Multivariable Linear Regression

A case study

Praveen Gowtham

A question from the king

- His munificence wishes to know the average life expectancy of his subjects and factors involved.
- Zamunda's health ministry has historically not kept a good record of many things.

Figure: His royal excellency Jaffe Joffer, king of Zamunda (long may he reign)

Figure: Map of the border of Zamunda

Acquired data

Health data sampled from local hospitals. Socioeconomic data borrowed from other agencies.

Data + definitions

Measurement	Value	Definition
Alcohol	6.45	Consumption (L) / capita
Heptatitis B	93.0	Immunization coverage
Measles	6	Cases/1000
BMI	33.2	Index Value
Polio	96	Immunization Coverage
Total expenditure	4.93	% Gov. expend. on health
HIV/AIDS	14.4	Deaths / 1000 births
GDP	5374	per capita (USD)
Population	1884238	Number
Thinness 1-19 years	9.6	%
Schooling	11.9	Num. of Years

Acquired data

Use World Health Organization (WHO) life expectancy data to estimate for Zamunda. Multivariable linear regression.

Data + definitions

Measurement	Value	Definition
Alcohol	6.45	Consumption (L) / capita
Heptatitis B	93.0	Immunization coverage
Measles	6	Cases/1000
BMI	33.2	Index Value
Polio	96	Immunization Coverage
Total expenditure	4.93	% Gov. expend. on health
HIV/AIDS	14.4	Deaths / 1000 births
GDP	5374	per capita (USD)
Population	1884238	Number
Thinness 1-19 years	9.6	%
Schooling	11.9	Num. of Years

Multivariate Linear Models

- Multivariable Linear Model:

$$Y = w_1 x_1 + w_2 x_2 + ... + w_0 + \epsilon$$

- Example from last time:

$$Sales \approx w_{TV} TV + w_{Radio} Radio + w_0$$

The challenge of multiple feature

- Multivariable model:

$$Y = w_1 x_1 + w_2 x_2 + ... w_n x_n + w_0 + \epsilon$$

- Which coefficients should be included in the model to improve approximating the data? How to tell?
- Can't visualize anymore.
- p-value testing: which w_i significantly different from 0?

Highly correlated variables

Correlation between features can cause problems:

$$Y = w_1 x_1 + w_2 x_2 + w_0$$

- Suppose x_1 and x_2 are highly correlated. (e.g, $x_1 \approx 2x_2$)

$$Y_0 - w_0 = \Big(w_1 + 2w_2\Big)x_1$$

- Makes the weights unstable/unreliable.
- Makes weights very sensitive to training data (overfitting).