

ST 762, HOMEWORK 4, FALL 2009

1. Do the “folklore” properties hold in finite samples? Consider our 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. Under these conditions, the theory in Chapter 9 says that, as $n \rightarrow \infty$,

- (i) The OLS estimator, the GLS estimator with $\boldsymbol{\theta}$ known and g correctly specified, and the GLS-PL estimator with g correctly specified are all consistent estimators for the true value of $\boldsymbol{\beta}$
- (ii) The GLS estimators are more precise than the OLS estimator
- (iii) The GLS estimator with $\boldsymbol{\theta}$ known and the GLS estimator with $\boldsymbol{\theta}$ estimated by PL are equally precise
- (iv) Valid approximate standard errors for the GLS estimators may be calculated using the “folklore” theorem as in Equation (9.9) on page 218, while approximate standard errors for OLS calculated assuming constant variance (i.e., using (9.9) with the matrix \mathbf{W} equal to an identity matrix) will lead to inappropriate assessment of the precision of the OLS estimator.

In this problem, we will investigate the relevance of these claims for fixed n .

When analytical arguments are intractable, a popular way to learn about the finite-sample properties of estimators is by Monte Carlo simulation. The objective of a simulation is to approximate the sampling distribution of an estimator by generating (via random deviate generation routines) some large number S independent data sets from a known situation and computing the estimator for each data set. The sample mean of the 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 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 (i)–(iv), we would do the following.

- Generate S data sets (see below).
- For each data set, estimate $\boldsymbol{\beta}$ in (1) by OLS, GLS with $\boldsymbol{\theta}$ set equal to the true value used to generate the data ($\boldsymbol{\theta}$ known), and GLS with $\boldsymbol{\theta}$ estimated using PL (GLS-PL).
- Estimate standard errors for the two GLS estimators by substituting the corresponding estimates in (9.9) assuming the model (1) is correct.
- Estimate standard errors for OLS by substituting the OLS estimates in (9.9), erroneously assuming that $\text{var}(Y_j|\mathbf{x}_j)$ is constant, so setting \mathbf{W} equal to an identity matrix in (9.9) with $\hat{\sigma}^2$ equal to the usual OLS estimator of the assumed constant variance.

We may then address (i)–(iv) as follows. Here, let $\boldsymbol{\beta}_0$ denote the true value of $\boldsymbol{\beta}$, with components $\beta_{0,k}$, $k = 1, \dots, p$, and let $\hat{\beta}_{k,s}$ denote the k th component of an estimate for β_k calculated from the s th generated data set.

- (i) If the 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 is close to the true value of β , with only minimal bias. To assess this based on the S observations from the sampling distribution, calculate the *Monte Carlo bias* for each component of an estimator $\hat{\beta}$, defined for the k th component as

$$S^{-1} \sum_{s=1}^S \hat{\beta}_{k,s} - \beta_{0,k}.$$

It is standard to report the “raw” Monte Carlo bias as given above. It is also standard to report this bias relative to the true value (so report

$$\frac{S^{-1} \sum_{s=1}^S \hat{\beta}_{k,s} - \beta_{0,k}}{\beta_{0,k}},$$

which can of course be problematic when the true value is very close to 0) and relative to the Monte Carlo standard deviation (see (iii) below), so that the size of the bias relative to the variation in the estimator can be assessed.

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

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

where $\bar{\beta}_k$ is the sample average of the $\hat{\beta}_{k,s}$. Note that MSE may thus be interpreted as sample variance over the S estimates plus observed bias, squared.

The ratio of estimated MSE values may be used as a measure of relative precision, similar to asymptotic relative efficiency.

- (iii) 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 $\hat{\beta}$, that is, the *Monte Carlo standard deviation*, to the average of the estimated standard errors for that component found using the “folklore” result (9.9). If the theory is relevant, we would expect the sample standard deviation, an approximation to the true sampling variation, and the average of estimated standard errors, to be “close.”
- (iv) To assess further how well the asymptotic sampling distribution approximates the true sampling distribution, for each estimator, calculate for each of the S data sets 95% Wald confidence intervals (using the usual critical value from the standard normal distribution of 1.96) for the true values of each component of β , and record the proportion of times that the intervals contain the true values. These proportions are *Monte Carlo coverage probabilities* – if the Wald intervals are reliable, we would expect them to be close to the nominal coverage probability of 0.95. If the Monte Carlo values are not close to 0.95, then using the large-sample approximation may be unreliable.

(a) In your favorite programming language, write a program to carry out such a simulation. These should have the following features.

- Each of S data sets should contain n observations generated for $j = 1, \dots, n$ as

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

where $x_j = 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.6, 2, 2.8, 4, 6, 8, 10, 12, 16, 20, 24, 30$ (so $n = 18$),

$$f(x_j, \beta) = e^{\beta_1} \exp(-x_j e^{\beta_2}),$$

(the monoexponential model parameterized to ensure positivity), where $\epsilon_j \sim N(0, 1)$ are independent, and the true values of the parameters are

$$\beta_0 = \log(9.5, 0.25)^T, \quad \sigma_0 = 0.01, \quad \theta_0 = 0.8$$

See below for suggestions on choosing S .

- For each data set, estimate β , σ , and θ using (a) OLS, (b) the 3-step GLS algorithm with $C = 10$ iterations with θ known and equal to the true value above (GLS-known); and (c) the 3-step GLS algorithm with $C = 10$ iterations and PL estimation of θ at step 2 (GLS-PL). You will find it convenient to use the same programs and conventions you have used previously for this. You may use the true values of β and θ as starting values.
- For each data set, also calculate estimated standard errors associated with each element of the OLS and the two GLS approaches using the folklore theory as described above. Note that the standard errors you compute for OLS are inappropriate, as they assume that the variance is constant when it clearly is not.
- Output the OLS, GLS-known, and GLS-PL estimates for β and the corresponding estimated standard errors for each component from each of the S data sets to a file so that you can compute summary statistics to address (i)–(iv) as discussed below.

(b) To address (i), calculate the raw bias and relative biases of the OLS, GLS-known, and GLS-PL estimators for each component of β as in (i). Comment on the implications of these results. Are the estimators approximately unbiased for this sample size?

(c) To address (ii), calculate the MSE for each component of the OLS, GLS-known, and GLS-PL estimators, and form the following ratios:

$$\frac{MSE(GLS - known)}{MSE(OLS)} \quad \text{and} \quad \frac{MSE(GLS - PL)}{MSE(OLS)};$$

each of these ratios is obviously an approximation to the asymptotic relative efficiency of OLS to that of each of the GLS estimators. Is there evidence that one or both of the the GLS estimators are more precise than the OLS estimator for this sample size?

(d) To also address (ii), form the ratio

$$\frac{MSE(GLS - known)}{MSE(GLS - PL)},$$

which measures the relative efficiency of the GLS estimator where θ is estimated (GLS-PL) to that of the ideal GLS estimator where θ is known. Is there evidence at this sample size

that having to estimate θ causes a degradation in precision, despite the implications of the “folklore” theory?

(e) To address (iii), calculate the sample standard deviation of each component of the OLS and the GLS estimates, and compare it to the average of the estimated standard errors for that component. Do you believe that the folklore theory yields reliable estimated standard errors for this sample size? For OLS, does erroneously assuming that the variance of the response is constant result in an erroneous assessment of precision of the OLS estimator?

(f) To address (iv), calculate the Monte Carlo coverage probabilities of the Wald intervals based on the OLS and GLS estimates. Are the OLS intervals reliable? Does the folklore theory yield a reasonable approximation for intervals based on the GLS estimators?

Some practical pointers:

- Both SAS and R allow generation of random deviates from the $\mathcal{N}(0, 1)$ distribution. It is a good idea to pick a “starting seed” for the random number generator to begin your simulation so that you can reproduce the results rather than let the random number generator start wherever it likes. The random number generator will update the seed automatically and internally each time you call it, so you only need to set the initial seed at the beginning of your program.
- Random number generators work by generating a sequence of random deviates that should be approximately independent. Thus, you do not want to change the seed yourself over the course of the simulation once it starts.
- One of the hazards of doing a simulation is that, occasionally you will encounter a data set for which the routine will not converge (it will “bomb”). This is usually quite rare. The usual convention is to delete the “bad” data set from consideration and move forward to the next data set in the sequence dictated by the random number generator. Thus, you will need to make sure you keep track of the value of the seed so that you can “restart” the simulation at the next data set in the event that it “bombs.” If this happens a *lot*, then something is wrong; if it happens once in a while, it is generally no cause for concern in terms of using the results to get a reliable idea of the properties of the estimator. *Do not* restart the simulation at another seed you pick yourself – this may lead to non-independent data sets.
- Following these principles, continue to generate data sets until you get S convergent data sets. Keep a count of the number of nonconvergent data sets you encounter (it should be very small).

For S , use a minimum of $S = 500$; $S = 1000$ is preferred. Depending on how efficiently you implement the simulation, it may move quickly or slowly. If yours is pretty quick, you will want to use a larger S . A simulation is just like any other experiment – the sample size (S) should be large enough to achieve acceptable precision of the estimates of the sampling characteristics. $S = 500$ should be sufficient for us to get a basic idea for our purposes. To read more about how to design a simulation study, see the guide “*Simulation study in statistics*” by Erning Li, Dennis Boos, and Marcia Gumpertz, available on the ST 810A web site <http://www4.stat.ncsu.edu/~davidian/st810a/>.