Code ▾

Data and Code

Several datasets and some supporting code will be used in the short course. They are below.

Datasets

The following datasets are used as examples.

DTI Data

The data description below draws heavily from Goldsmith et al (http://www.tandfonline.com/doi/abs/10.1198/jcgs.2010.10007) (2011):

Diffusion tensor imaging (DTI) tractography is a magnetic resonance imaging (MRI) technique that allows the study of white-matter tracts by measuring the diffusivity of water in the brain: in white-matter tracts, water diffuses anisotropically in the direction of the tract, while elsewhere water diffuses isotropically. Using measurements of diffusivity along several gradients, DTI can provide relatively detailed images of white-matter anatomy in the brain.

For each white-matter tract, DTI provides us several measures describing the diffusivity of water. One example of these measures is parallel diffusivity, which is the diffusivity along the principal axis of the tract. Parallel diffusivity is recorded at many locations along the tract, so that for each tract we have a continuous profile or function.

Several tracts are included in the DTI dataset, and each tract has profiles for several measures. Also included are demographic variables, results for cognitive tests, and indicators of disease status (multiple sclerosis vs healthy control). The data can be downloaded here (./DataCode/DTI.RDA).

Sugar Data

The data description below comes from Gertheiss et al (http://www4.stat.ncsu.edu/~staicu/papers/msFGRPL_R1.pdf) (2013):

In chemometrics there are often function-like absorbance or emission spectra given – in particular for food samples – that are used to determine the content of certain ingredients. Using the spectra is typically much cheaper than alternative chemical analysis.

268 samples of sugar were dissolved and the solution was measured spectrofluorometrically. For every sample the emission spectra from 275–560 nm were measured in 0.5 nm intervals (i.e., at 571 wavelengths) at seven excitation wavelengths: 230, 240, 255, 290, 305, 325, and 340 nm. In addition, there are laboratory determinations of the quality of the sugar given, such as ash content (in percentage). Ash content measures the amount of inorganic impurities in the refined sugar, cf. Bro (1999). The aim of the analysis is to study the association between the ash content and the fluorescence spectra.

The original data can be downloaded here (http://www.models.life.ku.dk/sugar_process), and a processed version is available here (./DataCode/Sugar.RDA).

Activity Data

The data description below draws heavily from Goldsmith et al (http://jeffgoldsmith.com/Downloads/HeadStart.pdf) (2016):

Accelerometers have become an appealing alternative to self-report techniques for studying physical activity in observational studies and clinical trials, largely because of their relative objectivity. During observation periods, the devices measure electrical signals that are a proxy for acceleration. ``Activity counts" are then devised by summarizing the voltage signals across a short period known as an epoch; one-minute epochs are common. Because accelerometers can be worn comfortably and unobtrusively, they produce around-the-clock observations of many kinds of activity.

Study participants were recruited from 50 Head Start centers in northern Manhattan, the Bronx, and Brooklyn, in neighborhoods with high rates of pediatric asthma. We used a survey instrument to collect data on the child’s age, race, sex, asthma symptoms and other medical conditions, birth order and family-related factors, and features of the home environment. Field staff measured the child's height, weight, and skin-fold thicknesses The staff then attached the accelerometer to the child's non-dominant wrist with a hospital band. We obtain $y_i(t)$ by averaging, for each t separately, across the six days of observation for each child. Additionally, we aggregate into 10 minute epochs.

For data confidentiality reasons, we will work with a simulated dataset. These data are simulated to mimic real data: they use realistic covariates, coefficients based on a fitted model, and residuals based on those in a full analysis. The activity data can be found here (./DataCode/Activity.RDA).

Additional code

In the limited time of the short course, we won't have time to carefully examine all available options for fitting functional regression models with variable selection. Below are some additional resources that may be of interest:

- The methods for generalized scalar-on-function regression found in Gertheiss et al (http://www4.stat.ncsu.edu/~staicu/papers/msFGRPL_R1.pdf) (2013) are implemented here (./DataCode/Gertheiss.zip). The zip file includes code to analyze the DTI and Sugar datasets.
- The methods for function-on-scalar regression using the LASSO found in Barber et al (https://arxiv.org/pdf/1610.07403.pdf) and in subsequent work are implemented here (./DataCode/Reimherr.zip).