Statistics 347, HW 8, due March 12

Discussion of homework problems among students is encouraged. However, all material handed in for credit must be your own work.

Please hand in each problem in a separate file with your name on it.

---

1. [Smoothing splines]

   Suppose that $N > 2$ and g is a natural cubic spline interpolant to the pairs $x_n, z_n, n = 1, \ldots, N$, with $a < x_1 < \ldots < x_N < b$. Namely g is a natural cubic spline with knots at each x_n and $g(x_i) = z_i$. Explain why such a function exists. (Assume the design matrix $N_{ij} = N_j(x_i)$ is invertible.) Let $\tilde{g}$ be any other twice differentiable function on $[a, b]$ that interpolates these N pairs.

   (a) Let $h(x) = g(x) - \tilde{g}(x)$. Use the boundary conditions on g and integration by parts to show that

   $$\int_a^b g''(x)h''(x)dx = -\sum_{n=1}^{N-1} g'''(x_n^+)[h(x_{n+1}) - h(x_n)], = 0$$

   and that therefore this expression is zero.

   (b) Conclude that

   $$\int_a^b \tilde{g}''(x)^2 dx \geq \int_a^b g''(x)^2 dx,$$

   and that equality is only possible when h is identically zero on $[a, b]$.

   (c) Let $\mathcal{F}$ be the space of functions with continuous second derivatives on $[a, b]$. Consider the penalized least squares problem

   $$\min_{f \in \mathcal{F}} \left[\sum_{n=1}^{N} (Y_n - f(X_n))^2 + \lambda \int_a^b f''(x)^2 dx \right],$$

   Show that the minimizer must be a natural cubic spline with knots at the points $x_n, n = 1, \ldots, N$.

2. The South African Heart Disease dataset 'SAheart.data' includes data from three areas of South Africa and is intended to study heart disease factors. The data include white males between the age of 15 and 64 and the response variable 'chd' is the presence or absence of myocardial infarction. There are 160 cases and 302 controls. You can ignore the variables 'adiposity' and 'typea'. The remaining variables are 'sbp' - systolic blood pressure, 'tobacco' - cumulative tobacco in kg, 'famhist' - family history of heart disease (Present, Absent), 'obesity' - obesity level, 'alcohol' - current alcohol consumption, 'age', 'ldl'.

   (a) Perform a logistic regression on the seven predictors and discuss the results. Can you identify a simpler model than the full model that fits the data well?

   (b) We now want to try a non-parametric fit of the dependence on the predictors. Use the function ns (for example ns(SAheart$age,df=4)) in R to produce a natural spline basis with 3 interior knots and two boundary knots for each of the continuous predictors. The result with be an $N \times 4$ matrix, with the y-value of each of 4 b-splines at each of the x-values of the predictor. If you want to see the splines you can't simply plot the columns of the matrix, you need to first order the predictor and use the same ordering on the column. Perform a logistic regression with this larger set of predictors. How many are there in total? Compare this fit to the simple linear fit? Is it significant? Investigate possible simplifications of the model by eliminating blocks of splines corresponding to a given original variable. Plot the function $f(x_i)$ for each of the original variables in the model (again remember to use the ordered variable). How does the dependence on the variables differ from the linear case.