STA 232A: Lab #7

Fall 2013 R. Beran

DUE in Discussion on 26 November 2013. Use class techniques and results to do the project. Visit `www.stat.ucdavis.edu/~beran/s232a/litter.dat` for the data. Attach your computer code to the report.

This lab explores by penalized least squares techniques the rat litter data from Lab #3. Each response recorded in the data-set is the (average) weight gain of a rat litter when the infants in the litter are nursed by a rat foster-mother. Factor 1 is the genotype of the foster-mother nursing the infants. Factor 2 is the genotype of the infant litter. The **general model** for the data is an unbalanced two-way layout: the observed weight-gains $y_{ijk} = m_{ij} + e_{ijk}$, where $1 \le i \le p_1$, $1 \le j \le p_2$, and $1 \le k \le n_{ij}$ and the errors $\{e_{ijk}\}$ satisfy the strong Gauss-Markov condition. The $\{m_{ij}\}$ define a matrix of means.

a) Let m, y and e be the vectorized means, observed weight gains, and errors. Construct incidence matrix C so that the general model can be expressed as $y = Cm + e$. Compute the **least squares** estimates $\hat{m}_{LS}$ of m and $\hat{\sigma}^2_{LS}$ of σ^2 in the general model. Report the values of both, expressing the first in the same matrix form as the means $\{m_{ij}\}$.

Let P_0, P_1, P_2, P_{12} denote the orthogonal projections, used in Lab #3, that define the standard two-way ANOVA decomposition. Let

$$Q(t) = t_0 P_0 + t_1 P_1 + t_2 P_2 + t_{12} P_{12} \quad \text{where} \quad t = (t_0, t_1, t_2, t_{12}) \in [0, \infty)^4.$$

The candidate **penalized least squares** (PLS) estimators of m are

$$\hat{m}_{PLS}(t) = [C'C + Q(t)]^{-1} C'y, \qquad t \in [0, \infty)^4.$$

Consider the associated candidate **hypercubed PLS** (HPLS) estimators $\hat{m}_{HPLS}(d)$, where $d \in [0,1]^4$. These were developed in class to complete and stabilize numerically the class of PLS estimators. Define the **risk** of $\hat{m}_{HPLS}(d)$ to be

$$p^{-1} \mathrm{E} |C\hat{m}_{HPLS}(d) - Cm|^2 \quad \text{where} \quad p = p_1 p_2.$$

b) Consider the 16 candidate estimators $\hat{m}_{HPLS}(d)$ that arise when d is restricted to vectors whose components are each either 0 or 1. Explain how each such candidate estimator is the least squares estimator of m in an ANOVA submodel of the general model. Make a table in which the first column lists the 16 values of d while the second column gives the associated ANOVA submodel.

c) Drawing on course results, state the formulas for $\hat{m}_{HPLS}(d)$ and for its **estimated risk** $\hat{R}(d)$. Compute and report the estimated risk for each of the 16 candidate estimators considered in part b. Let $\hat{m}_{AN}$ be the estimator with smallest estimated risk among these candidates. Report that smallest estimated risk and $\hat{m}_{AN}$, expressing the latter in the same matrix form as the means $\{m_{ij}\}$.

d) Next, consider **all** candidate estimators $\hat{m}_{HPLS}(d)$ that arise when d ranges over the entire hypercube $[0,1]^4$. For which of these d values is estimated risk smallest? Let $\hat{m}_{OPT}$ be the candidate HPLS estimator that achieves it. Report that smallest estimated risk and $\hat{m}_{OPT}$, expressing the latter in the same matrix form as the means $\{m_{ij}\}$.

e) Plot against j the first row of the matrix $\{\hat{m}_{LS,ij}\}$ found in part a, joining the points with line segments to guide the eye. On the same plot, do the same for each of the other rows of $\hat{m}_{LS,ij}$, using a **different** line type for each row. Label the axes appropriately and indicate the foster-mother genotype associated with each line-type. Make two more such plots, one for the matrix $\{\hat{m}_{AN,ij}\}$ found in part c and the other for the matrix $\{\hat{m}_{OPT.ij}\}$ found in part d. To aid comparisons, adjust the plots so that they each have the same range on the vertical axis. Print all three plots on one sheet of paper.

f) Basic questions regarding the data include: How does foster mother genotype affect mean weight gains? How does litter genotype affect mean weight gains? Discuss these questions in the light of your numerical results and plots.