AC 209a
Data Science

We Are What We Eat

Leonard Loo - Xindi Zhao - Yihang Yan

Introduction and Data Exploration
– Facts About Food and Diabetes

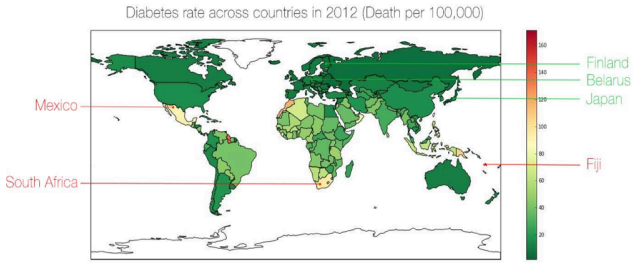
We have some general idea of "healthy food" or "unhealthy food", but have not really looked at this issue in a scientific way. In this project, we investigate the relationship between food and Diabetes.

Datasets

- **Disease data** - from the World Health Organization website

We chose to study Diabetes risk, given its increasing problem. We have data for the years 2000 and 2012 for the number of deaths by Diabetes per 100,000 people.

↑ 3% **increase** from 2000 to 2012



Diabetes rates are generally low for most countries (mostly green).
Highest Diabetes: **Fiji**, **Mexico**, and **South Africa**
Lowest Diabetes: **Finland**, **Belarus**, and **Japan**

- **Food data** - from the FAOSTAT

Consumption of food g/capita/day for 127 different food items (crops and livestock) for all countries from 1961 to 2013

At a glance, we can see what people in the world are eating. Rice, fish and maize are the top food per capita on average. Overall, carbohydrate intake is high.

Missing data

Diabetes data is clean, while food data have missing years for some food items. We experimented with the following ways to fill in missing food quantities:

- 2-Nearest-Neighbor: Take the average of two most recent years
- Linear regression: Fit on present years to predict missing years

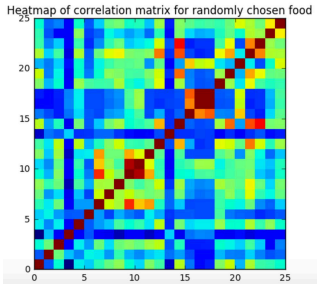
2-NN was computationally efficient, but we chose linear regression because most missing data were at the boundaries.

Models and Predictions
– How Food Impact Diabetes

Prediction Methodology

- Including all years of all food as the predictors, we will have thousands of predictors in total – yielding high dimensional issue.
- Instead take the mean consumption across a certain number of year before the disease data year.
- We illustrate using 10 years average, but the parameter can be tuned. That is, we use the average from 1991 to 2000 of food consumption to predict the disease rate in 2000, and use the average from 2002 to 2012 of food consumption to predict disease rate in 2012.
- **Train the model based on food items and Diabetes rate from year 2000 and test on 2012 food and Diabetes rates.**

Predictor Investigation



From the snap shot of the correlation plot between the 25 food items, we see most of the predictors are not very correlated.

Some are still reddish off-diagonal -> We are open to regularization methods for better prediction.

Model Selection

- **Baseline 1 – Null Model**
  - All predictions set to global average - always gives 0 R squared value
- **Baseline 2 – Linear Regression**
  - All food items as predictors – may have included too many terms
- **Ridge**
  - L1 regularization on loss function – do not allow predictor selection.
- **LASSO**
  - L2 regularization on loss function – allows variable selection.
- **PCA**
  - Dimension reduction by forming linear combination of the original predictors
- **Multivariate Adaptive Regression Spline**
  - Non-parametric regression technique that can be seen as an extension of linear models which automatically models nonlinearities and interactions between variables
  - May potentially yield overfitting issues

Model	Linear	Ridge	LASSO	PCA(n=40)	MARS
R^2	-0.250	0.513	0.556	0.539	-0.152
Training R^2	0.848	0.800	0.632	0.594	0.6906

Linear regression model may have too many predictors without penalty, and MARS may be overfitting the data with "hinge" functions so even though they both have high training R square, the testing performance is not desirable.
Ridge, LASSO and PCA all give satisfactory results, by including regularization or dimension reduction.
Overall, **Lasso** seems to be a promising model to use for prediction based on both the testing R square value and interpretability.

Analysis and Recommendation
– So What We Should Eat

By inspecting the coefficients of Lasso, we can see their relationship with Diabetes. As we standardize the predictors, we can compare across food items to see which are most and least positive:

Most positive

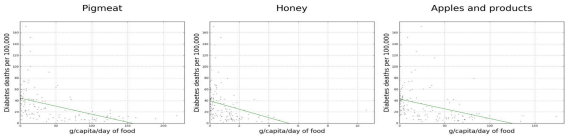
Food item	Coefficient
Peas	5.269141
Soyabean Oil	3.559297
Oilcrops	3.545669
Spices, Other	3.156651
Barley	2.915349
Maize	2.052394
Sugar	2.037754

Most negative

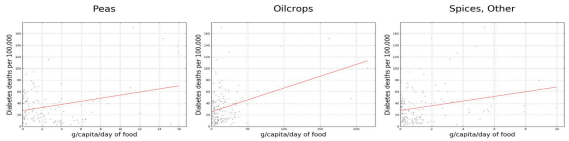
Food item	Coefficient
Potatoes	-5.826556
Pig meat	-4.317692
Eggs	-3.680632
Milk	-2.632306
Tomatoes	-1.309889
Apples	-1.159267
Fish	-1.034791

Positive coefficients are positively correlated with Diabetes and are not "healthy". Negative coefficients are negatively associated with Diabetes and are "healthy". We plotted the linear association of some of these food with Diabetes rates:

Most "Healthy" food



Least "Healthy" food



Recommendation

- A healthy diet that leads to a lower Diabetes rates should include plenty of **Potatoes, Pig meat, Eggs, Milk, Tomatoes, Apples** and **Fish**. This is in line with conventional wisdom.
- Try to avoid consuming too much **Peas, Spices, Barley, Maize** or **Sugar.**