Parameter and Confidence Interval Estimation in Dynamic Models: Maximum Likelihood and Bootstrapping Methods

Appendix

Jeroen Struben, John Sterman, David Keith

A0 Documentation Overview

Table A1. Documents, R-scripts, and data included in the online appendix

Document	Contents	Main R-script	Subordinate R-scripts used	Data
Main document	Theory Application Challenge	CH1_MLE_BOOT_Application.R	CH1_MLE_Functions.R CH1_BOOT_Functions.R CH1_LR_Interval_Functions.R	ServiceQualityData2.csv
Appendix	Further detailing of MLE theory, using linear model as an example	CH1_MLE_BOOT_LinEx.R	CH1_LinEx_Functions.R	-
Challenge Solutions	Solutions to the challenge in the main document	CH1_MLE_BOOT_CHALLENGE.R	CH1_MLE_Functions.R CH1_BOOT_Functions.R CH1_LR_Interval_Functions.R CH1_Challenge_Functions.R	Beer Game Subject 1.csv
Required folder to save:		Scripts	Scripts	Data

All documents except the main document are provided in the electronic supplement on the publishers' handbook website. The electronic documents can also be requested from one of the authors (jeroen.struben@mcgill.ca).

Start instructions:

1. Create a work folder for your analysis. (You may use different work folders for the application and the challenge).
2. Within the work folder create three subfolders: "Scripts", "Data", and "BookChapterOutput".
3. Save the provided documents, listed above, in the appropriate subfolders ("Scripts" or "Data"). Save any of your R-script files in the "Script" folder.
   Note: Figures will be saved in the folder "BookChapterOutput".

A1 Derivation of the Likelihood function

Suppose, for a function y, there is a vector $\boldsymbol{u}$ of k random variables, $\boldsymbol{u} = [u_1, \ldots, u_k]^\top$ and that we know the joint density function $g(\boldsymbol{u})$ of vector $\boldsymbol{u}$. If a function $\boldsymbol{y} = [y_1,\ldots,y_k]^\top$ is monotonic in $\boldsymbol{u}$, we can find the joint density function for **y** by multiplying the density function $\boldsymbol{g(u)}$ in $\boldsymbol{y}$, with the Jacobian of the transformation $J\left(u_i \rightarrow y_j\right)$. The Jacobian matrix, often used in optimization problems, provides the first order partial derivatives of u with respect to y, $\partial u_i/\partial y_j$. Hence, the pdf of y_i is $f(y_i) = f(u_i)|J|$, or

$$f(\boldsymbol{y}) = g\left(\boldsymbol{u}(\boldsymbol{y})\right)\left\|\frac{\partial u_i(\boldsymbol{y})}{\partial y_j}\right\| \tag{A1}$$

where the second term on the right-hand side is the absolute value of the Jacobian matrix of $u=f(y)$ (Wilks 1943).

Thus, to derive the density function for y as a function of the parameters we seek to transform (the unobservable) error distribution r_i into the observable y_i, through the monotone relation. If the error terms are independent, then the Jacobian $J\left(u_i \rightarrow y_j\right)$ is simply equal to one because $\partial u_i/\partial y_i = 1$ while $\partial u_i/\partial y_j = 0 \forall i \neq j$. Then the distribution we are looking for, $f(\boldsymbol{y})=g(\boldsymbol{u})=g(\boldsymbol{y}-f(\boldsymbol{x}, \boldsymbol{\theta}))$.

If we further impose the common assumptions of: (i) homoscedaskicity, i.e. the variance of the errors is constant for all x_i; (ii) the explanatory variable is measured without error; and (iii) normally distributed errors, then the error terms r_i are identically, independently and normally distributed. The pdf of y_i is then:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma}}|J|e^{-\frac{1}{2\sigma^2}u_i^2} = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2\sigma^2}\left(y_i - f(x_i,\boldsymbol{\theta})\right)^2}$$

with the likelihood function:

$$L\left(\boldsymbol{\theta},\sigma^2; x_{1,\ldots},x_n\right) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}\left(y_i - f(x_i,\boldsymbol{\theta})\right)^2} = \frac{1}{2\pi\sigma^2}^{\frac{1}{2}n}\ e^{\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(y_i - f(x_i,\boldsymbol{\theta})\right)^2\right)} \tag{A2}$$

which becomes the log-likelihood (*LL,* or *l*):

$$
\begin{aligned}
l\left(\mu,\sigma^2;x_{1,\ldots},x_n\right) &= \ln\left(L\left(\mu,\sigma^2;x_{1,\ldots},x_n\right)\right) \\
&= \ln\left(\frac{1}{2\pi\sigma^2}^{\frac{1}{2}n} e^{\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(y_j-f\left(x_j,\boldsymbol{\theta}\right)\right)^2\right)}\right) \\
&= \ln\left(\frac{1}{2\pi\sigma^2}^{\frac{1}{2}n}\right)+\ln\left(e^{\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(y_j-f\left(x_j,\boldsymbol{\theta}\right)\right)^2\right)}\right) \\
&= -\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\left(\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(y_j-f\left(x_j,\boldsymbol{\theta}\right)\right)^2
\end{aligned}
\tag{A3}
$$

A2 Likelihood function for a linear model with normal *iid* errors

In the case of models that are linear in parameters, and under the standard assumptions of *iid* normally distributed errors, the MLE is equivalent to traditional ordinary least squares (OLS) estimation. In fact, because of the linearity, we can actually solve the unconstrained maximization problem of the MLE using calculus and simple linear algebra. Solving this is identical to solving the first order equations for the MLE.

A2.1 Deriving the likelihood function

To illustrate, consider a true relationship between independent and dependent variables as follows: $y_i = a + bx_i + \varepsilon_i$ with true parameters: *a=20*, b=0.2 and independently, identically and normally distributed errors $\varepsilon_i \sim N(0,\sigma^2)$, with $\sigma^2 = 1$.

The likelihood function $L(a,b,\sigma^2|y_{1,\ldots},y_n)$ is defined as the joint pdf over all y_i's, $f(y_{1,\ldots,} y_{i,\ldots} y_n)$. Because the $y_{i's}$ are independent, this joint pdf is equal to their product $f(y_1)\ldots f(y_i)\ldots f(y_n)$ (equation 1). To obtain the Log likelihood function we take the logarithm of this product (equation 2), which yields:

$$
\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_i\left(y_i-\hat{a}-\hat{b}x_i\right)^2 \tag{A4}
$$

Finding the MLE consists of finding the parameters that maximize the log likelihood, $\max_{a,b,\sigma^2} \ln L\left(a,b,\sigma^2\middle|y_1,\ldots,y_n\right)$.

The maximization can be performed analytically if the system of equations is twice differentiable using the standard first- and second-order conditions. In the linear case, one can derive the results in two steps. First, solving the first order condition for $\partial \ln L \big/ \partial\hat{\sigma}^2$ produces exactly the sum of square errors, $SSE = 1\big/n\sum_i\left(y_i-\hat{a}-\hat{b}x_i\right)$. Further, the first order condition $dlnL/d\boldsymbol{\theta} = 0$ reduces to minimizing $\sum_i\left(y_i-\hat{a}-\hat{b}x_i\right)$. In more detail:

$$\max_{\mu,\sigma^2} l\left(\boldsymbol{\theta},\sigma^2;x_{1,\ldots},x_n\right) \quad \Rightarrow \frac{\partial l}{\partial a}=0 \wedge \frac{\partial l}{\partial b}=0 \wedge \frac{\partial l}{\partial \sigma^2}=0$$

$$\frac{\partial l}{\partial a}=\frac{1}{\sigma^2}\sum_{j=1}^{n}\left(y_j-a-bx_j\right)=0 \Rightarrow \hat{a}=\frac{1}{n}\sum_{j=1}^{n}\left(y_j-bx_j\right)=\langle y\rangle-\hat{b}\langle x\rangle$$

$$\frac{\partial l}{\partial b}=\frac{1}{\sigma^2}\sum_{j=1}^{n}x_j\left(y_j-a-bx_j\right)=0 \Rightarrow \text{using } \hat{a}=\langle y\rangle-\hat{b}\langle x\rangle \Rightarrow$$

$$\frac{1}{\sigma^2}\sum_{j=1}^{n}x_j\left(\left[y_j-\langle y\rangle\right]-\hat{b}\left[x_j-\langle x\rangle\right]\right)=0 \Rightarrow \hat{b}=\frac{\sum_{j=1}^{n}x_j\left[y_j-\langle y\rangle\right]}{\sum_{j=1}^{n}x_j\left[x_j-\langle x\rangle\right]}$$

$$\frac{\partial l}{\partial \sigma^2}=\frac{1}{2\sigma^2}\left(\frac{1}{\sigma^2}\sum_{j=1}^{n}\left(y_j-\hat{a}-\hat{b}x_j\right)^2-n\right)\underset{=0 \textit{ iff}}{\Rightarrow}\widehat{\sigma^2}=\frac{1}{n}\sum_{j=1}^{n}\left(y_j-\hat{a}-\hat{b}x_j\right)^2 \tag{A5}$$

Thus, $\hat{a}$, $\hat{b}$ and $\widehat{\sigma}^2$ are equal to the sample intercept, slope, and variance. To estimate a and b we can use standard linear OLS regression.

The OLS estimator minimizes the residual sum of squares (RSS, or, sum of squared errors, SSE), across all data points. The residual of a datapoint is then $r_i = y_i - \hat{y}_i = y_i - f\left(x,\hat{\theta}\right)$. Then, $RSS=\sum_i r_i^2=\sum_i\left(y_i-\hat{y}_i\right)^2=\sum_i\left(y_i-f\left(x_i,\hat{\theta}\right)\right)^2$. For the linear regression the OLS is simply: $\sum_i\left(y_i-\hat{a}-\hat{b}x_i\right)^2$. Hence, in this case, under linearity and with *iid*-distributed normal errors, the OLS is an efficient estimator. That is, in this case, least squares then corresponds exactly to maximizing the likelihood or probability that the parameters have been chosen correctly, given the data in the sample.

Further, because the model is linear we can derive the estimates "manually" using core statistics (sumY^2, sumY, sumX^2, sumX, sumXY). First we calculate $(XY)^T$*XY. Next we need to get the corrected sums (SSX, SSY, SSXY) which we can derive from the MLE estimate of b=SSXY/SSX. In the R-scripts we provide an example with a=20 and b=0.2. Given the sample we use, the manual estimation yields a= 20.0453134, b= 0.1956545, identical to the automated OLS result.

2.1.1 Univariate likelihood intervals (LR-univariate)

The *univariate likelihood* is equivalent to taking a slice of the LL function along the dimensions of the parameter of interest. The univariate method requires no new optimizations once the MLE has been found. The critical parameter value for θ_k is then simply: $LL\left(\theta_k;\hat{\boldsymbol{\theta}}_{-k}\right)=LL\left(\hat{\boldsymbol{\theta}}\right)-\chi_{1,1-\alpha}$, where $\hat{\boldsymbol{\theta}}_{-k}$ is the vector of MLE estimates for all parameter but

k.[1] A disadvantage of univariate confidence interval estimation, however, is that the parameter space is not fully explored, hence the effect of any interactions among parameters on LL is ignored.

2.1.2 Surface likelihood (LR-surface)

The *surface likelihood* makes full use of the actual curvature of the inverted multidimensional bowl to estimate the uncertainty in the estimates due to sampling error. As in the univariate case this method requires no new optimizations. The likelihood function surface is simply the k-dimensional likelihood function $LL(\boldsymbol{\theta})$, while the confidence region is defined by the curve that satisfies $LL(\boldsymbol{\theta}) = LL(\hat{\boldsymbol{\theta}}) - \chi_{k,1-\alpha}$. The confidence interval for each parameter is then the widest range of parameter values that is contained within the enclosed regions.

2.1.3 Profile likelihood (LR-profile)

Following the same logic one can construct the confidence region for individual parameters (Cox and Snell 1989). In this *profile likelihood* method one reduces the log likelihood to a single-parameter function of the parameter of interest, *j*, by treating the others, *~j,* as "nuisance parameters" over which the likelihood function is to be maximized for each value θ_j . Hence the profile likelihood function is $L(\theta_j) = \max_{\boldsymbol{\theta}_{-j}} L(\theta_j, \boldsymbol{\theta}_{\sim j})$. The method therefore requires new optimizations for each parameter, once the MLE has been found.

The ratio of the profile likelihood function and the likelihood function of the estimate L($\hat{\theta}$) again follows a χ^2 distribution (Equation 3).[2] Increasing the number of parameters must expand the parameter region that falls within a given confidence level. Indeed, since the mass of the χ^2 distribution shifts to right as the degrees of freedom increase (with the distribution approaching the normal distribution for large k), the confidence interval narrows when constraining the number of parameters.

[1] For *univariate* confidence intervals, the χ^2 test has one degree of freedom. Formally, the degrees of freedom are given by the number of restrictions of the alternative, "constrained" model compared to the "unconstrained" null. In univariate contexts, the optimizations in the interval construction occur over all the other k-1 parameters, being constrained only by the parameter for which the interval is constructed. Hence, the degrees of freedom are n-k-1. Then for the χ^2 test, df =(n-k)-(n-k-1)= 1; when finding the confidence region for two parameters jointly, df = 2, and so on.

[2] Under profile confidence interval estimation the "null" involves maximizing all k parameters, while for the alternative we maximize over k-i parameters, with i the number of parameters of interest. Hence, Hence, df=i. If we seek the parameter region for all parameters jointly, df=i=k, while df=i=1 if we seek the profile for a single parameter.

A3 Additional output for the application in the paper

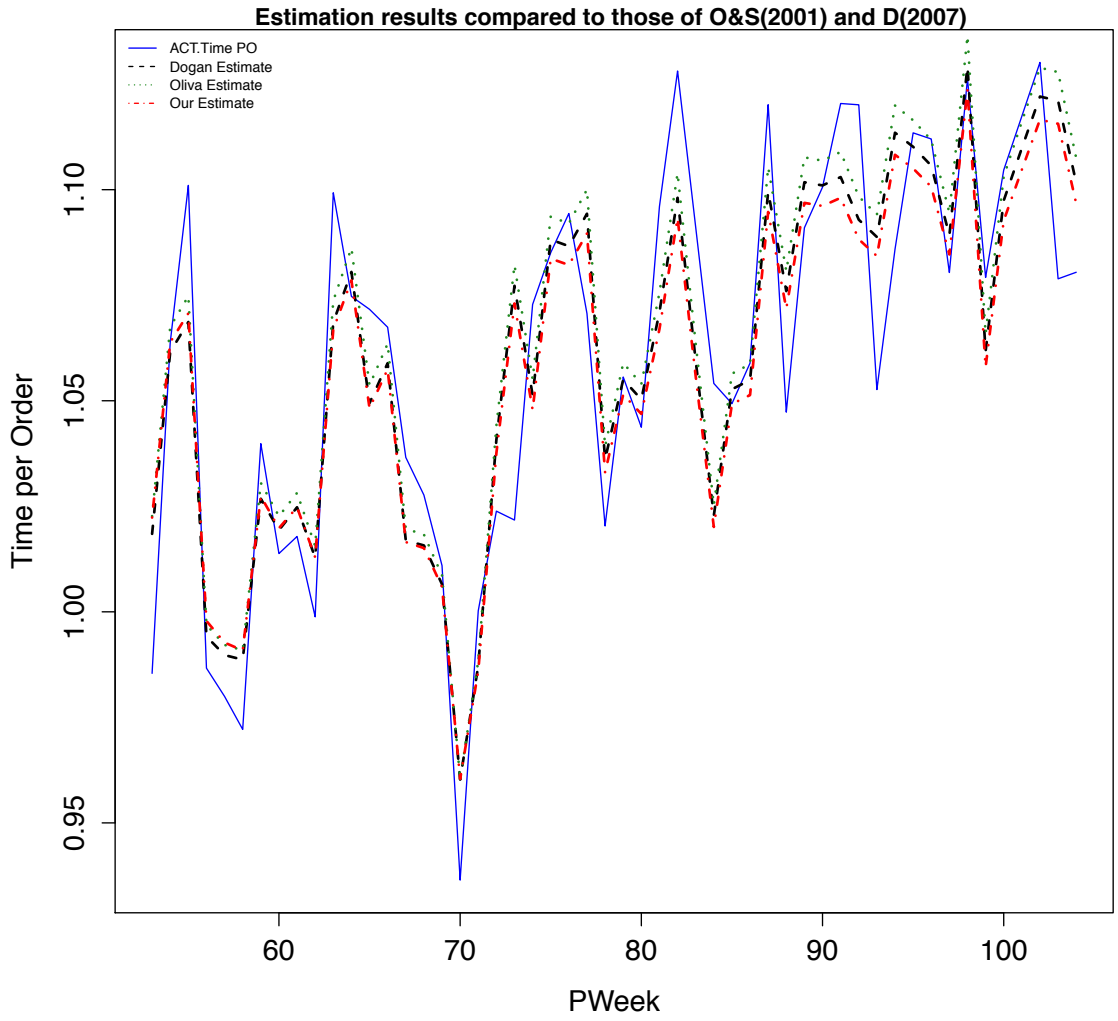

Figure A1. Comparing estimates of time per order of OS01, D07, and our estimates with the actual time per order

A4 Additional tests for normality and independence of the errors

Additional plots of residual statistics (note shown here) suggest patterns consistent with the main conclusions: the quantile distribution plot suggests a median that is approximately equal to the mean, while the qq (quantile-quantile) plot, ranking samples from the distribution against a similar number of ranked quantiles from the normal distribution, points to tails that are fatter than normal, but only for the extreme points. Likewise, plotting residuals of adjacent times, r(t) vs r(t+1), does not suggest strong first-order autocorrelation. These tests are provided in the R-script.

A5 Using MLE LR for Hypothesis testing

We can use the MLE LR ratio for hypothesis testing to examine whether, when comparing models, some can be rejected (model comparison). In the service quality example we can examine whether we can reject a simpler model that assumes the time constants for the adjustment of the standard time per task are equal. We compared the model with asymmetric norm adjustment to one in which τ_d and τ_i are constrained to be equal. Thus, in equation 6 we simply estimate τ_o (See the R-script). The resulting estimated τ_o =1.8e4, with LL=128.35. This LL value is very close to the MLE of the full model and, since with df=1 for a 95% confidence level the critical value is $3.8>2(LL_0-LL_{alt})$, we cannot reject the simpler model that the time constants are equal. Similarly, we cannot reject a simpler model that assumes both time constants to be fixed and equal to our estimated decrease-time constant ($\tau_o=\tau_d^*$=19.76), for which LL=125.33. Note that since df=2, the critical value is $6<2(LL_0-LL_{alt})$, so we can (just) reject this model. Any smaller values for τ_o are clearly rejected.

A6 Beer Game Challenge

The challenge involves a problem with moderate non-linearity in variables and parameters. Participants in the Beer Game choose how much beer to order each period in a simulated supply chain. The challenge is to estimate the parameters of a proposed decision rule for participant orders (Croson et al. 2014; Dogan 2007; Sterman 1989). Following the ordering decision rule proposed in Sterman (1989), orders O_t placed in week t are given by:[3]

$$O_t = \max\left[0, CO_t^e + \alpha\left(S' - S_t - \beta SL_t\right) + \varepsilon_t\right] \quad \text{(A6)}$$

where S_t is actual on-hand inventory, SL_t is the supply line of on-order inventory, S' is desired on-hand and on-order inventory, and CO^e_t is expected customer orders (the order that participants expect to receive next period from their immediate customer). Expected orders are given by exponential smoothing of actual incoming orders, IO:

$$CO_t^e = \gamma IO_{t-1} + (1-\gamma)CO_{t-1}^e \quad \text{(A7)}$$

The parameters to be estimated are γ, the weight on incoming orders in demand forecasting, S', the net desired on-hand and on-order inventory, α, the fraction of the gap between desired and actual on-hand and on-order inventory ordered each week, and β, the fraction of the supply line the subject accounts for.

Dogan (2007) illustrates the use of bootstrapping in the beer game using data for one participant in the experiment reported in Croson et al. (2014). We use the same data to compare MLE to bootstrapping. In the electronic supplement we provide the data set with the inventory position and ordering decision for 1 player of the beer distribution game (Appendix A0).

The challenge is to estimate the values and confidence intervals for α, β, γ, and S' for the subject for which data have been provided. Scripts are provided to help you in answering the questions. You can make use of the predefined functions provided in the electronic supplement.

A6.1 Tips for the questions

Question 1. Data and Replication Model and Estimation

- The data and estimated orders by Dogan (2007) are provided in the electronic supplement.

[3] Sterman (1989) estimates the parameter set for each individual independently. Since we work with a dataset for a single individual we omit here indices associated with individuals or teams.

Question 2. Estimation

- Do not yet produce a likelihood function. Consider what method to use. Can you use a linear optimization?

Question 4. Validity of asymptotic assumptions

- Hint for question c: if you do need to correct for autocorrelation, construct a "new" $y_{actual.corrected}$, knowing that $y_{predicted} = y_{actual} - residuals$. Then, to find the $y_{actual.corrected}$ you first have to construct residuals corrected for autocorrelation with an autoregressive (AR) model with appropriate time lags.