Statistics 232A: Lab #6

Fall 2013 R. Beran

DUE in Discussion on 19 November 2013. Use class techniques and results to do the project. Visit `www.stat.ucdavis.edu/~beran/s232a/motor.dat` for the data. Attach your computer code to your report.

The file `motor.dat`, adapted from Silverman (1985), reports $n = 133$ observations of motorcycle acceleration against time in a simulated motorcycle accident. The $p = 277$ possible observation times constitute the vector $t = (1, 2, \ldots, 277)$. Accelerations were observed at only $q < p$ of these equally spaced times, sometimes with replication.

A linear model for this incomplete, unbalanced, one-way layout is $y = Xm + e$. Here y is the $n \times 1$ vector of accelerations recorded and e is the $n \times 1$ vector of experimental errors. Moreover, $m = (m_1, m_2, \ldots, m_{277})'$ denotes the unknown mean accelerations at the times from 1 to 277 while X is the design matrix that maps selected components of m into the observed accelerations. The components of e are assumed to be independent, identically distributed random variables, with mean 0 and unknown variance σ^2.

The problem is to estimate $\eta = \mathrm{E}(y) = Xm$ and the entire vector m.

a) Construct the design matrix X for the motorcycle data. Report numerically its row and column dimensions and its rank. Justify mathematically the claimed rank. Report the vector that gives the *number* of acceleration observations made at the successive times in the vector t.

b) Report the q-dimensional subvector t_{obs} of times in t at which one or more accelerations were observed. State the numerical value of q. Report too the complementary subvector t_{miss} of times in t at which *no* accelerations were observed.

c) State a valid formula for the least squares estimator of $\eta = Xm$ and apply it to the data. Thereby, compute and report the associated least squares estimate $\hat{\sigma}^2_{ls}$ of σ^2.

Let D_2 denote the *second-difference* matrix with p columns. Let $\hat{m}_{pls}(\lambda)$ be *any* value of $m \in R^p$ that minimizes the *penalized least squares* (PLS) criterion $|y - Xm|^2 + \lambda|D_2 m|^2$, where $\lambda \geq 0$ is a scalar penalty weight. The PLS estimator of $\eta = Xm$ is algebraically

$$\hat{\eta}_{pls}(\lambda) = X\hat{m}_{pls}(\lambda) = A(\lambda)y,$$

where $A(\lambda) = X(X'X + \lambda D_2' D_2)^+ X'$ and the superscript $^+$ denotes the Moore-Penrose pseudoinverse.

d) Explain algebraically why $\hat{m}_{pls}(0)$ *is not unique* for the motorcycle data.

The *estimated quadratic risk* of any symmetric linear estimator Ay for $\eta = Xm$ is

$$\hat{R}(A) = q^{-1}[|y - Ay|^2 + (2\mathrm{tr}(A) - n)\hat{\sigma}^2_{ls}].$$

e) Compute and report the estimated risk of $\hat{\eta}_{pls}(\lambda)$ for $\lambda = 0, 1000, 2000, 3000, 4000, 5000$.

f) Compute and report the value λ_{opt} that minimizes the estimated risk of $\hat{\eta}_{pls}(\lambda)$ over $\lambda \geq 0$; and the value of that smallest estimated risk.

g) Plot the components of the empirically best PLS estimator $\hat{\eta}_{pls}(\lambda_{opt})$ against the pertinent observation times, using the plotting character "o". Add appropriately to the plot the components of y, using the plotting character "x". Adjust plot size parameters to make the plot visually clear.

h) Explain algebraically why $\hat{m}_{pls}(\lambda_{opt})$ *is unique* for the motorcycle data.

i) Plot $\hat{m}_{pls}(\lambda_{opt})$ versus the times t. Use the plotting character "o" for the components of $\hat{m}_{pls}(\lambda_{opt})$ at the times in t_{obs} and the plotting character "x" for the remaining components at the times in t_{miss}. Adjust plotting parameters to make the plot visually clear. The goal is to reveal the nature of the PLS interpolation between the times at which acceleration was observed.

j) Plot the residuals $y - \hat{\eta}_{pls}(\lambda_{opt})$ against the observation times. Comment on what you learn, relative to the linear model fitted.