

ST 762, HOMEWORK 5, FALL 2007

These problems are to be turned in on the due date.

1. Does how one estimates θ really matter? Consider the usual mean-variance model

$$E(Y_j|\mathbf{x}_j) = f(\mathbf{x}_j, \boldsymbol{\beta}), \quad \text{var}(Y_j|\mathbf{x}_j) = \sigma^2 g^2(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}_j), \quad (1)$$

where the (Y_j, \mathbf{x}_j) , $j = 1, \dots, n$ are independent, and suppose that both f and g are correctly specified.

Recall the following implications of the theory in Chapters 9–12 of the class notes.

- (i) The folklore theory for GLS says that, in large samples, the GLS estimator $\hat{\boldsymbol{\beta}}$ is equally precise whether $\boldsymbol{\theta}$ is known or estimated.
- (ii) Furthermore, the folklore theory says that, if one estimates $\boldsymbol{\theta}$, how one estimates it is of no consequence, so that all GLS estimators with $\boldsymbol{\theta}$ estimated are equally precise, regardless of how $\boldsymbol{\theta}$ was estimated (e.g., by using different transformations of absolute residuals).
- (iii) The theory for variance function estimation implies that, in large samples, estimators for $\boldsymbol{\theta}$ have different properties that depend on those of the underlying true distribution of the data (i.e., of $Y_j|\mathbf{x}_j$).

The big question is whether these implications of the first-order large sample theory are relevant in finite samples. In this problem, you will carry out simulation studies to investigate.

See Homework 4 for a review of the rationale behind simulation studies and details of how to carry out a simulation and summarize the results.

In all simulation studies you will perform, data will be generated according to the following model:

$$Y_j = f(x_j, \beta) + \sigma f^\theta(x_j, \beta) \epsilon_j,$$

where $x_j = 1.25, 2.5, 5, 10, 20, 40, 60, 80, 120, 160, 200, 300, 400$, (so $n = 13$),

$$f(x_j, \beta) = \frac{\beta_1}{1 + \beta_2/x}$$

ϵ_j are independent with mean 0 and variance 1, and the true values of the parameters are

$$\boldsymbol{\beta}_0 = (22, 30)^T, \quad \sigma_0 = 0.04, \quad \theta_0 = 0.8.$$

For each of the two simulation scenarios described below, you will do the following:

- Generate S data sets according to the model above.
- For each data set, estimate $\boldsymbol{\beta}$ assuming the model above using the three-step GLS algorithm with $C = 10$ with θ fixed at the true value. Call this estimator $\hat{\boldsymbol{\beta}}_{TRUE}$.
- For each data set, estimate $\boldsymbol{\beta}$ assuming the model above using the three-step GLS algorithm with $C = 10$ and θ estimated by the quadratic PL estimator ($\lambda = 2$). Call the estimators $\hat{\boldsymbol{\beta}}_{PL}$ and $\hat{\theta}_{PL}$.
- For each data set, estimate $\boldsymbol{\beta}$ assuming the model above using the three-step GLS algorithm with $C = 10$ and θ estimated using the identity transformation estimator ($\lambda = 1$). Call the estimators $\hat{\boldsymbol{\beta}}_{ID}$ and $\hat{\theta}_{ID}$.

- For each data set, estimate β assuming the model above using the three-step GLS algorithm with $C = 10$ and θ estimated using the log transformation estimator ($\lambda = 0$). Call the estimators $\hat{\beta}_{LOG}$ and $\hat{\theta}_{LOG}$.
- For each data set, save the 4 GLS estimates and the 3 estimates of θ to a file.
- From the results for all S data sets, calculate the Monte Carlo bias and MSE of $\hat{\beta}_{TRUE}$, $\hat{\beta}_{PL}$, $\hat{\beta}_{ID}$, and $\hat{\beta}_{LOG}$. For each component of β , plot a histogram of the S estimates of each type.
- From the results for all S data sets, calculate the Monte Carlo bias and MSE of $\hat{\theta}_{PL}$, $\hat{\theta}_{ID}$, and $\hat{\theta}_{LOG}$. Plot a histogram of the S estimates of each type.

Simulation 1: Generate the ϵ_j from a standard normal distribution.

Simulation 2: Generate the ϵ_j from a standardized *contaminated normal distribution*. (see Homework 3, Problem 1). That is, for given values α and b , generate each ϵ_j as follows:

- Generate a uniform random variable U on $(0,1)$.
- If $U \leq 1 - \alpha$, generate X from a $\mathcal{N}(0, 1)$ distribution. Otherwise, if $U > 1 - \alpha$, generate X from a $\mathcal{N}(0, b^2)$ distribution.
- Form $\epsilon_j = X\{(1 - \alpha) + \alpha b^2\}^{-1/2}$, which from Homework 3 Extra Problems, Problem 1, satisfies $E(\epsilon_j) = 0$, $\text{var}(\epsilon_j) = 1$.

For this simulation, take $\alpha = 0.05$ and $b = 3$.

From the results of each of Simulations 1 and 2, answer the following questions.

- (a) Do all of the 4 estimators for β appear to be unbiased? Does the sampling distribution of each component appear to be approximately normally distributed? *Explain.*
- (b) For the k th component of β , $k = 1, 2$, compute the MSE ratios

$$\frac{MSE(\hat{\beta}_{k,TRUE})}{MSE(\hat{\beta}_{k,\ell})},$$

where $\ell = PL, ID$, and LOG , respectively. According to the first-order theory, to what value should each ratio be approximately equal? Does there seem to be a loss of efficiency in either component associated with having to estimate θ rather than knowing it? Does this depend on which estimator for θ was used? *Explain.*

- (c) For the k th component of β , $k = 1, 2$, compute the MSE ratios

$$\frac{MSE(\hat{\beta}_{k,PL})}{MSE(\hat{\beta}_{k,\ell})},$$

where $\ell = ID$ and LOG , respectively. According to the first-order theory, to what value should each ratio be approximately equal? Does it seem to matter which estimator for θ is used? *Explain.*

- (d) From (a)–(c), do you think that the first-order theory for GLS estimators for β is relevant for this situation and sample size? *Explain.*
- (e) Do all of the 3 estimators for θ appear to be unbiased? Does the sampling distribution of each estimator appear to be approximately normally distributed? *Explain.*

(f) Compute the MSE ratios

$$\frac{MSE(\hat{\theta}_{PL})}{MSE(\hat{\theta}_\ell)},$$

where $\ell = ID$ and LOG . According to the theory, these should be approximately equal to the relevant values in Table 12.1 on page 308 of the notes when σ is “small.” How well do these MSE ratios compare with the relevant AREs given in the table? *Explain.*

(g) From (e)–(f), do you think that the first-order theory for estimators for θ is relevant for this situation and sample size? *Explain.*

2. Human exposure to chemicals, pollutants, and other nasty stuff is an important matter of concern in public health. The impact of such substances on the unborn is of special interest, and studies to evaluate the effect of exposure to the substances through the mother prior to birth, usually referred to as developmental toxicity studies, are commonly performed in animals. The motivation is that substances identified as dangerous to animals may also pose a threat to humans; thus, learning of the exposure risks in animals provides a model for the problem in humans.

The data in the file `dymedat.dat` on the class web page are from a small such study. 15 pregnant female mice were randomly assigned to be exposed to one of 5 doses of diethylene glycol dimethyl ether (DYME) ranging from 0.0 (control) to 0.5 g/kg/day, with 3 mice per dose. DYME is a widely-used organic solvent in a class of chemicals that has been linked to a variety of developmental toxicities. At birth, 3 baby mice were selected at random from the litter of each mother and weighed (measured in grams). Thus, for mother i , $i = 1, \dots, m = 15$, we may envision a trivariate random vector Y_i ($n_i = 3$) consisting of the birthweights for her 3 randomly selected offspring, for a total of $N = 45$ observations.

The investigators wished to characterize the relationship between birthweight and dose of DYME to which the mother was exposed during pregnancy. For the j th baby mouse from mother i , who was exposed to DYME dose x_i , $j = 1, 2, 3$, they postulated the following model for expected mean birthweight:

$$E(Y_{ij}|x_i) = \exp(\beta_1 + \beta_2 x_i) = f(x_i, \boldsymbol{\beta}), \quad \boldsymbol{\beta} = (\beta_1, \beta_2)^T. \quad (2)$$

Furthermore, based on previous studies, the investigators modeled the marginal variance as

$$\text{var}(Y_{ij}|x_i) = \sigma^2 f(x_i, \boldsymbol{\beta}). \quad (3)$$

Finally, the investigators assumed that the elements of \mathbf{Y}_i , conditional on x_i , follow a compound symmetric correlation pattern with correlation parameter α . Thus, in the notation of the notes, the covariance parameters are $\boldsymbol{\xi} = (\sigma, \alpha)$.

In this problem, you will fit this model.

(a) Write down the covariance matrix $\text{var}(\mathbf{Y}_i|x_i)$ under these assumptions.

(b) To implement the fit, you will use a linear estimating equation to estimate $\boldsymbol{\beta}$ and a quadratic estimating equation to estimate α ; that is, by solving the system of equations (14.21) on page 386 of the notes.

Using your result in (a), write down the vectors \mathbf{u}_i and $\mathbf{v}_i(\boldsymbol{\beta}, \boldsymbol{\xi})$ and the gradient matrix $\mathbf{E}_i(\boldsymbol{\beta}, \boldsymbol{\xi})$ that apply to this problem. *Hint:* Note that there are no unknown variance parameters; the only covariance parameter to be estimated is the correlation parameter α . Thus,

including the squared terms in \mathbf{u}_i would be redundant, as discussed on the bottom of page 379.

(c) To specify the quadratic estimating equation, you will also need to make some assumption about the “working covariance matrix” $\mathbf{Z}_i(\boldsymbol{\beta}, \boldsymbol{\xi})$. Write down the form of $\mathbf{Z}_i(\boldsymbol{\beta}, \boldsymbol{\xi})$ for this problem under the “Gaussian working assumption,” where all entries of $\mathbf{Z}_i(\boldsymbol{\beta}, \boldsymbol{\xi})$ are calculated assuming that the $\mathbf{Y}_i|x_i$ are normally distributed.

(d) In part (e) below, you will use the `nlinmix` SAS macro to solve the estimating equations characterized by the specifications in (a)–(c). To find starting values for $\boldsymbol{\beta}$, fit the above model using `proc genmod`, which uses simple moment methods to estimate correlation parameters. *Hint:* You should be able to figure out an appropriate combination of the `link` and `distribution` options of the `model` statement to represent the model; see the documentation.

(e) Now use the `nlinmix` SAS macro to solve the estimating equations characterized by the specifications in (a)–(c). Write down the estimates of $\boldsymbol{\beta}$ and α resulting from this implementation. Compare the results from using `proc genmod` in (d).