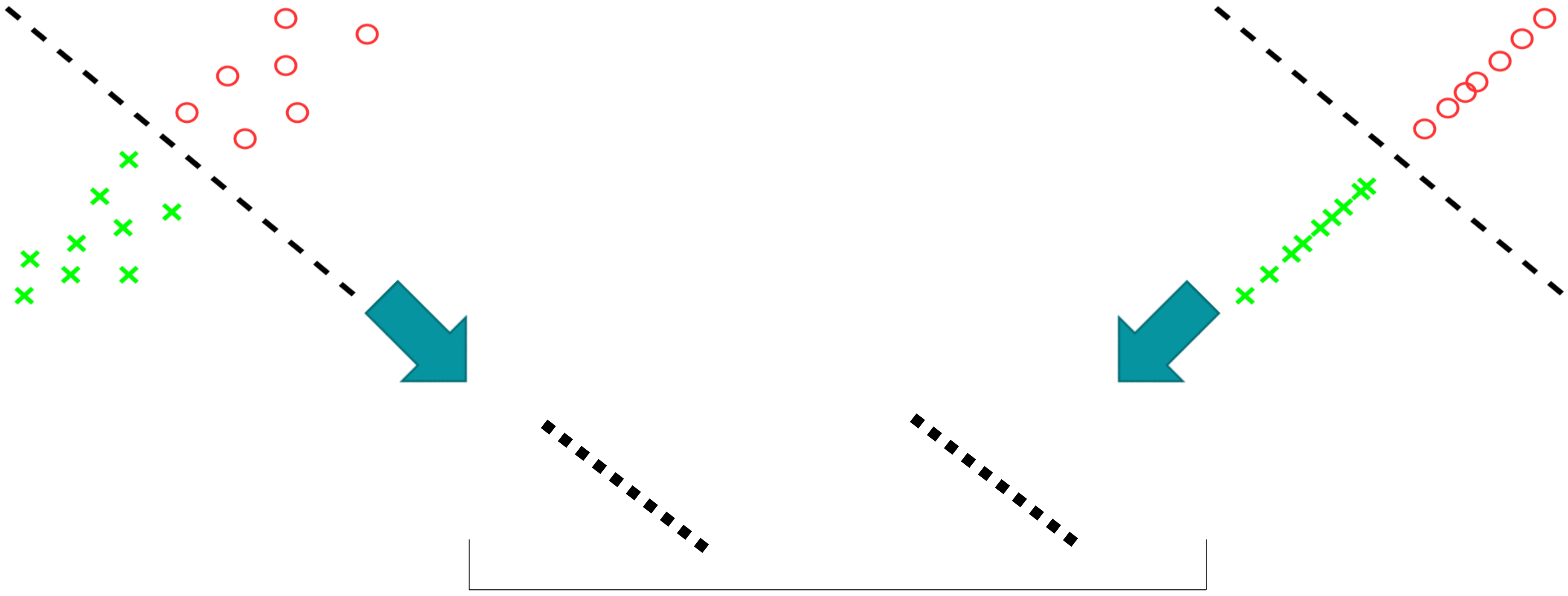
Australian
National
University

Validation

- For each condition, we construct **two logistic regression models** to predict the target condition
  - One uses the *original data* as input.
  - One uses *data reconstructed from the index* as input.

- We then **compare the linear coefficients** from each of the models
  - If the index represents the original variables faithfully we should get the *same* coefficients.
- We measure the *faithfulness* of the index as the cosine similarity between the coefficients.

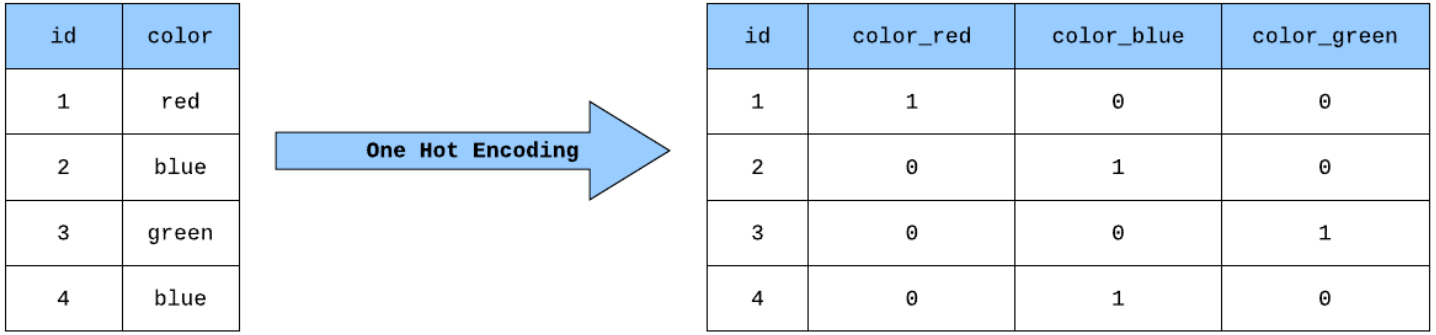
Faithfulness



Compare

Variable Encoding

- We used raw numeric values instead of binary indicator variables with thresholds (e.g. income vs income levels).
  - All variables are normalized to have mean 0 and standard deviation 1.
- Categorical variables were one-hot encoded
  - Only the 5 most common values are encoded, all others are marked as ‘other’.



id	color
1	red
2	blue
3	green
4	blue

One Hot Encoding

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Handling Missing Values

- Many of our variables have a significant proportion of missing values.
- We investigated 3 strategies for handling them
  1. Impute with 0
  2. Impute with mean

Imputation	Accuracy	Faithfulness
Zero	65.2	0.238
Mean	57.2	0.117

Aggregation

- We wish to construct an index of households, so we need to aggregate records from individuals who are living together in the same household.
- We investigate 2 strategies for aggregation
  1. Mean
  2. Maximum

Aggregation	Accuracy	Faithfulness
Max	65.2	0.238
Mean	65.1	0.234

Results – Constructed Index

Variable	Loading
Income	0.452
Rent	-0.179
Total Welfare Received	-0.322
No Occupation	-0.382
Number of Children	-0.192
Married	-0.218
Does not Live in Shared Accommodation	-0.231
Private Health Insurance Cover	0.393
Duration on Income Support	-0.433

Results – Prediction Accuracies

Condition	Original Accuracy	Index Accuracy
Diabetes (Type 2)	71.8	57.2
High Cholesterol	55.9	53.9
High Sugar Levels	67.5	54.0
Heart Attack	68.0	56.1
Depression	61.9	55.0
Alcohol and Drug Problems	75.6	68.7

Future Work

- Investigate different strategies of aggregation and imputation
  - For example, impute income with mean but impute rent with 0.

- Explore adding additional variables in the index.

- Apply to other health care outcomes (e.g. death, mental health).

Thank you

Australian Government
Department of Health

Australian National University