

**MA/ST 810, HOMEWORK 2, FALL 2009**  
*Due November 3, 2009*

Recall the chemostat model from Homework 1:

$$\begin{aligned}\frac{dN}{dt} &= r(c)N - qN \\ \frac{dc}{dt} &= q(c_0 - c) - \frac{1}{y}r(c)N,\end{aligned}\tag{1}$$

where  $N(0) = N_0$ ,  $c(0) = 0$ , and

$$r(c) = \frac{R_{\max} c}{K_m + c}, \quad q = \frac{Q}{V}.$$

Typical values for the parameters in this model are

$$V = 20 \text{ L}, \quad c_0 = 2.5 \text{ g/L}, \quad K_m = 12.3 \mu\text{g/mL}, \quad R_{\max} = 0.85/\text{hr}, \quad y = 10.6 \text{ L/g}$$

Take  $N_0 = 0.025 \text{ g}$  as in Homework 1.

In this assignment, you will carry out statistical *Monte Carlo simulation studies* of the OLS and GLS estimators for parameters in (1) in order to approximate and compare the sampling distributions of the estimators for finite sample sizes and to see how relevant the large-sample theory is in practice. In particular, as you did in Homework 1, you will generate samples of data of size  $n$  involving random noise from a situation where the variance of an observation is not constant but rather is of the form

$$\text{var}(Y_j|U) = \sigma^2 \{f^{(\ell)}(t_j, U, \theta)\}^2\tag{2}$$

(so  $\gamma$  in the handouts is known and equal to 1), where  $f^{(\ell)}(t_j, U, \theta)$  is the relevant component of the solution of (1) as described below at  $t_j$ .

You will consider the following two estimators for parameters in (1):

- *Ordinary least squares* (OLS). This is the estimator solving the minimization problem on Slide 6 of the handout “Statistical Inference for Independent Data” (equivalently, the estimator solving the corresponding estimating equation on Slide 19). This is the estimator you used in Problems 2 and 3 of Homework 1.
- *Generalized least squares* (GLS). This is the estimator solving the estimating equation (6) on Slide 20 of the same notes. For a given data set, solution of this equation for known  $\gamma$  may be implemented as follows (although there are other, more computationally efficient, ways that you are welcome to use):
  - (i) Find the OLS estimate; set  $c = 0$  and call this  $\hat{\theta}^{(0)}$ .
  - (ii) Form “estimated weights”  $\hat{w}_j = \{f^{(\ell)}(t_j, u, \hat{\theta}^{(c)})\}^{-2}$ .
  - (iii) Minimize  $\sum_{j=1}^n \hat{w}_j \{y_j - f^{(\ell)}(t_j, u, \theta)\}^2$  in  $\theta$  to obtain  $\hat{\theta}^{(c+1)}$ , where  $y_j$  is the data value at  $t_j$ . Set  $c = c + 1$  and go to (ii).

Continue to cycle between (ii) and (iii) until the relative change between  $\hat{\theta}^{(c)}$  and  $\hat{\theta}^{(c+1)}$  is “small” (less than  $10^{-6}$ , say).

The simulation study you design and carry out will address the following questions; how it will do this is discussed below:

- (a) Are these estimators both consistent?
- (b) Is the OLS estimator inefficient relative to the GLS estimator, as predicted by the large-sample theory?
- (c) Are the approximations to the sampling distributions of the OLS and GLS estimators given in (7) on Slide 26 for OLS and in (9) on Slide 29 for GLS relevant in finite samples? That is, are the sampling distributions approximately normal with the given sampling variances (standard deviations)? In particular, if we use OLS when the data really have nonconstant variance, and then blindly use the large-sample approximate sampling distribution (7) on Slide 26 that is based on constant variance, will our impression of sampling variance be erroneous?
- (d) Are approximate *confidence intervals* for the true values of the components of  $\theta$  based on the large sample theory reliable? That is, if we construct confidence intervals using the estimated standard errors from the large sample theory according to the technique on Slide 19 of the notes “Principles of Statistical Inference,” with  $\alpha = 0.05$ , if such intervals are reliable, the true value of the component should be contained in the endpoints for  $100(1 - \alpha)\% = 95\%$  of all data sets of size  $n$ .

*Brief description of the rationale and implementation of a Monte Carlo simulation study:* Large-sample theory is useful for gaining insight into the sampling properties of estimators and how estimators compare (via asymptotic relative efficiency). However, such theory is admittedly idealistic, as it is derived under the condition  $n \rightarrow \infty$ . Alternatively, a popular way to learn about the finite-sample properties of estimators and whether the implications of the large-sample theory hold in finite-sample situations is by Monte Carlo simulation.

The objective of a simulation is to approximate the sampling distribution of an estimator by generating some large number  $S$  of independent data sets from a known situation and computing the estimator for each data set; usually,  $S \geq 500$  or thereabouts in order to achieve an accurate approximation. (Recall that the sampling distribution is the distribution of all possible values of the estimator across all possible data sets, so we are approximating this distribution based on a sample of  $S$  data sets.) The sample mean of the  $S$  estimates over all  $S$  data sets is an estimate of the mean of the sampling distribution of the estimator; similarly, the standard deviation of the  $S$  estimates over the  $S$  data sets is an estimate of the standard deviation of the sampling distribution. (How good these quantities are at capturing the true features of the sampling distribution obviously depends on the size of  $S$ ).

To carry out a simulation that addresses (a)–(c) above, we would do the following.

- Generate  $S$  data sets under some particular situation (see below).
- For each data set, estimate  $\theta$  in (1) by OLS and GLS.

- Estimate standard errors for GLS based on (9) on Slide 29, so assuming correctly that the variance is nonconstant as in (2). (You will need to estimate  $\sigma^2$  as on Slide 21.)
- Estimate standard errors for OLS based on (7) on Slide 26, so assuming incorrectly that the variance is constant. (Because you are assuming the variance of  $Y_j$  is constant across  $j$ , you should use the estimator for  $\sigma^2$  given on Slide 9.)

We may then tackle (a)–(d) above as follows. Here, let  $\theta_0$  denote the true value of  $\theta$  from which the data were generated, with components  $\theta_{0,k}$ ,  $k = 1, \dots, p$ , and let  $\hat{\theta}_{k,s}$  denote the  $k$ th component of an estimator for  $\hat{\theta}_k$  calculated from the  $s$ th generated data set.

- (a) If the OLS and GLS estimators are consistent, we would hope that they would be approximately unbiased in finite samples. Thus, we would hope that the mean of the sampling distribution for each one is close to the true value of  $\theta$ , with only minimal bias. To assess this based on the  $S$  observations from the sampling distribution of each estimator, take the average of the  $S$  estimates for each method and then inspect its difference from  $\theta_0$ ; this difference,  $(S^{-1} \sum_{s=1}^S \hat{\theta}_{k,s}) - \theta_{0,k}$ , is the *bias*. To put this on a “dimensionless” basis, also calculate the *relative bias* for each component of an estimator  $\hat{\theta}$ , expressed as a percentage, defined for the  $k$ th component as

$$\frac{(S^{-1} \sum_{s=1}^S \hat{\theta}_{k,s}) - \theta_{0,k}}{\theta_{0,k}} \times 100\%;$$

this measures the size of the bias relative to the size of the thing being estimated. One would hope that this relative bias is small in magnitude (on the order of a few percent) if the estimator is consistent.

- (b) To compare the precision of the OLS and GLS estimators based on the  $S$  estimates of each, we could compare their sample variances across the  $S$  data sets, thus mimicking the idea of asymptotic relative efficiency. However, because the estimators may exhibit some *bias* for finite  $n$ , as characterized in (a), it is standard instead to take this into account and compute the *mean square error* (MSE) for each estimator. For an estimator  $\hat{\theta}$ , the estimated MSE based on the  $S$  estimates  $\hat{\theta}_s$ , say,  $s = 1, \dots, S$ , for the  $k$ th component, is defined as

$$S^{-1} \sum_{s=1}^S (\hat{\theta}_{k,s} - \theta_{0,k})^2 = S^{-1} \sum_{s=1}^S (\hat{\theta}_{k,s} - \bar{\theta}_k)^2 + (\bar{\theta}_k - \theta_{0,k})^2,$$

where  $\bar{\theta}_k$  is the sample average of the  $\hat{\theta}_{k,s}$ . Note that MSE may thus be interpreted as sample variance over the  $S$  estimates plus observed bias, squared. If bias is small, then MSE is approximately equal to the estimate of the variance of the sampling distribution based on the  $S$  estimates.

The ratio of estimated MSE values may be used as a measure of relative precision, similar to asymptotic relative efficiency; that is, one may compute “MSE ratios” for each component of  $\hat{\theta}$  as approximations to the true ARE values.

- (c) To assess how well the estimated standard errors approximate the true sampling variation, one may compare the sample standard deviation of each component of the  $S$  estimates  $\hat{\theta}$  to the average of the estimated standard errors over the  $S$  data sets for that component found using the large sample theory approximations. If the theory is relevant, we would expect the sample standard deviation, an approximation to the true standard deviation of the sampling distribution, and the average of the estimated standard errors over the  $S$  data sets, each of which is trying to estimate the true standard deviation, to be “close.” Furthermore, as the large-sample approximate sampling distributions are normal, we would also expect that histograms of all  $S$  estimates for each of OLS and GLS would look “bell-shaped.”
- (d) To assess the performance of confidence intervals, one may construct a 95% confidence interval for each of the  $S$  data sets for a component of  $\theta$  using the estimated standard errors and calculate the proportion of the  $S$  data sets for which the interval so calculated “covers” the true value of that component. If the interval is reliable, we would hope that this proportion would be close to 0.95.

To carry out a Monte Carlo simulation study to address the issues above, you should do the following.

- Based on your experience in Homework 1, choose a value of  $Q$  in the range 14 to 16 (L/hr).
- Choose a set of time points  $(t_1, \dots, t_n)$  for some value  $n$  that is not too large (so representing a “finite sample”) but for which you did not have trouble finding the OLS estimate in Homework 1, Problem 3.
- Simulate the forward solution to (1) for your chosen scenario to obtain  $f(t_j, U, \theta) = \{f^{(1)}(t_j, U, \theta), f^{(2)}(t_j, U, \theta)\}^T$  at each  $t_j$ , i.e., the solutions for  $c(t)$  and  $N(t)$ , respectively, at a particular  $t_j$  for a particular value of  $\theta$ .
- Focus on  $N(t)$  as the response of interest, so that the data you will generate and fit are univariate. Thus, the component of  $f(t_j, U, \theta)$  of interest here at each  $t_j$  is  $f^{(2)}(t_j, U, \theta)$ , so that  $\ell = 2$  in the discussion above.
- Identify  $\theta = (R_{\max}, K_m)^T$ , and treat all other parameters in (1) as fixed.
- Choose a value of  $\sigma = 0.05$  or  $0.10$  as in Homework 1.
- Generate  $S$  data sets, where  $S$  is “large” as described above.
- To obtain a single data set, obtain univariate data  $y_j$ ,  $j = 1, \dots, n$ , according to the statistical model

$$Y_j = f^{(2)}(t_j, U, \theta) + \sigma f^{(2)}(t_j, U, \theta)\epsilon_j,$$

where the  $\epsilon_j$  are generated independently for  $j = 1, \dots, n$  according to a distribution of your choice with variance 1; e.g., one of those in Problem 3 of Homework 1: (a)  $\epsilon_j$  uniformly distributed on  $(-\sqrt{3}, \sqrt{3})$ , or (b)  $\epsilon_j$  standard normal (mean 0, variance 1).

- Estimate  $\theta$  using OLS and GLS for each data set as described above and obtain estimated standard errors for each.
- For each method, obtain the average and standard deviation of the  $S$  estimates, compute the bias, relative bias, and MSE; obtain the average of the  $S$  estimated standard errors; and construct confidence intervals. Based on these results, draw conclusions on the relevance of the large-sample theory as described in (a)–(d) above.

The diligent group will carry out several simulations, each of size  $S$ , varying  $\sigma$ ,  $n$ , the  $(t_1, \dots, t_n)$ , and the distribution of the  $\epsilon_j$ .