Lab #3

Fall 2013 R. Beran

DUE 22 Oct 2013 in Discussion. Visit `www.stat.ucdavis.edu/~beran/s232a/litter.dat` for the data. Use class linear algebra techniques to analyze the data. Attach your computer code to your report.

The rat litter data, printed on p. 140 of H. Scheffé's text, comes from a Ph.D. thesis *The Inheritance of Maternal Influences on the Growth of the Rat* by D. W. Bailey (1953). The response measured in the experiment is the (average) weight gain of an infant 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. We model the data as 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 $\{e_{ijk}\}$ are independent, identically distributed, $N(0, \sigma^2)$ random variables.

The **general** model puts no conditions on the means $\{m_{ij}\}$. The unrestricted means can be expressed as $m_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, subject to the constraints $\alpha_. = \beta_. = \gamma_{i.} = \gamma_{.j} = 0$.

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 estimate of m and the usual estimate of σ^2 under the general model.

b) Compute the residual vector $\hat{e} = y - C\hat{m}$, where $\hat{m}$ is the least squares estimate of m under the general model. Plot the ordered residuals against the corresponding quantiles of the standard normal distribution. Add to the plot the straight line that passes through the plotted points associated with the first and third quartiles of the residuals. The R functions `qqnorm` and `qqline` carry out these respective tasks. Does this plot of the residuals cast serious doubt on the normal error model?

c) As described in class, use orthogonal projection matrices and the Moore-Penrose pseudoinverse to compute the least squares estimate (LSE) of m under each of the following submodels.

- *Additive effects model:* $m_{ij} = \mu + \alpha_i + \beta_j$
- *Foster-mother effects model:* $m_{ij} = \mu + \alpha_i$
- *Infant effects model:* $m_{ij} = \mu + \beta_j$
- *No effects model:* $m_{ij} = \mu$

The constraints imposed by the general model on the $\{\alpha_i\}$, $\{\beta_j\}$ and $\{\gamma_{ij}\}$ still hold in these submodels.

d) Use the F-test p-values to assess the fit of each submodel in part c. Given the results, which of the five competing submodels seems most reasonable? (This is a traditional misuse of hypothesis testing theory to do submodel selection).

e) Use the method developed in class to estimate, *under the general model*, the normalized quadratic risk of the general model least squares fit and the risk of each submodel least squares fit in part c. Which of the five competing fits has smallest estimated risk? Explain carefully what this particular fit says about the effects of the two factors on the mean weight-gain of infant rat litters.