Advanced Methods in Biostatistics 2 (140.752)
Winter 2005

Homework Assignment #7
(Due Monday, December 19, 2005)

Please hand in a hard copy of your R code, and send an electronic version of it to Kenny (kshum@jhsph.edu).

1. Get `www.biostat.jhsph.edu/~iruczins/teaching/140.752/data/hw7.dat.1.R` which contains the blood pH readings on male mouse litter mates of two strains that had been selected for high and low blood pH. Only litters with at least 4 males were considered, and 4 males were selected at random whenever more than 4 males were present in the litter. Data are presented on seven litters for each strain.

Strain	Litter		pH readings		
pHH	387	7.43	7.38	7.49	7.49
	388	7.39	7.46	7.50	7.55
	389	7.53	7.50	7.63	7.47
	401	7.39	7.39	7.44	7.55
	402	7.48	7.43	7.47	7.44
	404	7.43	7.55	7.44	7.50
	405	7.49	7.49	7.51	7.54
pHL	392	7.40	7.46	7.43	7.42
	408	7.35	7.40	7.46	7.38
	413	7.51	7.39	7.42	7.43
	414	7.46	7.53	7.49	7.45
	415	7.48	7.53	7.52	7.43
	434	7.43	7.40	7.48	7.47
	446	7.53	7.47	7.50	7.53

Write down and explain your model, plot and analyze the data, and explain your results.

2. (a) Calculate the power to detect $\sigma_1^2 > 0$ in a balanced one-way ANOVA random effects model as a function of σ_1^2/σ^2, the number of groups p, and the number of observations per group r. Plot the power curves for $p = 2, 3, 4, 5$ and $r = 5, 10, 15, 20$.

   (b) In a blocked 'design', let σ_A^2 be the variance component of interest, and let let σ_B^2 be the block variance component. What is the power to reject $H_0 : \sigma_A^2 = 0$ when you have a levels for A, b blocks, and no replicates (i. e. $r = 1$)?

   (c) What is the power to reject $H_0 : \sigma_A^2 = 0$ when you do not account for extra variation due to blocks? Write down the distribution function of the test statistic and convince yourself that carrying out the actual power calculation analytically is hard.

   (d) Use simulations to estimate the power to reject $H_0 : \sigma_A^2 = 0$ for $a = 4, b = 5$, and $\sigma^2 = \sigma_A^2 = \sigma_B^2 = 1$. Confirm your results from (c) by simulating from the distribution function of the test statistic as well as simulating data and calculating an F-statistic with `aov()` in R. Compare the estimated power to the power for the correct model in (b).

3. Get `www.biostat.jhsph.edu/~iruczins/teaching/140.752/data/hw7.dat.2.R` which contains the daily cardiovascular disease mortality for Pittsburgh on the log scale (column one, `y`) and the daily pollution for Pittsburgh (column two, `x`), both obtained after removing effects from other variables such as temperature and dew point. We are interested in the model

$$y_t = \sum_{u=0}^{6} \beta_u x_{t-u} + \varepsilon,$$

i. e. we want to relate the current day mortality to the previous week's pollution. Note that since the data have been detrended, it is not necessary to fit an intercept in the model. In particular, we are interested in

$$\gamma = \sum_{u=0}^{6} \beta_u,$$

the distributed lag cumulative effect.

(a) Find the maximum likelihood estimates for $\beta_0, \ldots, \beta_6$ and σ^2 subject to the constraint $\sum_{u=0}^{6} \hat{\beta}_u = \gamma$ for an arbitrary γ.

(b) Write an R function that uses the above to plot the profile likelihood for γ defined as

$$L_p(\gamma) \quad = \quad \max_{\beta_0,\ldots,\beta_6,\sigma^2,\sum_{u=0}^{6}\beta_u=\gamma} L(\beta_0, \ldots, \beta_6, \sigma^2).$$

(c) Provide an interpretation of γ.