Home-Assignment

This is an individual home assignments. Copied or identical submissions receive zero points. The home assignment (R-code and PDF) has to be submitted until March 31, 2020, by email to anthony.strittmatter@unibas.ch.

The Tesco Grocery 1.0 dataset records 420 million food items purchased by 1.6 million fidelity card owners who shopped at the 411 Tesco stores in Greater London in the year 2015. To preserve anonymity, the data is aggregated at the Lower layer Super Output Area (LSOA) census areas. The data contain the variables:

- `area_id`: LSOA area identifier.
- `area_name`: Name of LSOA area.
- `alcohol`: Grams of alcohol in the area's average product.
- `weight`: Weight of the area's average food product, in grams.
- `volume`: Volume of the area's average drink product, in liters.
- `energy`: Nutritional energy of the area's average product, in kcals.
- `energy_density`: Concentration of calories in the area's average product, in kcals/gram.
- {`nutrient`}: Weight of {nutrient} in the area's average product, in grams. Possible nutrients are: carbs, sugar, fat, saturated fat, protein, fibre. The count of carbs include sugars and the count of fats includes saturated fats.
- `energy_`{`nutrient`}: Amount of energy from {nutrient} in the area's average product, in kcals.
- `f_`{`category`}: Fraction of products of type {category} purchased. Possible categories are: dairy, eggs, fats and oils, fish, fruit and vegetables, grains, red meat, poultry, readymade, sauces, soft drinks, sweets, tea and coffee, and water.
- `f_`{`category`}`_weight`: Fraction of total product weight given by products of type {category}.
- `num_transactions`: Total number of products purchased by fidelity card owners who are resident in the area.
- `man_day`: Cumulative number of man-days of purchase (number of distinct days a customer has purchased something, summed all individual customers).

Where applicable, measures are accompanied with their standard deviation (variables with suffix `_std`).

The data set `tesco.csv` is enriched with population data coming from the census (source: Office for National Statistics). In particular, the additional variables are:

- `mortality_2016`: Mortality rate (in %) among residents aged 65 years or more (census 2016).
- `population`: Total population of residents in the area (census 2015).
- `male`: Total male population in the area (census 2015).
- `female`: Total female population in the area (census 2015).
- `age_0_17`: Total number of residents between 0 and 17 years old (census 2015).
- `age_18_64`: Total number of residents between 18 and 64 years old (census 2015).
- `age_65`: Total number of residents aged 65 years or more (census 2015).
- `avg_age`: Average age of residents (census 2015).
- `area_sq_km`: Surface of the area in km^2 (census 2015).
- `people_per_sq_km`: Population density per km^2 (census 2015).

Exercises:

1. Load the `tesco.csv` data. How large is the correlation between `alcohol` and `mortality_2016`?

2. Generate a training and test sample. Why do we create training and test sample and do not estimate everything in the same data?

3. Predict the `mortality_2016` using a Lasso and Ridge estimator. What are the tuning parameters? How do you select the tuning parameters? Has the Lasso or Ridge better prediction power? How do the Lasso and Ridge estimators deal with very highly correlated variables?

4. Predict the `mortality_2016` using a Principal Component Regression (PCR). For this purpose, estimate the principal components and use them as regressor to predict `mortality_2016` in a Lasso approach. What is the main intuition of a Principal Component Analysis? How many principal components are in the data? Does the Lasso select the principal components in any systematic way? Has the PCR approach a better prediction power than the Lasso or Ridge?

5. Predict the `mortality_2016` using a Random Forest estimator. What are the tuning parameters? How do you specify the tuning parameters? Has the Random Forest estimator a better prediction power than Lasso, Ridge, or PCR? What are the advantages and disadvantages of the Random Forest estimator compared to the Lasso and Ridge estimators?

6. Plot the variable importance measure of the Random Forest estimator. How do we have to interpret the variable importance measure? Are the same variables for the Random Forest and Lasso important? What does this imply?

7. Use the Double Selection Procedure to estimate the causal effect of `alcohol` on `mortality_2016`. What are the main identifying assumptions (under a selection-on-observables strategy)? Are they credible in this application? How large is the causal effect? Is the causal effects significantly different from zero? What is the difference in the interpretation of the `alcohol` parameter in the prediction and causal model? Make an example of a naive (not identified) estimation approach for causal effects.