# CYO project: Obesity and Macronutrients

*bart-g*

*6/10/2020*

## Contents

1. Overview . . . . . . . . . . 1
   - Goal . . . . . . . . . . 1
   - Methods . . . . . . . . . . 1
   - Definitions . . . . . . . . . . 2
   - Source of data . . . . . . . . . . 3
2. Analysis . . . . . . . . . . 3
   - 2.1 Data preparation & cleaning . . . . . . . . . . 3
   - 2.2 Data partitioning . . . . . . . . . . 4
   - 2.3 Data exploration . . . . . . . . . . 4
   - 2.4 Regression methods . . . . . . . . . . 9
3. Results . . . . . . . . . . 23
4. Conclusions . . . . . . . . . . 24

## 1. Overview

### Goal

The goal of this project is to explain obesity, as defined by percentage of high BMI individuals in society (body-mass-index), using macronutrients' supply in all countries available.

### Methods

Regression methods used in the study:

- k-Nearest Neighbors,
- General Additive Model using Splines,
- General Additive Model using Loess,
- Stochastic Gradient Boosting

Computational environment, as used by author:

- Windows 10 OS,
- Intel i5 CPU,
- 8 GBs of RAM,
- Microsoft R Open distribution was used (improved performance considerations)

Minimum memory requirement is 4GB. Expected computation time was 10 minutes on author's desktop PC.

### Definitions

(1) Macronutrient

Macronutrients, the vital substances, necessary for organism to survive are included as independent variables:

- plant proteins,
- animal proteins,
- fats,
- carbonhydrates.

Water, though being a macronutrient was not included in the research.
Macronutrient variables are defined as supply of kilo calories (kcal) per day per person, sampled annually.
Notion of supply does mean availability of the macronutrient however doesn't measure the factual intake of calories. Yet, food supply, when categorised with macronutrients, offers the insight of growing intake potential of energy (calories).
When intake is not met by balanced expenditure of energy, we observe growing figures of obesity in our societies.
Which, brings to the idea of capturing overweight and obesity at extreme.

(2) Body mass index

Body mass index (BMI) is defined as:

$$BMI = \frac{weight[kg]}{height^2[m^2]}$$

People of BMI $<$ 25 are considered normal,
those with BMI $<$ 30 overweight,
whereas BMI $>=$ 30 indicates obesity.

### Is BMI the right measure?

Its definition can be questioned, since weight to height ratio isn't exactly explaining body fat ratio. Secondary shortcomings, the lack of age and gender differentation are not part of the BMI model. For instance, women typically have higher body fat percentage than men for the same BMI. Another aspect, considering professional sports, is the higher weight of muscle mass which exaggerates BMI score for actually healthy individuals, carrying less fat tissue thank others.

**Source of data**

Set 1: **Share of obese adults per country**

https://ourworldindata.org/obesity#13-of-adults-in-the-world-are-obese

Sampling period: 1975-2016.

Availability: local file in csv format

File name: share-of-adults-defined-as-obese.csv

Set 2: **Diet compositions by macronutrient per country**

https://ourworldindata.org/diet-compositions#diet-compositions-by-macronutrient

Original data comes from FAO, and kindly postprocessed by OutWorldInData members. Imporantly, carbonhydrates were deduced from total caloric supply.

http://www.fao.org/faostat/en/#data/FBSH

Sampling period: 1961-2013.

Availability: local file in csv format.

File name: daily-caloric-supply-derived-from-carbohydrates-protein-and-fat.csv

## 2. Analysis

### 2.1 Data preparation & cleaning

Data is loaded from two CSV formatted files.

Share of obese adults, snapshot:

| country | year | y_obese |
|---|---|---|
| Afghanistan | 1975 | 0.5 |
| Afghanistan | 1976 | 0.5 |
| Afghanistan | 1977 | 0.6 |
| Afghanistan | 1978 | 0.6 |
| Afghanistan | 1979 | 0.6 |
| Afghanistan | 1980 | 0.7 |

Daily caloric supply [kcal], snapshot:

| country | year | animal_proteins_supply | plant_proteins_supply | fat_supply | carbonhydrates_supply |
|---|---|---|---|---|---|
| Afghanistan | 1961 | 54.12 | 285.52 | 337.59 | 2321.77 |
| Afghanistan | 1962 | 53.92 | 278.00 | 338.49 | 2246.59 |
| Afghanistan | 1963 | 56.80 | 251.68 | 347.13 | 2042.39 |
| Afghanistan | 1964 | 57.32 | 276.64 | 350.55 | 2268.49 |
| Afghanistan | 1965 | 59.76 | 275.68 | 357.57 | 2262.99 |
| Afghanistan | 1966 | 64.44 | 252.24 | 359.55 | 2060.77 |

Share of obese adults as uploaded from file:

| Obesity Share (total records) | First year | Last year |
|---:|---:|---:|
| 8316 | 1975 | 2016 |

Daily caloric supply as uploaded from file:

| Macronutrients Supply (total records) | First year | Last year |
|---:|---:|---:|
| 8981 | 1961 | 2013 |

Annually merged obesity data across all countries:

| Obesity data across countries (total records) | Countries count | First year | Last year |
|---:|---:|---:|---:|
| 6305 | 167 | 1975 | 2013 |

## 2.2 Data partitioning

Accordingly train and test set were randomly split 80/20.

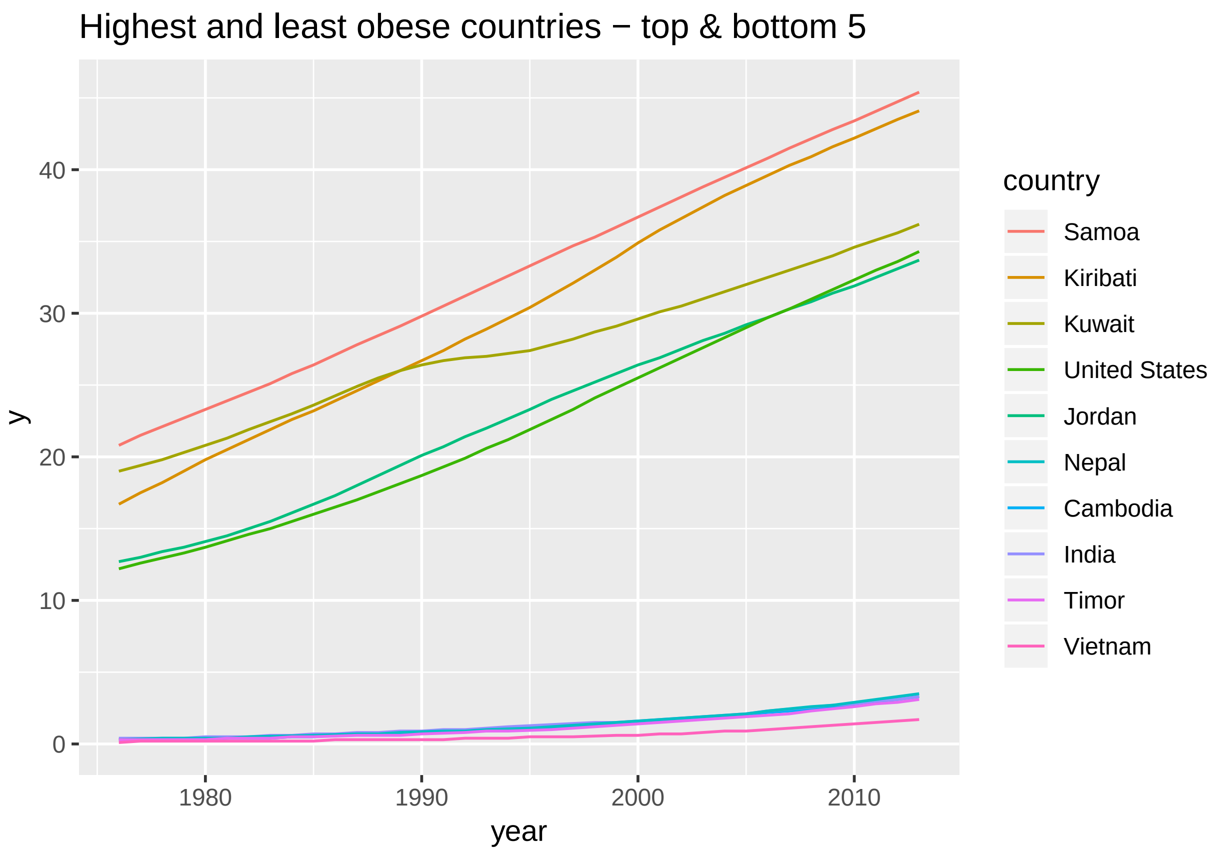
After all, given combined obesity data and macronutrients supply data, the most recent year we can use for further research is 2013.

## 2.3 Data exploration

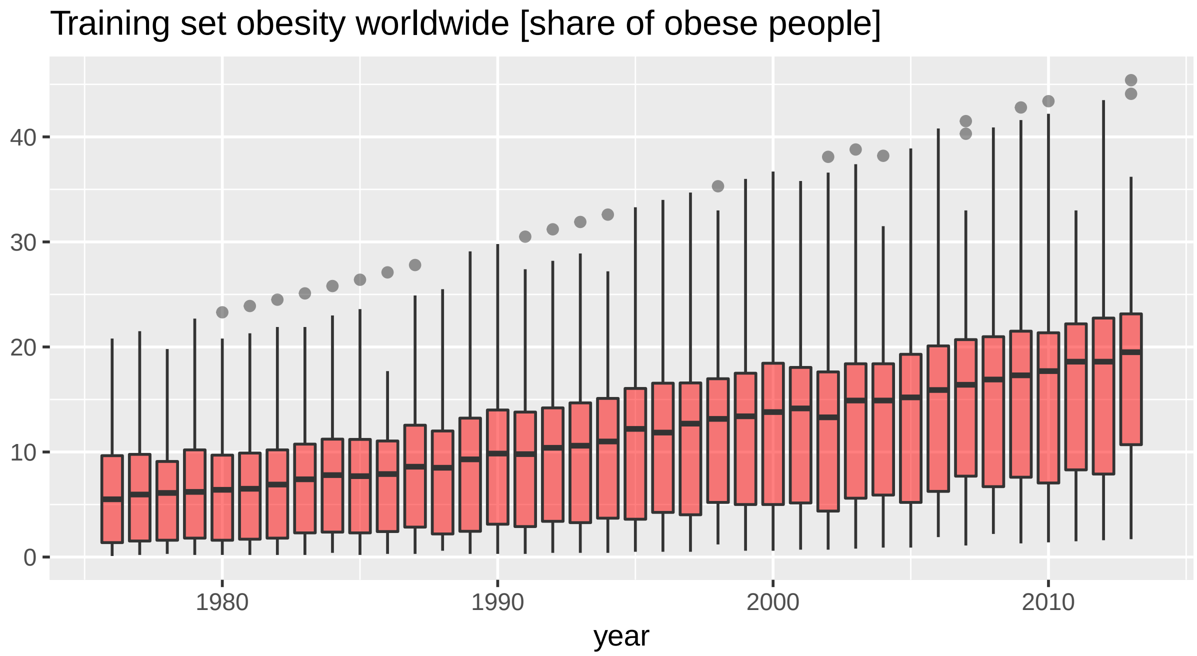
### Is there a commonality among countries observing obesity?

Just looking at top 5 and lowest 5 countries per obesity share among adults the picture emerges.

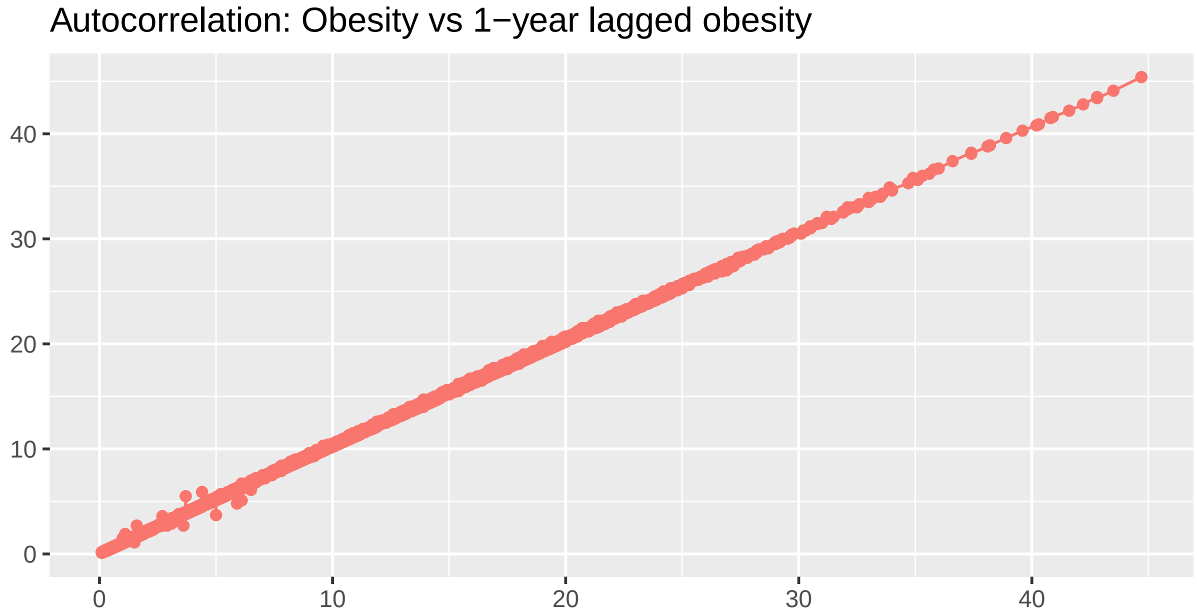
The most and the least obese societies observe upwards slope over the time. The least obese has very weak slope compared to the most dynamically obese countries.



Level of obesity just by pure glimpse at the extreme samples, suggests that past level of obesity impacts the future. It may encourage us to capture this autoregressive element if necessary. Here is another, more global snaphot of adults' obesity over the years.



QQ plotting shows strong autocorrelation of obesity with its lagged variable.



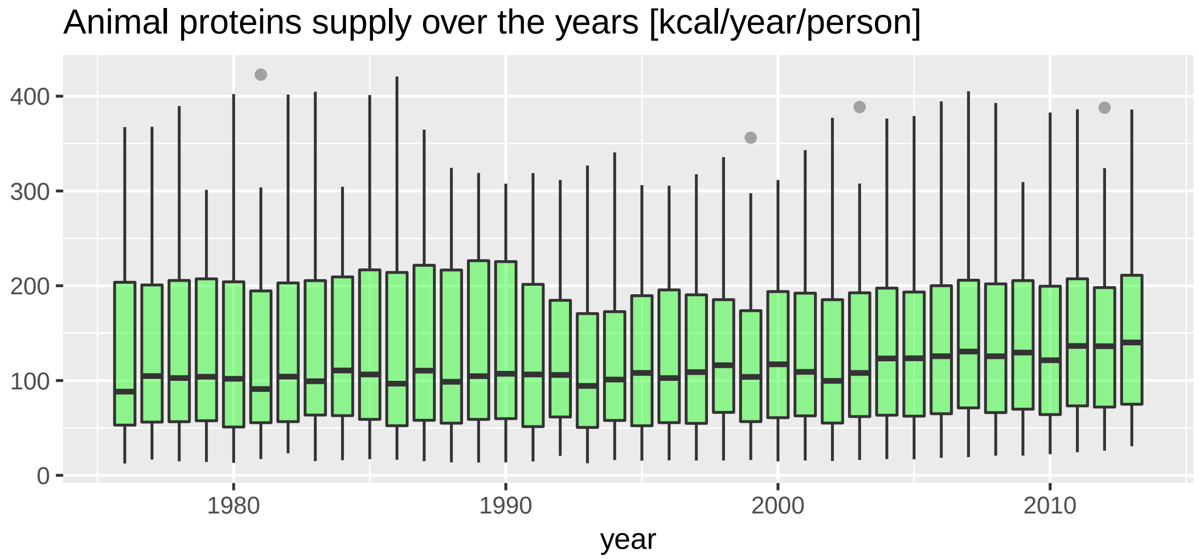
For that reason data set is enhanced with lagged variable:

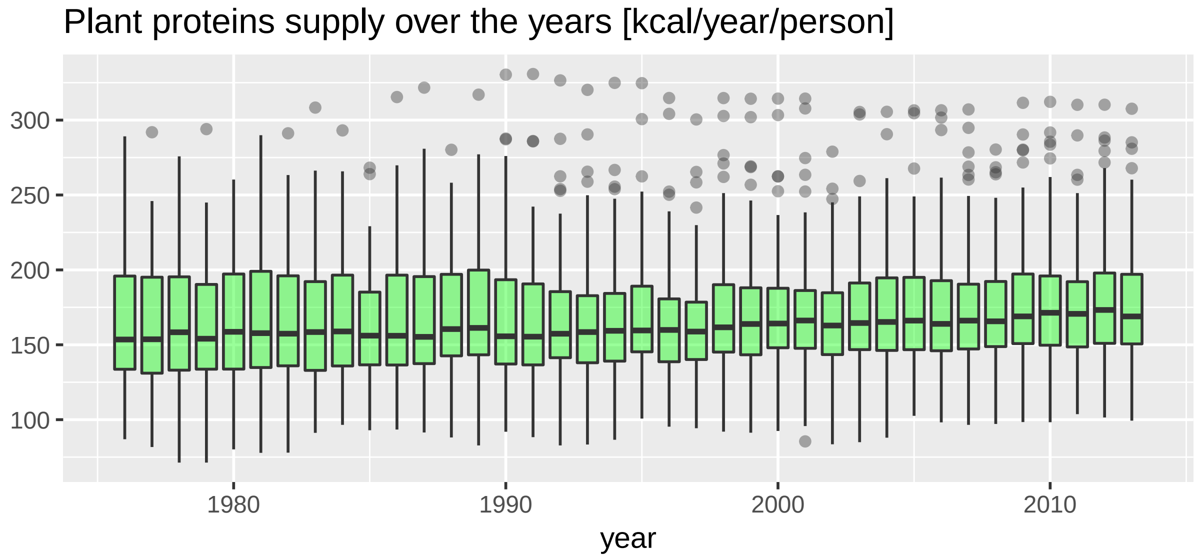
- last year obesity

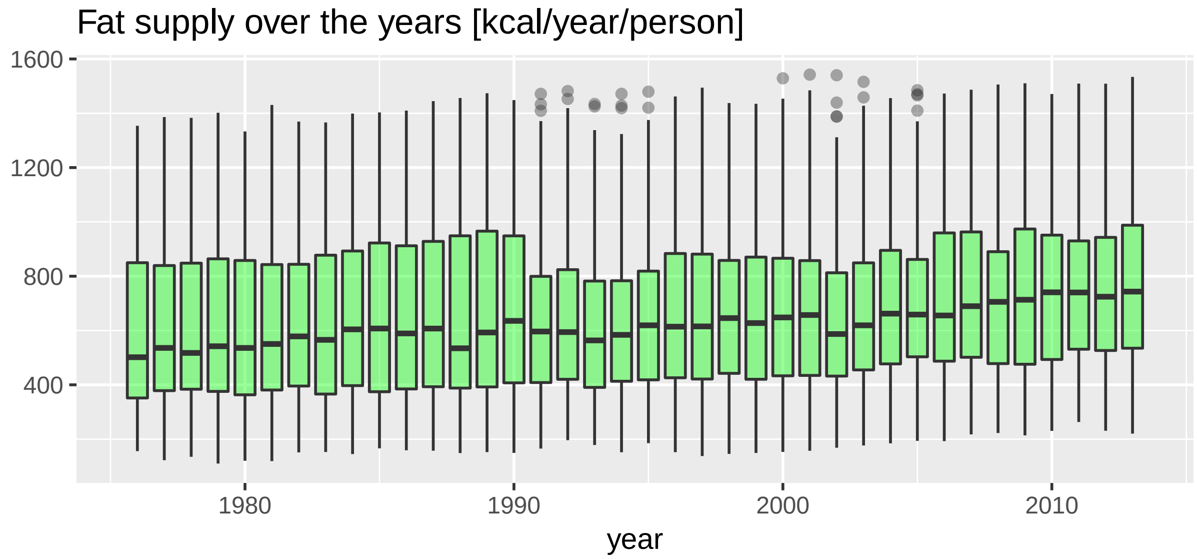
Yet, initial time-series starting in 1975, for almost all countries, had to be reduced since previous year data cannot be obtained for 1975. On the other hand filling up missing data, can cause biased results and criticism. Either when past obesity is zeroed or made to match 1975 levels. Limited loss of one point on the time-scale is acceptable. Auto-regressive enhancement will be included in modeling.

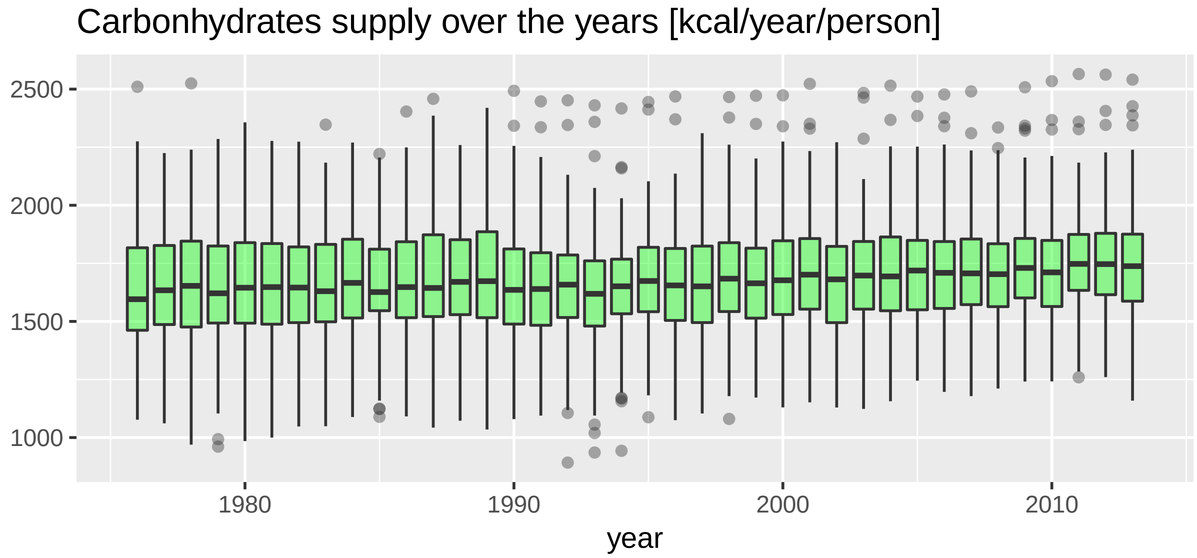
**How supply of macronutrients progressed over the years?**

One can note that not all macronutirents were in monotone supply.


Animal proteins supply over the years [kcal/year/person]


Plant proteins supply over the years [kcal/year/person]



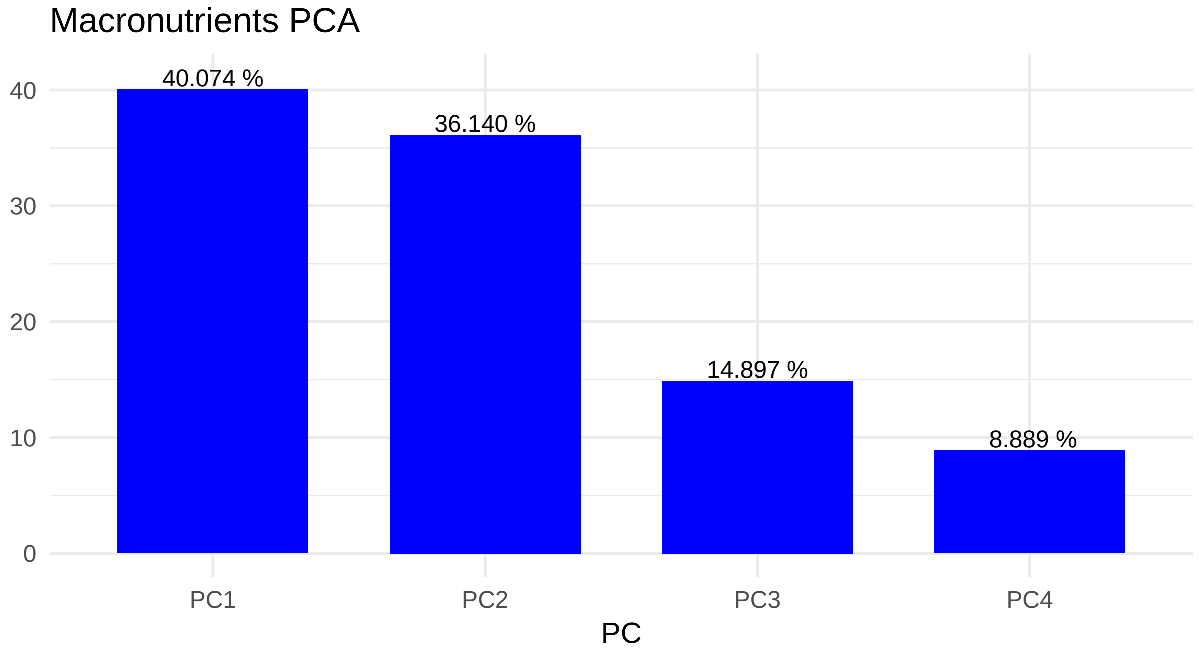


**How representative is macronutrients data?**

Looking at correlation of macronutrients we observe that heavy correlated pair does indeed exist. Fat and animal proteins supply expose colinearity.

| | animal proteins | plant proteins | fat | carbonhydrates |
|---|---|---|---|---|
| animal proteins | 1.00 | -0.1 | 0.89 | 0.2 |
| plant proteins | -0.10 | 1.0 | 0.00 | 0.7 |
| fat | 0.89 | 0.0 | 1.00 | 0.2 |
| carbonhydrates | 0.20 | 0.7 | 0.20 | 1.0 |

Principal Component Analysis suggests that 3 major components could simplify the features data set, but at the expense of portion of variability explained by macronutrient features. Had it been necessary, one could accept 92% variability.



However, compared to least obese countries, roughly 8% variability loss (PC4) could render explanatory capacity for those societies where few obesity share points are observed. Give that, it shouldn't be discarded easily for the sake of simplifying the features model, ie. reliance on first 3 PCs.

### 2.4 Regression methods

For the task of regression of obesity share among adults, onto macronutrients, few non-linear methods were selected:

- k-Nearest Neightbors,
- General Additive Model using Splines,
- General Additive Model using Loess,
- Gradient Boosting.

The last method, tree based modelling, was selected as the most advanced performer with the aim of greatly flexible tuning effort, yet the hope of outperforming its predecessors.

### Modelling variants

Linear representation of features used in regression study:

(a) macronutrients (supply in [kcal/year/per person])

$$y(i) = \beta_1 x_{plant.proteins}(i) + \beta_2 x_{animal.proteins}(i) + \beta_3 x_{fat}(i) + \beta_4 x_{carbonhydrates}(i)$$

(b) autoregressive obesity & macronutrients

Given discovered autocorrelation of dependant obesity variable, uses lagged variable.

$$y(i) = \beta_0 y(i-1) + \beta_1 x_{plant.proteins}(i) + \beta_2 x_{animal.proteins}(i) + \beta_3 x_{fat}(i) + \beta_4 x_{carbonhydrates}(i)$$

**Error function**

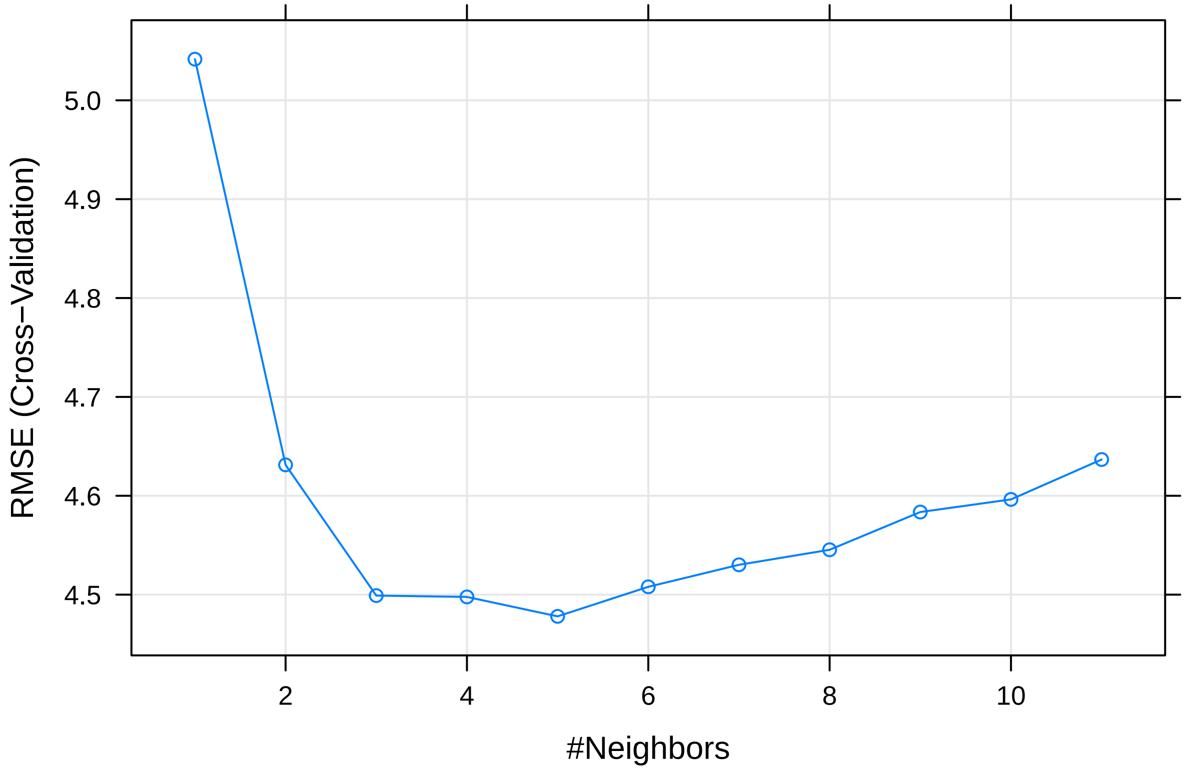
Standard error function of RMSE is used for accuracy assessment for all methods in question.

$$RMSE = \sqrt{\frac{1}{n}\sum_i (y(i) - \hat{y}(i))^2}$$

### 2.4.1 k-Nearest Neighbors method (knn)

**(a) features: macronutrients**

Variables comprise of macronutrients only. Below is the cross-validated training output where number of neighbors was limited within the range [1,11].



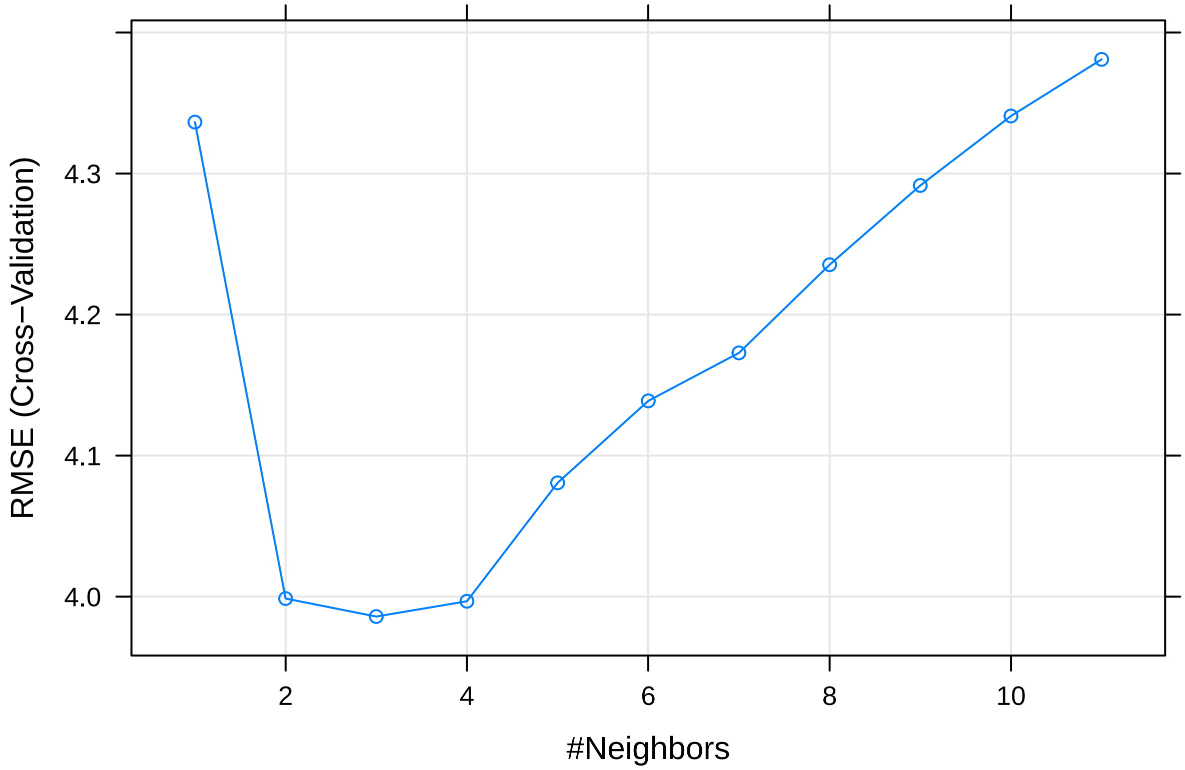
k-NN's best result is identified:

|  | k | RMSE | Rsquared | MAE | RMSESD | RsquaredSD | MAESD |
|---|---|---|---|---|---|---|---|
| Best fit | 5 | 4.478036 | 0.6622149 | 3.252623 | 0.1479049 | 0.0149455 | 0.1319095 |

Notably poor statistics of the fit (low Rˆ2 and high average error) render it unsatisfactory.

### (b) features: auto-regressive obesity & macronutrients

Here, macronutrient variables and lagged dependant variable are applied. Similarly 5-fold cross-validation is used in training.



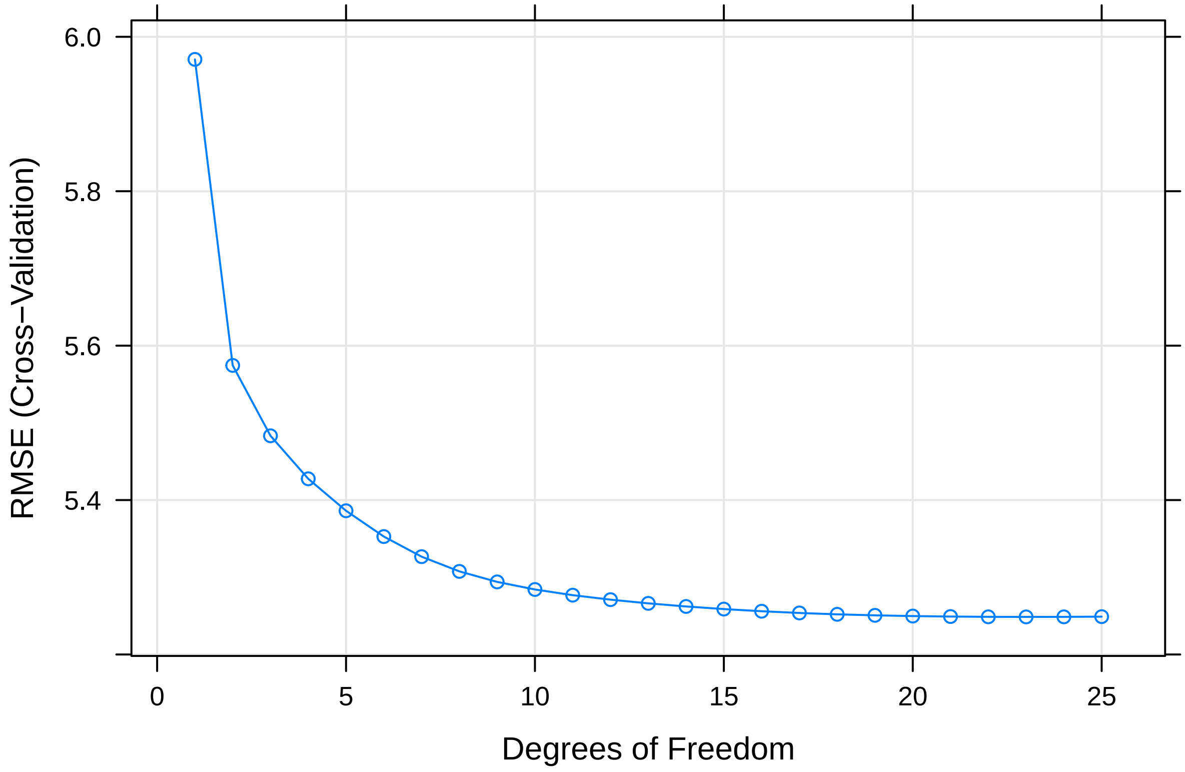
k-NN's best result is identified, with moderately improved fit (roughly 1% obesity share on average):

|  | k | RMSE | Rsquared | MAE | RMSESD | RsquaredSD | MAESD |
|---|---|---|---|---|---|---|---|
| Best fit | 3 | 3.985902 | 0.7333827 | 2.736543 | 0.1370495 | 0.0129346 | 0.1346445 |

### 2.4.2 Generalized Additive Model using Splines

### (a) features: macronutrients

Splines function for this type of task might seem an unusual attempt, to capture so much data in multiple dimensions. Cross-valided training is conducted by increasing degrees of freedom in range [1,31].



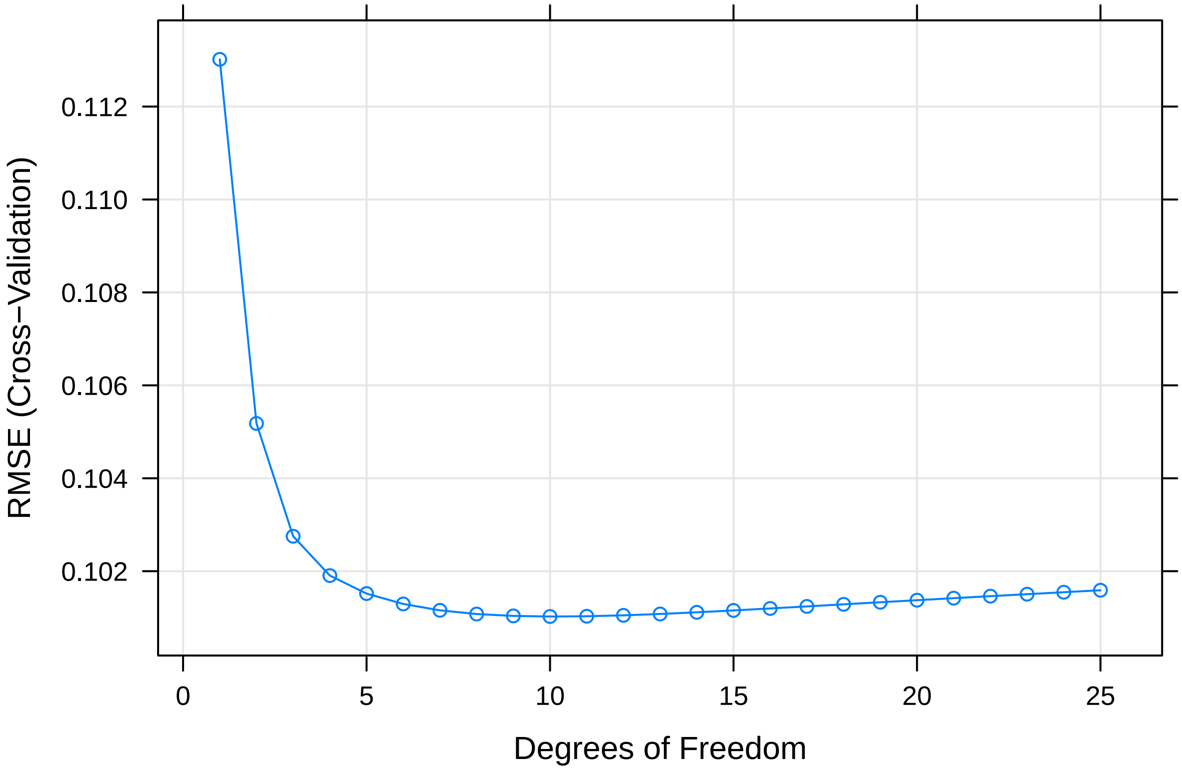
GAM splines best-fit is identified:

| | df | RMSE | Rsquared | MAE | RMSESD | RsquaredSD | MAESD |
|---|---|---|---|---|---|---|---|
| Best fit | 23 | 5.248652 | 0.534976 | 4.020683 | 0.2575897 | 0.0360316 | 0.1601781 |

### (b) features: auto-regressive obesity & macronutrients

Lagged dependant obesity variable is added to the set of independent variables. 5-fold cross-validation is conducted.

Splines function for this type of task might seem an unusual attempt, to capture so much data in multiple dimensions. Cross-valided training is conducted by increasing degrees of freedom in range [1,31].



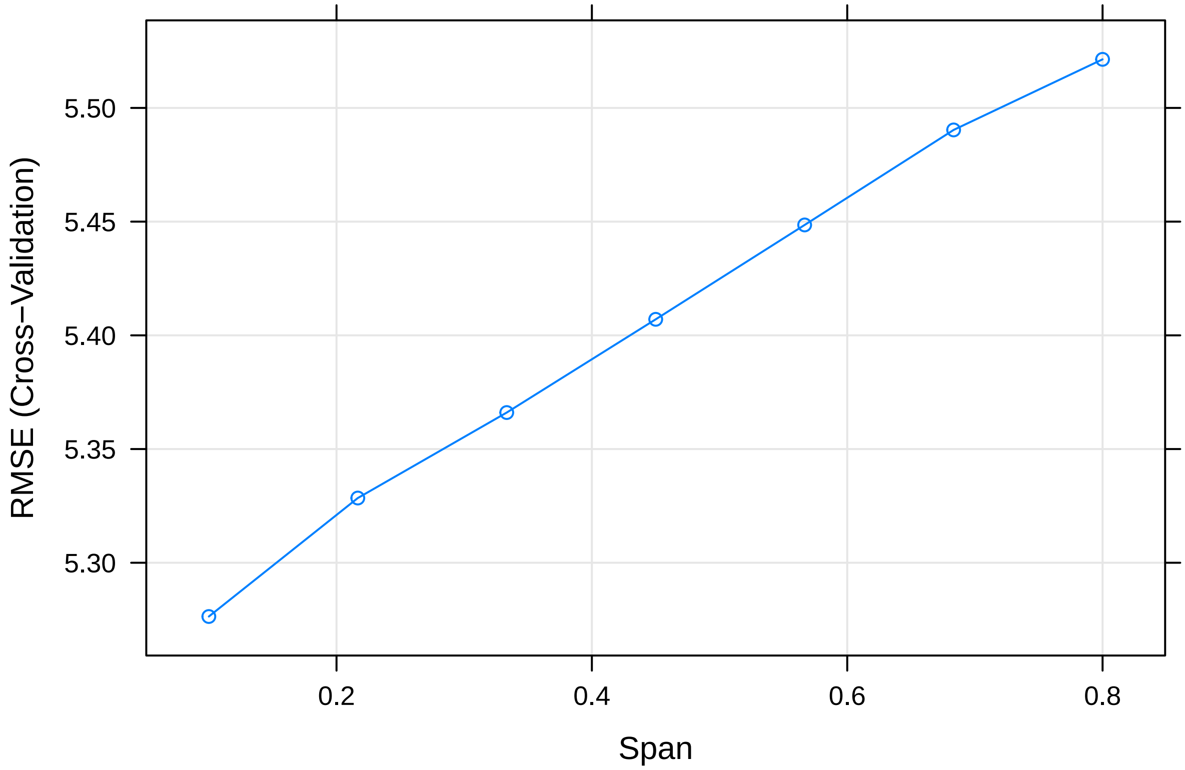
GAM splines best-fit is identified, which bring tremendous improvement of explanatory power (Rˆ2), and average error below 1 % share of adults.

| | df | RMSE | Rsquared | MAE | RMSESD | RsquaredSD | MAESD |
|---|---|---|---|---|---|---|---|
| Best fit | 10 | 0.1010245 | 0.9998269 | 0.0673018 | 0.0100512 | 3.34e-05 | 0.0021098 |

### 2.4.3 Generalized Additive Model using Loess

#### (a) features: macronutrients

Loess is a locally weighted regression method, making it more flexible than linear regression. Both degrees of freedom and span are costly to be calibrated. Unstable "gamLoess" implementation prevented and variations of degree parameters. Consequently degree was set to 1, as in the default configuration of the method. Important to note that larger span means smoother fit: - each neighborhood consisting between 10% and 80% of the observations



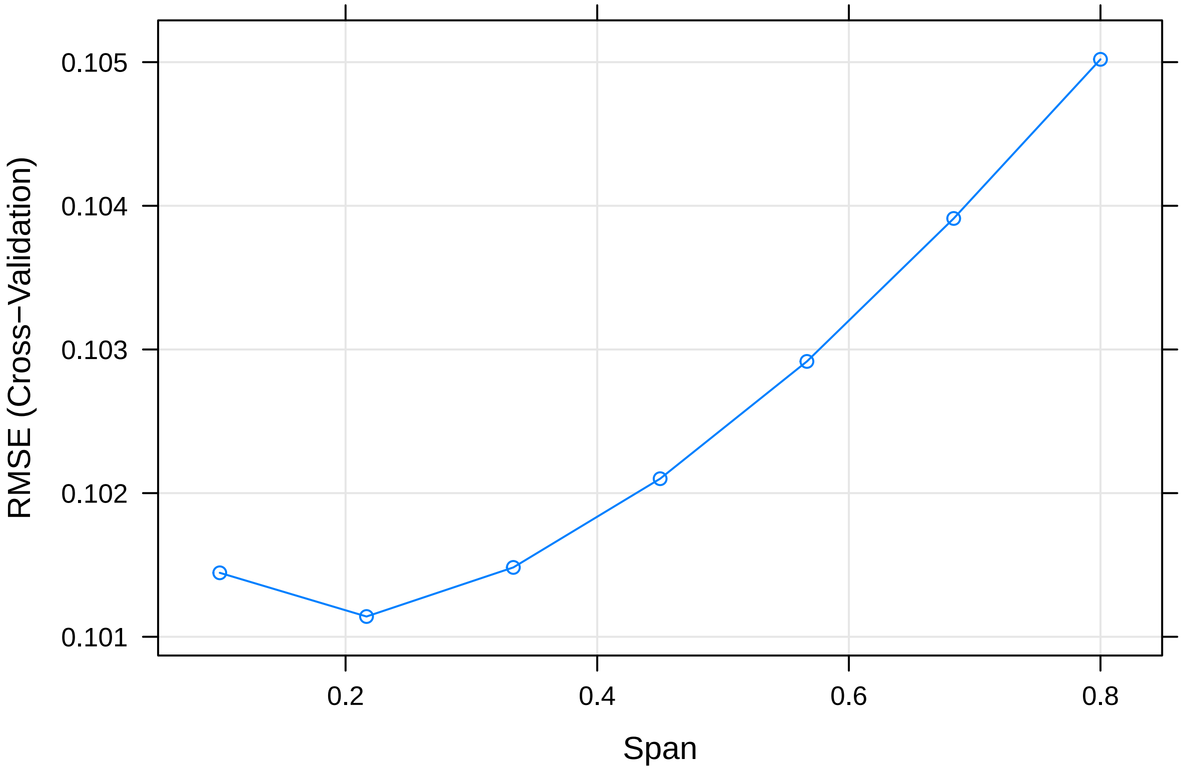
GAM loess best-fit is identified:

| | span | degree | RMSE | Rsquared | MAE | RMSESD | RsquaredSD | MAESD |
|---|---|---|---|---|---|---|---|---|
| Best fit | 0.1 | 1 | 5.276358 | 0.5299251 | 4.061266 | 0.2609941 | 0.0369757 | 0.1587851 |

That outcome is hardly explanatory, given the lowest Rˆ2 and very poor average error, which undermines applicability to countries where share of adults is less than the RMSE achieved by this method & features selected.

### (b) features: auto-regressive obesity & macronutrients

Here, the Loess method is provided the lagged obesity apart from all macronutrient variables. Similarly 5-fold cross-validation is applied.



GAM loess best-fit is identified:

| | span | degree | RMSE | Rsquared | MAE | RMSESD | RsquaredSD | MAESD |
|---|---|---|---|---|---|---|---|---|
| Best fit | 0.2166667 | 1 | 0.1011411 | 0.9998265 | 0.0674766 | 0.0100171 | 3.34e-05 | 0.0021287 |

### 2.4.4 Gradient Boosting Machines

Gradient boosted machines (GBMs) are powerful algorithms, building ensemble of shallow but weak successive trees, that learn from their predecessors. So each tree learns and improves from the previous tree. Such powerful ensemble carries similarity to random forest concept. However the latter relies on the independence of the deep trees that were built.

Boosting unlike bagging, improves the bias (often over-emphasizing outliers), which may lead to over-fitting. Notion of weakness in boosted learning comes from ability to always learn anything, with the outcome better than chance. Though, the memory and compution intensive with large number of trees (thousands), often finds itself superior against other methods.

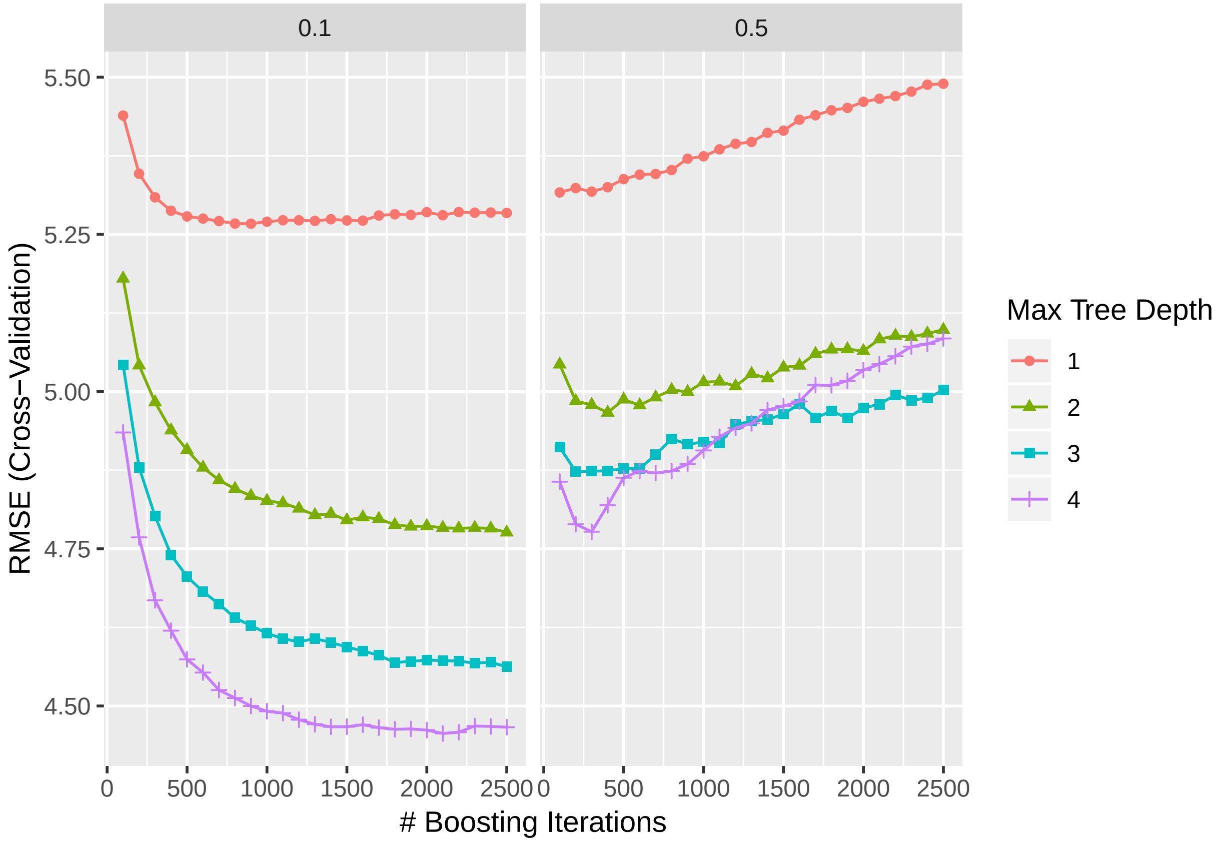
#### (a) features: macronutrients

Gradient boosting with Gaussian loss function was applied. Required tuning evolved around 3 parameters:

- interaction depth (beyond the additive interaction at depth 1),
- number of trees (empirically increased beyond 1000),

- shrinkage (learning rate, empirically narrowed to contrast results' evolution)

Rather sadly, the costly method is not thread-safe to permit parallel execution, and hence speeding up results recovery. This is particularly an impediment since cross-validation is used in training combined with wide range of tress and interactions.

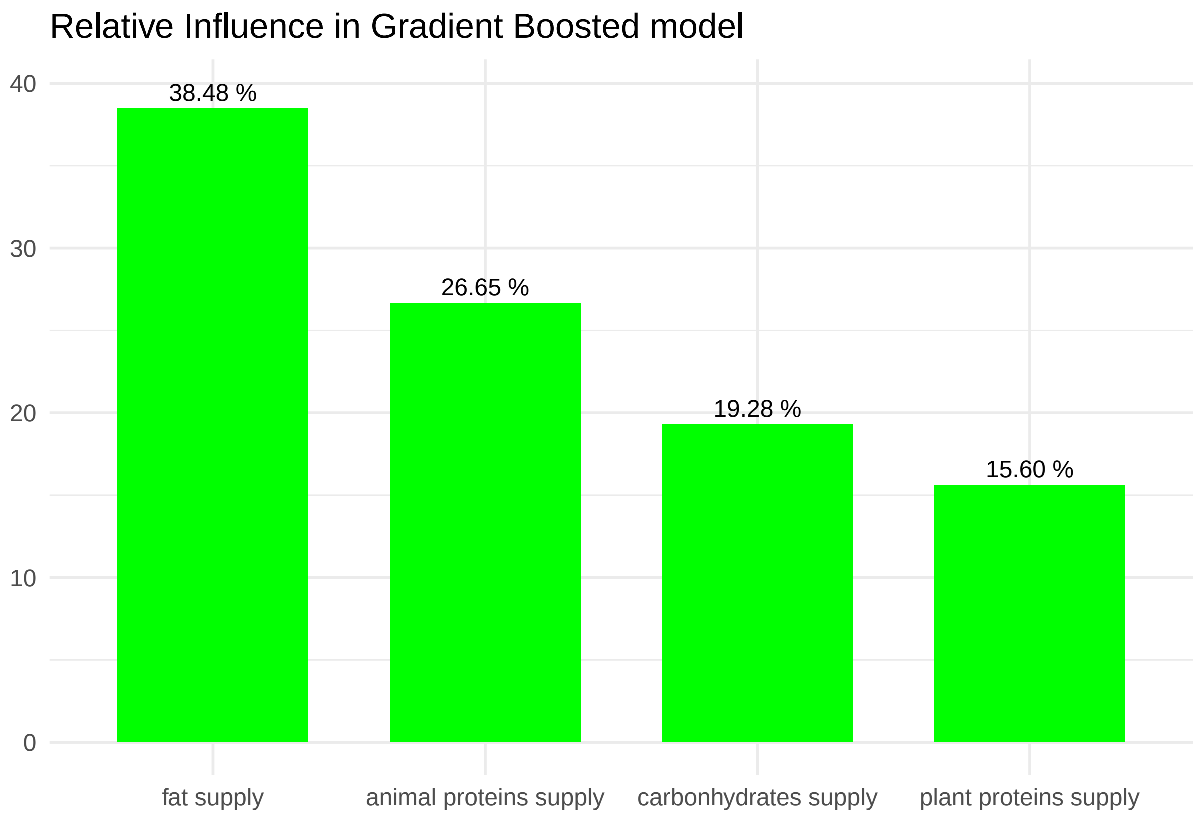


Boosted best-fit is identified with features' influence report:

|  | n.trees | interaction.depth | shrinkage | n.minobsinnode |
|---|---|---|---|---|
| Best fit | 2100 | 4 | 0.1 | 10 |

|  | RMSE | Rsquared | MAE | RMSESD | RsquaredSD |
|---|---|---|---|---|---|
| Best residuals | 4.456236 | 0.6657995 | 3.374119 | 0.1897722 | 0.0213544 |

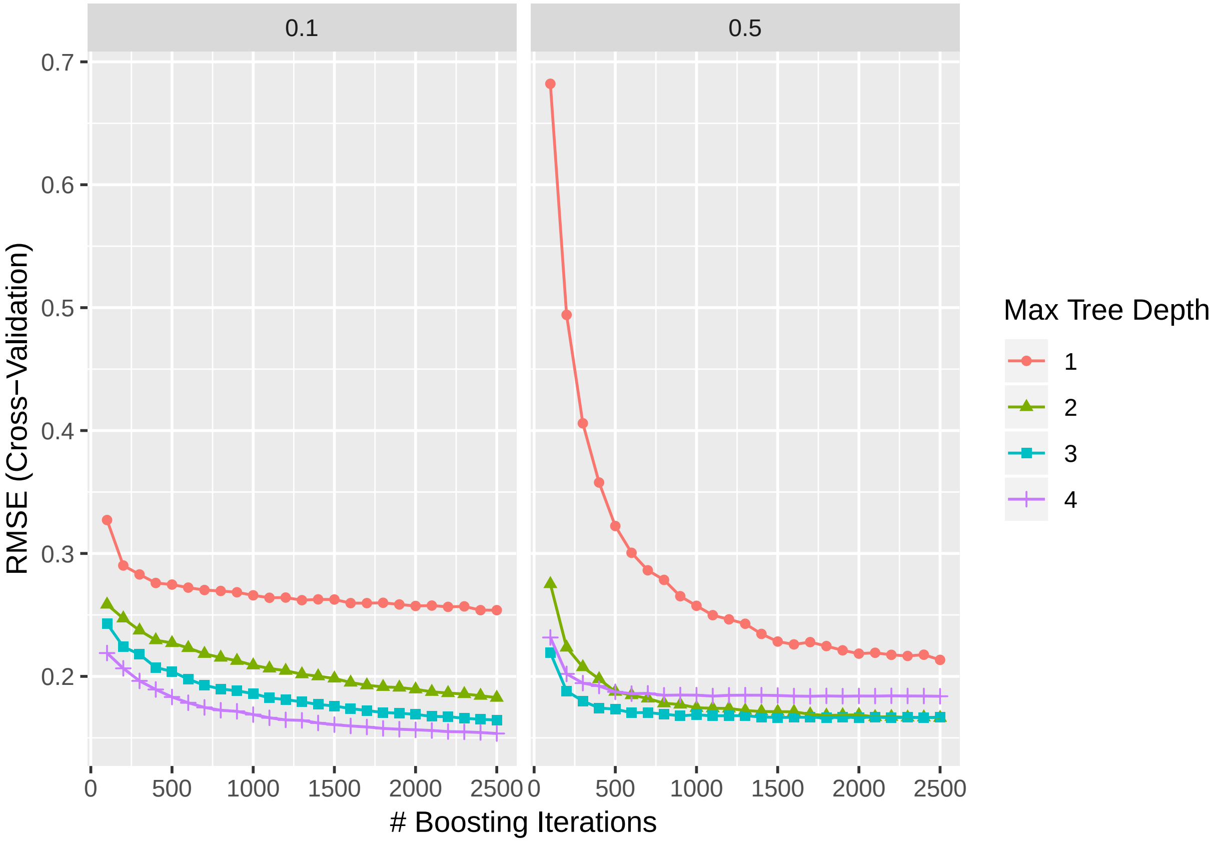
```
## n.trees not given. Using 2100 trees.
```



RMSE as well as R^2 are not impressive, overshadowing observed obesity share.

### (b) features: auto-regressive obesity & macronutrients

5-fold cross-validation is applied as before, 4 tuning parameters.

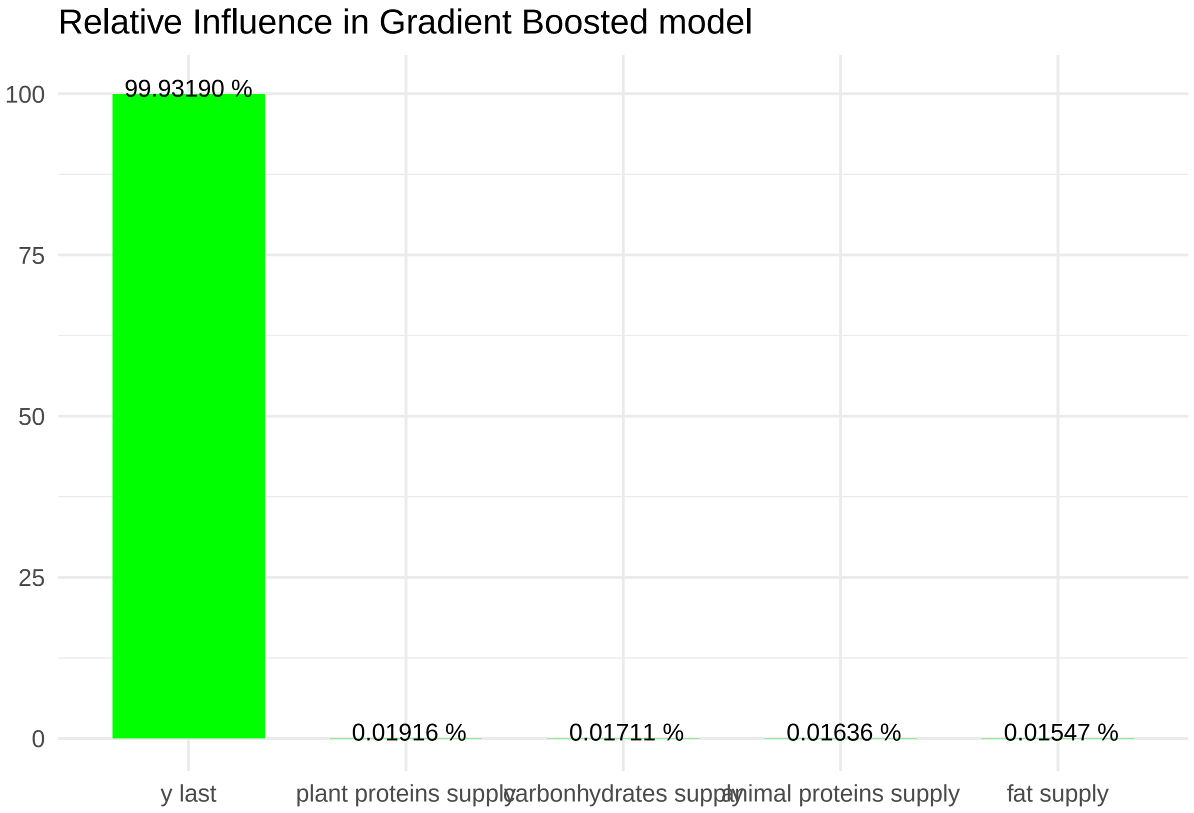


Boosted regression's best-fit parameters and residuals:

|  | n.trees | interaction.depth | shrinkage | n.minobsinnode |
|---|---|---|---|---|
| Best fit | 2500 | 4 | 0.1 | 10 |

|  | RMSE | Rsquared | MAE | RMSESD | RsquaredSD |
|---|---|---|---|---|---|
| Best residuals | 0.1535116 | 0.9996012 | 0.0929578 | 0.0215434 | 0.0001019 |

```
## n.trees not given. Using 2500 trees.
```



Once again, as for other methods, when auto-regressive model construction is employed, the residuals and explanatory power improve tremendously.

### Extended modelling variants

Linear representation of features used in regression study:

(c) macronutrients (supply in [kcal/year/per person])

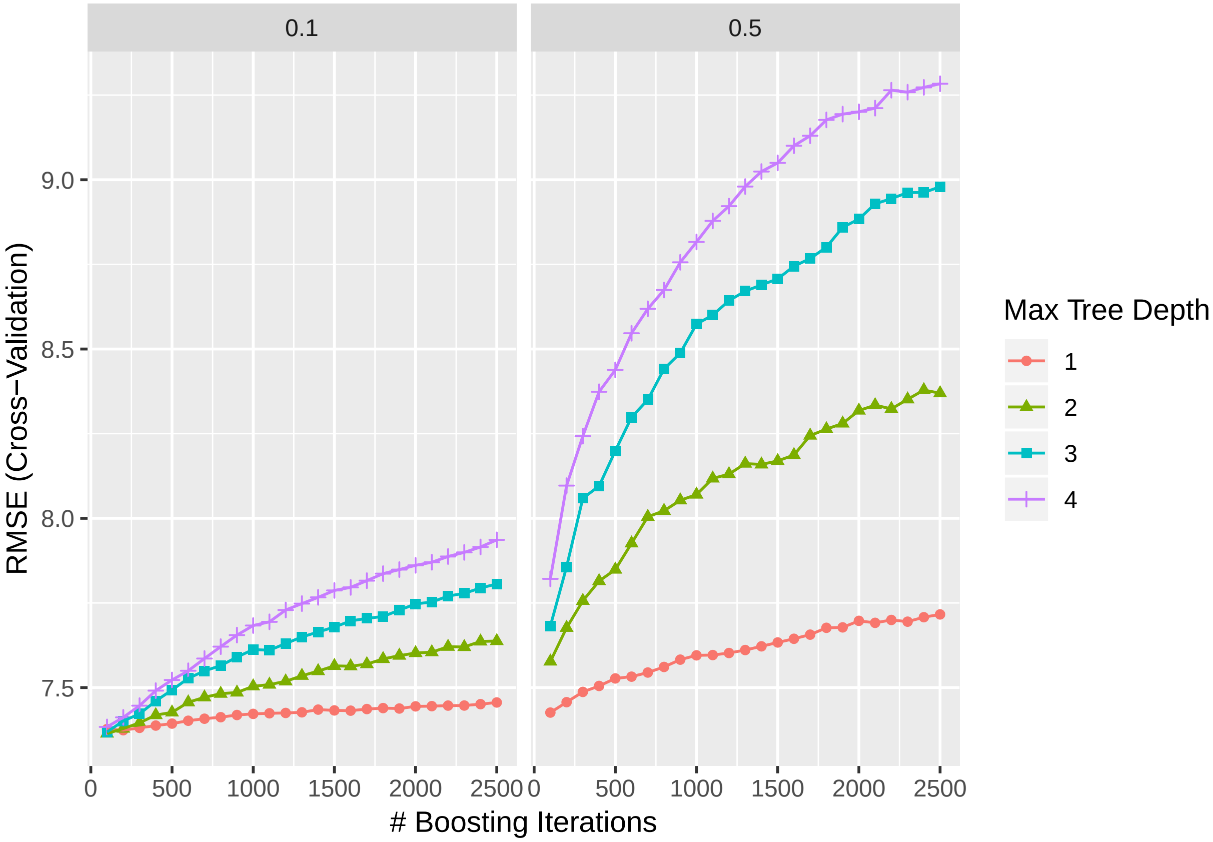
$$y(i) = \beta_1 \Delta x_{plant.proteins}(i) + \beta_2 \Delta x_{animal.proteins}(i) + \beta_3 \Delta x_{fat}(i) + \beta_4 \Delta x_{carbonhydrates}(i)$$

(d) autoregressive obesity & macronutrients

Given discovered autocorrelation of dependant obesity variable, uses lagged variable.

$$y(i) = \beta_0 y(i-1) + \beta_1 \Delta x_{plant.proteins}(i) + \beta_2 \Delta x_{animal.proteins}(i) + \beta_3 \Delta x_{fat}(i) + \beta_4 \Delta x_{carbonhydrates}(i)$$

### (c) features: change in macronutrients

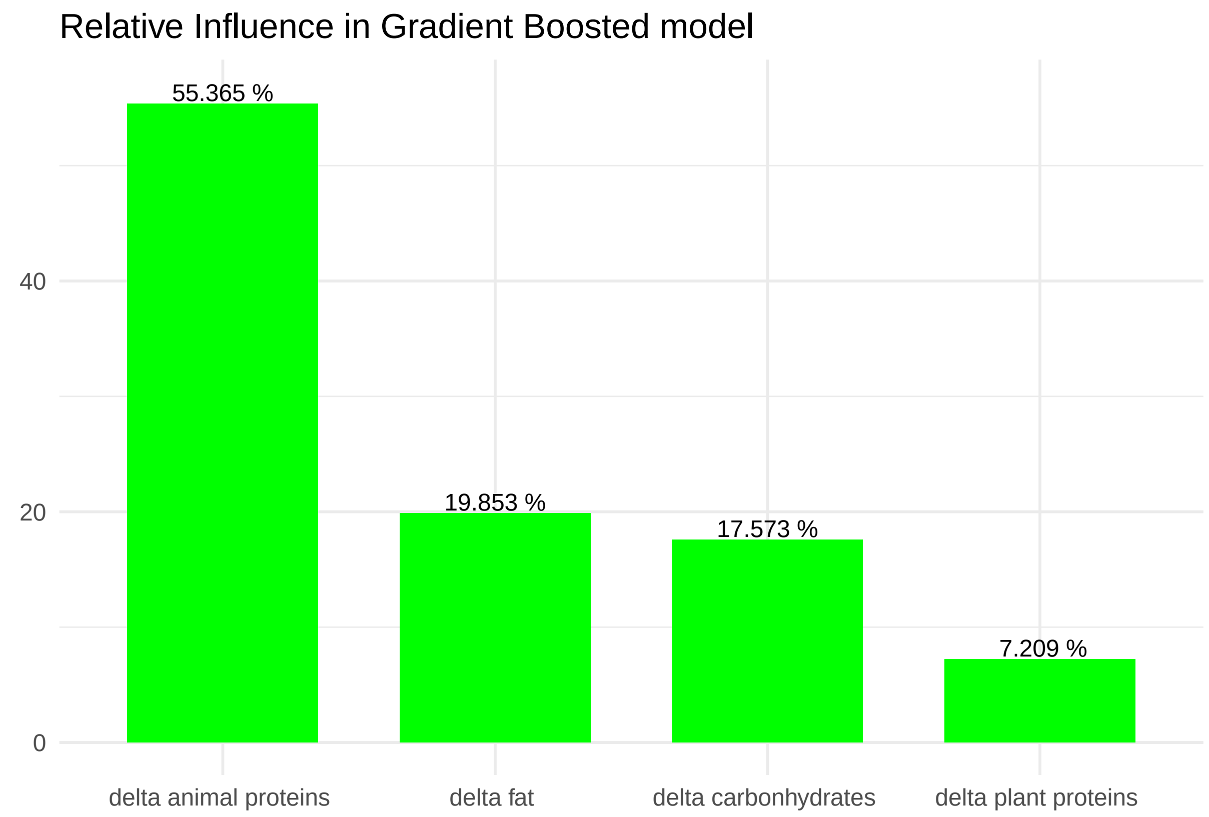


Boosted best-fit is identified:

| | n.trees | interaction.depth | shrinkage | n.minobsinnode |
|---|---|---|---|---|
| Best fit | 100 | 2 | 0.1 | 10 |

| | RMSE | Rsquared | MAE | RMSESD | RsquaredSD |
|---|---|---|---|---|---|
| Best residuals | 7.364376 | 0.0849641 | 6.088954 | 0.0893414 | 0.0151016 |

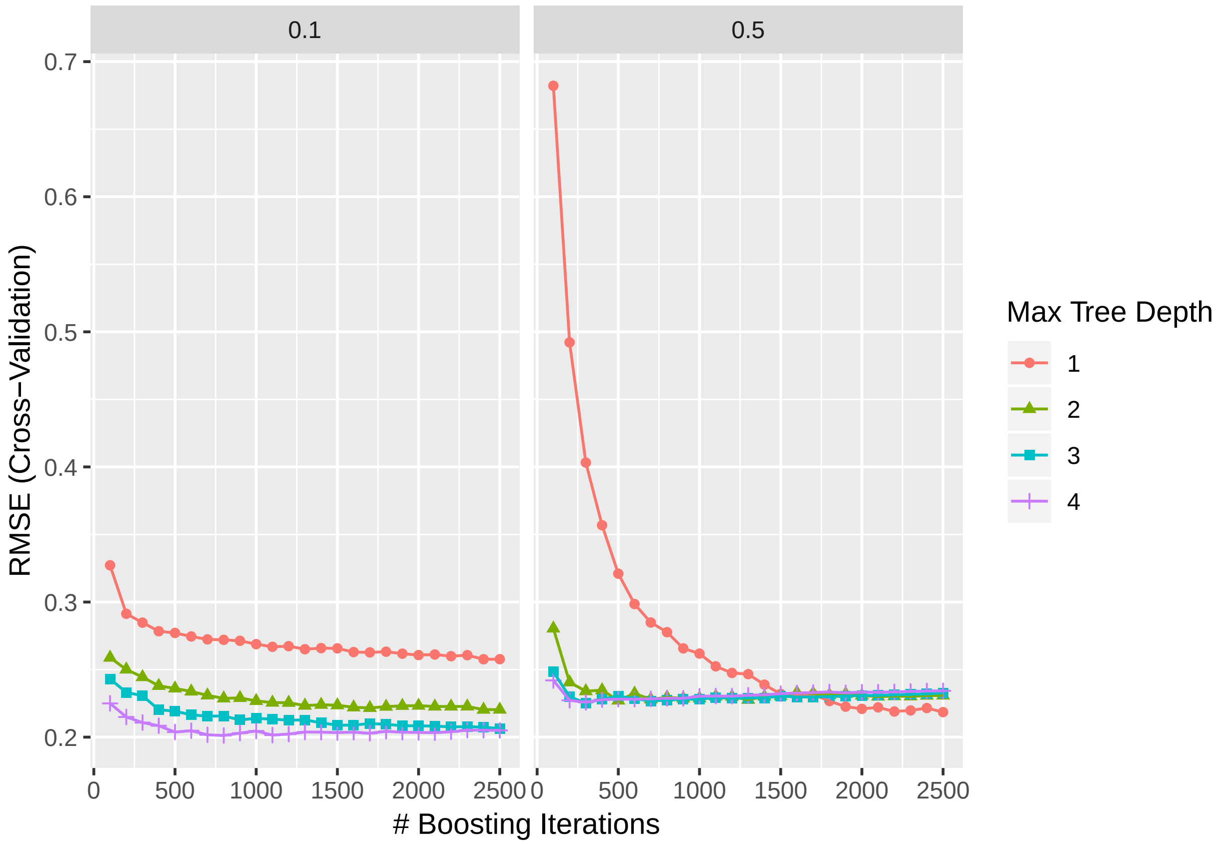
```
## n.trees not given. Using 100 trees.
```



RMSE as well as Rˆ2 are not impressive, overshadowing observed obesity share.

### (d) features: auto-regressive obesity & change in macronutrients

5-fold cross-validation is applied as before with 4 tuning parameters.

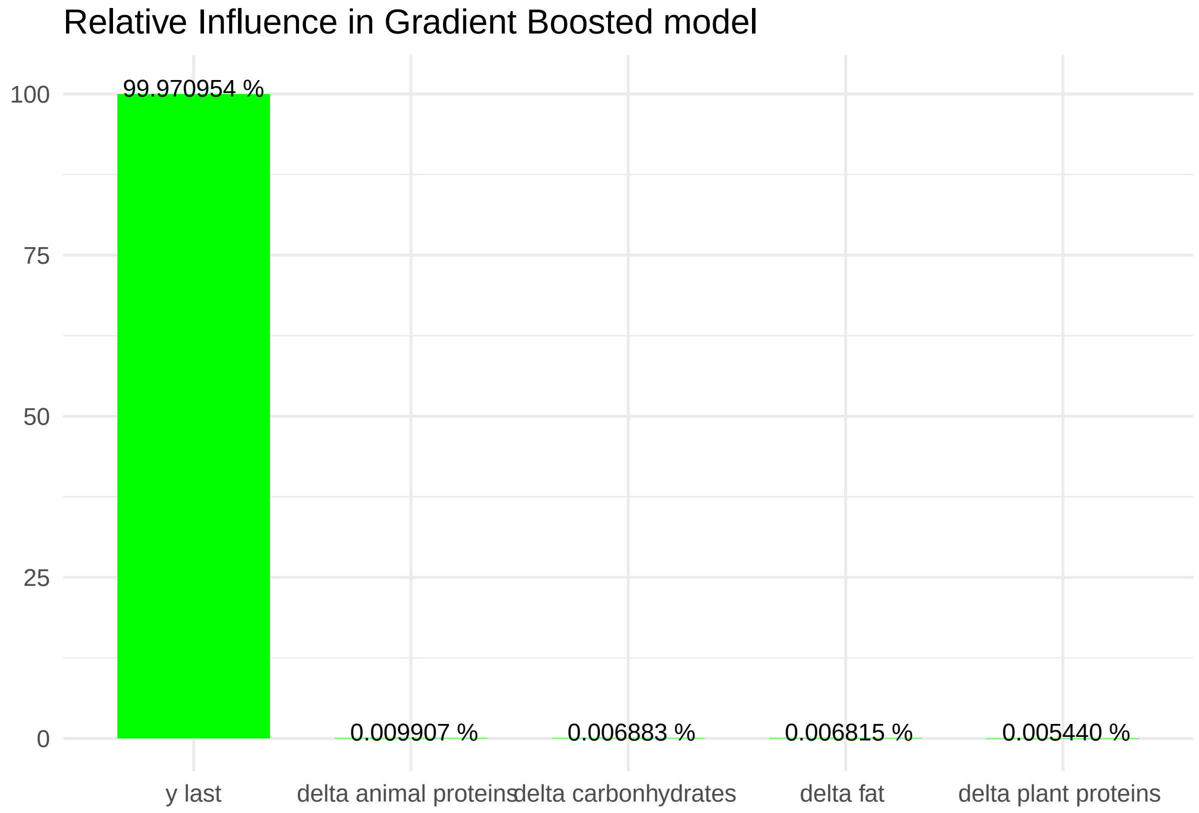


Boosted regression's best-fit parameters and residuals:

| | n.trees | interaction.depth | shrinkage | n.minobsinnode |
|---|---|---|---|---|
| Best fit | 800 | 4 | 0.1 | 10 |

| | RMSE | Rsquared | MAE | RMSESD | RsquaredSD |
|---|---|---|---|---|---|
| Best residuals | 0.2011893 | 0.9993041 | 0.1132786 | 0.0393743 | 0.0002652 |

```
## n.trees not given. Using 800 trees.
```



## 3. Results

Models comparison

Various methods and feature models are tested now, pretrained using cross-validation across the board. Their performance out-of-sample is compared with training set RMSE figures.

| Method | Features | RMSE(test) | RMSE(training) | % Change |
|---|---|---|---|---|
| kNN | macronutrients | 4.3295778 | 3.5018817 | 23.635752 |
| kNN | AR(1) & macronutrients | 3.6833285 | 2.5183697 | 46.258448 |
| GAM with Splines | macronutrients | 5.3450721 | 5.1483295 | 3.821484 |
| GAM with Splines | AR(1) & macronutrients | 0.1112117 | 0.1002895 | 10.890746 |
| GAM with Loess | macronutrients | 5.3621411 | 5.1957797 | 3.201857 |
| GAM with Loess | AR(1) & macronutrients | 0.1115616 | 0.1004656 | 11.044583 |
| GBM | macronutrients | 4.3443421 | 2.7839846 | 56.047631 |
| GBM | AR(1) & macronutrients | 0.1659730 | 0.0688965 | 140.901773 |
| GBM | macronutrients change | 7.2931745 | 7.2208529 | 1.001565 |
| GBM | AR(1) & macronutrients change | 0.2031050 | 0.1243121 | 63.383218 |

General Additive Models, both using Splines and Loess, appear to provide the best results among all the methods:

- out-of-sample perfromance was reduced by around 11% only (RMSE),

- training set and test set RMSE was below one point of the dependant variable (obesity share).

There was an attempt to contrast regression methods with different feature models:

- auto-regressive model of obesity proved to outperform the others

Gradient Boosted Machines haven't produced the expected results:

- RMSEs were comparable with GAM modeling (below 1 point at best),
- yet performance on the test set was down by 140% and 63% when auto-regressive obesity was tested.

## 4. Conclusions

Summary:

- buidling global model of obesity share proved to be a challenging task,
- selection of macronutrient features and their form had little impact (supply vs 1st difference change),
- performance of the methods was important, but of secondary importance,
- fundamental inclusion of lagged obesity made the biggest impact on out-of-sample performance,
- k-NN method when applied to non-classification task performed poorly, compared with GAM's and GBM's type methods,
- Stochastic Gradient Boosting, being prone of over-fitting experienced big drop in performance on test set.

Limitations of the study:

- limited number of features, mainly macronutrients,
- country specific effects were not tested (hindered by small sample availability per country),
- available CPU power didn't permit employment of memory and CPU intensive methods (Neural Networks),
- no residuals study was conducted for each of the methods/model pairs of research.

Potential for the future:

- obesity research conducted with macronutrients exclusively is not satisfactory, further exploration is needed,
- different number of features should be mined and explored (eg. economic variables, demographic structure, and actual food products),
- residuals analysis could be employed: heterscedasticity and serial correlation testing, during model development,
- due to small sample size per country (<100) and total sample size (6k) bootstapping outght to be considered,
- future research can be conducted with limited number of methods, given the relative success of General Additive Modeling in this study.