

Final Exam – ST790R – December 2008

***** Do (1) and either (2) or (3) *****

1. (Follow-up to Homework 7, Problem 5) North Carolina health statistics are compiled by county and often reported as rates. For example, infant mortality is reported both by number of deaths and as deaths per 1000 live births. For evaluating public health programs, researchers often investigate the relationship between death rates and county characteristics, especially economic and access to health care. In the file **ncinfmt.dat** are given the following information for each county in North Carolina: county name, county population (in thousands), infant death rates (deaths/1000 residents), household income (in thousands), per capita personal income (in thousands), poverty rate, and doctors per 1000 residents. (The first column, county, is dropped in the file **ncinfmt.cdat**.) State policy makers are interested in the effect of income and pose a regression model with infant death rate as the dependent variable and the others (aside from population) as covariates, as

$$\text{infdr}_i = \beta_0 + \beta_1 \text{hinc}_i + \beta_2 \text{pcpi}_i + \beta_3 \text{pov}_i + \beta_4 \text{mdpk}_i + e_i,$$

for $i = 1, \dots, N = 100$, where infdr_i is the infant death rate in county i , and hopefully the others are self-explanatory. But I'm interested in the variance, and propose modelling the variance as $\text{Var}(e_i) = V_{ii} = \sigma^2 + \gamma^2/\text{pop}_i$.

- a) First, compute least squares estimates for the regression model above with standard (Gauss-Markov) homoskedastic assumptions. Give your estimates and standard errors for the coefficients. (Compute them directly; you can use other software, e.g. *lm* only to check.)
- b) Compute preliminary estimates for σ^2 and γ^2 using a simple linear regression of the square of the residuals on the reciprocal of population, that is, \hat{e}_i^2 on $1/\text{pop}_i$.
- c) Compute maximum likelihood estimates for the generalized least squares model described above with diagonal covariance matrix \mathbf{V} , with $V_{ii} = \sigma^2 + \gamma^2/\text{pop}_i$. (Hint: the density for $\mathbf{y} \sim N_N(\mathbf{X}\mathbf{b}, \mathbf{V})$ is

$$(2\pi)^{N/2} |\mathbf{V}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{b})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b})\right\}.$$

- d) Perform a likelihood ratio test of the hypothesis $H : \gamma^2 = 0$ versus $A : \gamma^2 > 0$, at level $\alpha = 0.05$.

Do (2) or (3) but, please, not both

2. Following on (1), test this hypothesis a different way, by obtaining the distribution of the MLE for γ^2 when the hypothesis is true. Using your Gauss-Markov estimates as truth, generate M new datasets, that is, new $\mathbf{y}^{(i)}$'s, where $\mathbf{y}^{(i)} \sim N_N(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I})$ for $i = 1, \dots, m$, and compute the MLE's as in (1c) above. If you do this, say, M times, does the MLE for γ^2 from (1c) fall in the tail (more than the 95th percentile) of the distribution? (You can use the LRT statistic instead of your MLE for γ^2 if you would like.)
3. In the file **oring.dat** are observations on the o-ring failures for the space shuttle vehicles prior to the Challenger disaster in 1986. (The Challenger flight is the last observation with missing response.) The relevant columns are the last two, the number of failures (y_i) and the temperature (t_i). (The file **oring.cdat** has just the last two columns, and drops the last observation.) Following Example 14.10, do a permutation test of whether the failures are related to temperature, by computing the statistic $\sum y_i t_i$ and finding the probability that a random permutation of the y 's (or t 's) would give a larger (or smaller) sum. Generate random permutations and shuffle the y 's (or t 's).