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

Jeroen Struben, John Sterman, David Keith

Solutions to the Challenge

In this challenge you will explore how to estimate the parameters for a decision rule for orders in the beer distribution game (Sterman 1989). This challenge involves a problem with moderate non-linearity in variables and parameters. For the R-code to replicate the solutions, see "CH1_MLE_Boot_Challenge.R". Below we summarize the steps and explain the results. As with the service quality example in the main paper we follow the process map of Figure 3.1.

CS.1 Data and Model

First we read in the data, develop the model and attempt to reproduce the Beer Game estimation results in Dogan (2007). The function *bg.model* is developed to replicate the hypothesized model structure for the decision rule using R. Next we plot the results: the actual order rate data, the Dogan (2007) estimates (provided in the excel spreadsheet), and the estimated order rate using Dogan (2007) parameters reproduced in R with our model (Figure CS.1.1).

*** FIGURE CS.1.1***

CS.1.1 Replicating the Dogan model

The model replicated in R produces results that are very close to the Dogan estimates. There is a single discrepancy in the estimated orders, in week 21, but without the source code for Dogan (2007) we could not explain the difference. Note that in the MLE estimation we do not impose rounding of orders to integers (whereas in the game they are). Doing so would make it difficult to use various estimation procedures. Other than that the results are exact and we can have confidence that we have developed an appropriate model.

CS.1.2 Observation of the data and implications for estimation

Because of the nonnegativity constraint on orders, the estimation function $f(\boldsymbol{\theta},\boldsymbol{x})$ is non-linear in parameters and cannot be solved analytically. Figure CS.1 reveals that the nonnegativity constraint on the participant's actual orders is binding for a substantial time. Further, the large range of orders, from 0 to 20 case/week, and the cycle in orders suggest the possibility of heteroscedasticity. Specifically, the model errors are likely to be smaller when actual orders are smaller and larger when orders are larger. Finally, autocorrelation is likely as order decisions may be anchored on past decisions. These likely violations of standard assumptions suggest least squares and asymptotic confidence interval methods may not be appropriate. We explore these issues below.

CS.2 Estimation

CS.2.1 Estimation and Standard Statistics

CS.2.1.1 Estimation

To produce our estimates we follow the OLS approach, which was the estimation procedure in Dogan (2007). We do not yet have to worry about the likelihood function at this point, though it is fine to do so; the results should be identical. R offers a standard non-linear least squares function *nls* that is appropriate for the least squares estimation.

Following the theory, we constrain the lower and upper bounds for the parameters to their natural limits, specifically, $0 \leq S'$ and $0 \leq \alpha, \beta, \gamma \leq 1$ (Sterman 1989; Dogan (2007).

In general, one must consider the possibility of multiple local optima. To test for this possibility we carry out the estimation using multiple starting points in the parameter space, including some close to and others far from the expected estimates. Most converged to the same values, specifically, $\alpha = 0.45760$, $\beta = 0^*$, $\gamma = 1^*$, $S' = 2.80418$, and none had lower RSS (or, similarly, SER), suggesting a single global optimum for the likelihood function. Figure CS.2.1 replicates Figure CS.1.1 but includes also our *nls* estimates, showing the estimated parameters and orders are nearly identical to the original Dogan (2007) values. This suggests that the minor deviations in the reproduction of the model are not important.

*** FIGURE CS.2.1***

CS.2.2 Summary Statistics

Next we summarize the estimation results. We use the function *nls* to provide the default statistics by running *summary*(*bg.nls.est*). We use our custom-defined function to generate additional statistics on the residuals (RSS, SER, and ESS). See Table CS.2.1.

****TABLE CS.2.1 AROUND HERE****

CS.2.3 Interpreting results

We can carry out basic hypothesis tests using the standard statistics. The main parameters of interest in this model are α, the strength of the response to inventory discrepancies, and β, the fraction of the supply line taken into account. The null hypothes that the true value of α given the data is zero is convincingly rejected ($t(44) = 6.4$, $p < 0.0000$). However, the best estimate of β *is* zero, so the null hypothesis that $\beta = 0$ cannot be rejected at any confidence level. Note, however, that the estimated standard error for β, 0.047, is small, indicating that the value of β is highly likely to be close to zero. In particular, we can strongly reject the hypothesis that the true value of β equals the optimal value of one ($t(44) = 21.2$, $p < 0.0000$). Similarly, the hypothesis that the true value of α given the data is the optimal value of one is also strongly rejected ($t(44) = 7.6$, $p < 0.0000$).

Turning to the residuals, it is straightforward to examine the fit visually, for example, plotting predicted against actual orders, or the residuals (Figure C.2.2). The residuals suggest some trend associated with the order value (consistent with the discussion of the data in Figure C.1.1), rather than independence. We explore that possibility further below.

*** Figure C.2.2.***

Note that we can also run standard statistics on Dogan (2007)'s results and compare those statistics with the ones produced in that paper. To do so, simply run the estimation forcing the parameters to take the values estimated by Dogan by using bounds identical to his estimates. We provide the *nls* estimate in the paper and leave further analysis to the reader.

CS.3 MLE Confidence Intervals

We now estimate the confidence intervals for the estimated parameters via maximum likelihood methods, using both asymptotic (AS) and likelihood-ratio (LR) based intervals.

CS.3.1 AS intervals

Confidence intervals assuming asymptotic normality (AS Wald), based on the standard error, and implying a parabolic approximation of the likelihood function in the neighborhood of the best estimates, are produced by the default function *overview()* from the *nls()* estimate. The results are produced in Table CS3.1.

We also use the manual AS method based on the Hessian, as discussed in the chapter. The statistics of the *optim* function allows calculation of the AS interval. To do so we now construct a likelihood function on which we can run *optim*.

CS.3.1.1 Likelihood function

We estimate the model assuming the error terms are iid normally distributed, in which case we can specify a convenient likelihood function. We know that under the assumption of iid normal errors the likelihood function corresponds with OLS. We will show here that results are identical.

Before examining the confidence intervals we need to specify the model in terms of the likelihood function. The likelihood function is defined in *bg.model.ll*. Note that the *optim* function generates the minimum of the function, so to find the maximum of the likelihood, we minimize its negative.

As expected, the results of the MLE estimation under the assumption of iid normal errors is identical to the estimate using *nls, overview(bg.nls.est)*.

CS.3.1.2 Manual AS

Note that the asymptotic method, due to the assumption of normally distributed errors, yields confidence intervals that are symmetric around the estimate, even for a model that is nonlinear in parameters.

CS.3.1.3 Likelihood Ratio

To construct the LR confidence intervals we now explore the curvature of the LL surface.

For the LR intervals we plot the LL function and identify the parameters for which $2\ln R \le q_k(1-\alpha)$ with $R = L(\hat{\theta}) \big/ L(\theta^*)$, as shown in the main paper. We can use our custom-made *LL.multipar.plot* function to study the univariate and profile likelihood intervals. Figure CS.3.1 shows the univariate curvature and intervals. Intervals are reported in Table CS3.1 as well. Figure CS.3.2 shows the MLE's and likelihood profile for the four parameters.

**** FIGURE CS.3.1. ABOUT HERE ****

**** TABLE CS.3.1 ABOUT HERE ****

CS.3.1.4 Interpreting the results

The example shows that LR methods, in contrast to AS methods, can handle more complex situations. Unlike the asymptotic method, the likelihood functions and resulting confidence intervals are not symmetric and, in some cases, not parabolic. For example, the likelihood function for γdrops off steeply for $\gamma < 0$. Such values would, nonsensically, imply that recent orders are given negative weight in the demand forecast. Also note that, while the MLE is constrained to the admissible regions ($0 \le$ S' and $0 \le \alpha, \beta, \gamma \le 1$), the LR intervals are not. The example thus illustrates that the LR method can detect and correctly capture important non-linearity in parameters that the AS method, which by construction assumes the symmetric parabolic approximation around the MLE, does not.

Asymmetry in confidence regions is more likely when parameters interact. The surface of two parameters of particular interest, α and β, (Figure CS.3.3), also produced with the *LL.multipar.plot* function, shows how the confidence interval for α (the gain of the negative feedback that corrects inventory discrepancies) interacts with the value of β (the fraction of the supply line taken into account), increasing as β moves further away from zero.

**** FIGURE CS.3.3 ABOUT HERE ****

Asymmetries in the confidence intervals illustrate the differences between LR and AS methods. These results highlight that relying solely on just AS or univariate LR intervals may be problematic.

However, in this case, the differences in assumptions across methods have little impact on the confidence intervals (Table CS.3.1). Irrespective of the approaches, for example, we may conclude that the value of β is highly likely to be close to zero, and strongly reject the hypothesis

that $\beta = 1$, that is, we can be highly confident that the subject in this example ignored the supply line of unfilled orders. Thus, at least in this example, the second order (parabolic) approximation of the likelihood function in the neighborhood of the best estimates is reasonable.

CS.4 Validity of asymptotic assumptions

So far we have, optimistically, assumed normality and independence of the errors (no autocorrelation in the errors). We now assess the appropriateness of these assumptions by analyzing the residuals. The R function *nlsResiduals* provides, by default, four classic plots of residuals (Delignette-Muller and Baty 2012): non-transformed residuals against fitted values, standardized residuals against fitted values, auto-correlation plot of residuals ($i+1^{th}$ residual against i^{th} residual), and qq-plot of the residuals. Of particular interest are normality and autocorrelation.

CS.4.1 Normality

The histogram plot of the residuals (Figure CS.4.1), obtained by selecting nlsResiduals index 5, suggests a distribution that deviates from normality.

**** FIGURE CS.4.1 ABOUT HERE ****

Other basic residual statistics plots (Figure CS.4.1) indicate similar patterns: the quantile distribution plot suggests a median that is smaller than 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.

**** FIGURE CS.4.2 ABOUT HERE ****

To confirm what the plots of the residuals suggest we carry out the standard tests for normality, specifically the Shapiro-Wilk test. The analysis strongly rejects the hypothesis that the residuals follow are normally distributed (p = 0.0007613 (in similar fashion we find p = 0.0006047 for the Dogan (2007) model).

CS.4.2 Autocorrelation

The residual statistics (Figure CS.4.2. bottom left) also suggest autocorrelation. To assess whether the residuals show significant autocorrelation we plot the autocorrelation by lag (Figure CS.4.3). The autocorrelation in residuals with lag k = 1 is high.

The test for any autocorrelation involves rejecting autocorrelation for any lag. This combined tests requires a smaller significance level that one desired for the overall multiple test (Barlas 1994). Following Dogan (2007), using 0.01 we can reject the null hypothesis that the residuals are not autocorrelated if $\pm z^{(0.995)}$ z(0.995) and |t(k)| >2.58 for at least 1 of the lagged elements. The formal analysis shows that the autocorrelation is significant at the 5% level only for lag k = 1, suggesting a first-order autoregressive (AR[1]) error process (Hamilton 1994).

CS.4.3 Preparing for bootstrapping

To deal with the non-normality of the errors one could construct a more appropriate MLE, or use bootstrapping. We focus on bootstrapping. Given the characteristics of the error-distribution (non-normality), we use non-parametric residual-based bootstrapping. Alternatively one could proceed with parametric bootstrapping, using an approximation to the empirical residual distribution to generate the errors (including the heteroscedasticity, and with a lower bound). Reshuffling the observed error-terms is inappropriate in the presence of autocorrelation unless that is also corrected.

To prepare for bootstrapping, we (i) remove the first-order autocorrelation from the observed error-terms; (ii) examine whether the standard deviation of the adjusted error-terms differs importantly from that of the observed error-terms; (iii) if needed, correct the standard deviation so that it is equal to that of the observed error-terms; (iv) test that autocorrelation is not significant in the corrected error-terms after the standard deviation correction; finally, (v) re-estimate the model, as described above, with the ***y***'s adjusted for the corrected but not-resampled error-terms and examine the new parameters and error-terms to test that the corrections above have worked. The process is carefully described and tools to analyse this are provided in the R-script.

CS.4.4 Adjusting for autocorrelation

We conclude that an AR[1] process is an appropriate correction. Figure CS.4.4 shows the histogram and autocorrelation for the adjusted process. More formally, we cannot now reject the hypothesis that there is no autocorrelation in the residuals. The Shapiro-Wilk test, however, shows that we still must reject normality (p = 0.002).

Table CS.3.1. also shows the parameter estimates and AS/LR confidence intervals for the corrected dataset (ACT.ORD.corr).

CS.5 Bootstrapping

CS.5.1.1 Bootstrap results

In this analysis, because the solution can be found using nls, we can use a standard bootstrapping function, specifically designed for non-linear least-squares, *nlsBoot*. *nlsBoot* uses non-parametric bootstrapping with mean centered residuals. For each new data set the original non-linear regression model is fitted and the resulting parameters stored. Bootstrap estimate distributions of this function can be visualized using the function *plot.nlsBoot* either by plotting the bootstrap sample for each pair of parameters or by displaying the boxplot representation of the bootstrap sample for each parameter. Our Table CS3.1 reports the result (for reference we also report the uncorrected results).

The nlsBoot approach, with 2000 iterations, did not converge 621 times. Mean estimated values for ($\alpha, \beta, \gamma, S'$) were (0.46, 0, 1, and 3.1) and (0.43, 0, 1, 3.49) for the uncorrected and corrected data respectively. We can also perform boot.ci (as in the application). All results are provided in Table CS.3.1

CS.5.2 Suggestions for follow up

As a follow up: (i) implement multiple starts in the automated boot; (ii) see if you can generate the same results with the manual boot (for both see the service quality example in the main chapter).

CS.6 Interpreting the Results

The differences in confidence intervals across the different methods are minor and have no impact on key questions such as whether the decision maker took the supply line into account. The largest differences across the methods are in the estimated confidence intervals for γ. These differences are not surprising given that the likelihood function for γ(Figure 4.1) shows the greatest departure from the parabola assumed by AS methods, has the widest confidence intervals, and because the first-order smoothing used to form the expectation of incoming orders affects the degree of autocorrelation.

**** TABLE 4.2 AROUND HERE ****

Overall, even though the model and data violate maintained hypotheses of MLE including independence and normality of the errors, MLE provides estimates of the parameters and confidence intervals around them consistent with the bootstrapping estimates.

Regarding the procedures, while bootstrapping, including correction for autocorrelation, is fairly straightforward, the process involves several manual steps that require judgment, and can be tedious. In our case the reestimated parameters are close to the original values while the autocorrelation disappears. If bootstrapping were not feasible, and one concluded that it was necessary to deal with non-normality of the errors, one could construct a more appropriate MLE, and simulate with an error generation process that captures the observed distribution (Train, 2003).

Note: The beer game example, though nonlinear, is relatively simple and analogous to typical regression studies in that the estimation problem did not involve an explicit feedback system. The beer game is, of course, a complex system with multiple feedbacks, but because the data are collected in the context of an experiment, all the explanatory variables needed to estimate the decision rule for orders—incoming orders, inventory, the supply line—are directly measured. Few systems of interest in dynamic modeling offer such complete data.

Tables and Figures

Table CS.2.1 – Estimation results, beer distribution game example (95% level)

Parameter	Estimate	Std. Error	t value	Pr(>|t|)	Estimate (Dogan 2007)
α	0.45760	0.07130	6.418	8.2e-08 ***	0.5
β	0.00000	0.04724	0.000	1.0000	0.01
γ	1.00000	0.58425	1.712	0.0940 .	0.95
S’	2.80418	1.59008	1.764	0.0848 .	1.96
RSS/SSE	294.932 on 44 degrees of freedom				302.1
SER	2.5891				2.6202
ESS	573.0907				572.7023

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table CS.3.1 Comparative confidence intervals, beer game example (95% level)[1]

Method/Parameter	duration	α	β	γ	S'
Uncorrected Error terms					
MLE Estimate		**0.458**	**0**	**1**	**2.80**
Confidence Interval, AS					
• Wald test	negligible	(0.31, 0.60)	(-0.10, 0.10)	(-0.18, 2.18)	(-0.4, 6.0)
• Manual, Hessian-based	negligible	(0.32, 0.59)	(-0.09, 0.09)	(0.31, 1.69)	(-0.25, 5.9)
Confidence Interval, LR					
• Univariate *df=1*	7 sec	(0.34, 0.60)	(-0.07, 0.05)	(0.46, 1.72)	(0.69, 4.65)
• Profile *df=1*	1 min	(0.33, 0.76)	(-0.14, 0.07)	(0.48, 1.81)	(-1.31, 4.89)
• Bivariate (α,β) *df=2*	15 sec	(0.29, 0.70)	(-0.12, 0.09)	(0.46, 1.72)	(0.69, 4.65)
Confidence Interval, Boot (2000 resamplings)					
• nlsBoot (mean-centered)	4.5 min	(0.33, 0.71)	(0*, 0.09)	(0.02,1*)	(0.90, 6.03)
• nlsBCI	4.5 min	(0.34, 0.67)	(0*, 0.08)	(0.02,1*)	(0.93, 5.74)
AR(1) Corrected					
MLE Estimate		**0.429**	**0**	**1**	**3.12**
Confidence Interval, AS					
• Wald test	negligible	(0.30, 0.56)	(-0.1, 0.10)	(-0.18, 2.18)	(-0.12, 6.4)
• Manual, Hessian-based	negligible	(0.31, 0.55)	(-0.09, 0.09)	(0.25, 1.75)	(0.10, 6.1)
Confidence Interval, LR					
• Univariate *df=1*	7 sec	(0.31, 0.55)	(-0.07, 0.06)	(0.44, 1.74)	(1.14, 5.1)
• Profile *df=1*	1 min	(0.33, 0.90)	(-0.1, 0.12)	(0.60, 1.87)	(-0.04, 5.4)
• Bivariate (α,β) *df=2*	15 sec	(0.27, 0.82)	(-0.11, 0.13)	(0.44, 1.74)	(1.14, 5.1)
Confidence Interval, Boot (2000 resamplings)					
• nlsBoot (mean-centered)	4.5 min	(0.32, 0.60)	(0*, 0.09)	(0.04,1*)	(1.12, 6.26)
• normal	6 min	(0.26, 0.57)	(-0.07, 0.03)	(0.63,1.8)	(-0.06, 4.70)
• percentile	6 min	(0.32, 0.62)	(0*, 0.09)	(0.01,1*)	(1.7, 6.50)
• BCa	6 min	(0.30, 0.57)	(-)	(0.03, 1*)	(0.03, 4.65)
• nlsBCI	4.5 min	(0.32, 0.63)	(0*, 0.09)	(0.016, 1*)	(1.17, 6.13)
AR(1) Corr, Dogan (2007)					
OLS Estimate		**0.5**	**0.01**	**0.95**	**1.96**
Confidence Interval, Boot (1000 resamplings)					
• Percentile	NA	(0.27, 0.92)	(0, 0.14)	(0.02, 1)	(0, 7.19)
• BCI	NA	(0.26, 1.00)	(0, 0.14)	(0.44, 1)	(0, 6.05)

[1] Confidence bounds indicated by * means that the admissible values were reached and not overruled by the particular method.

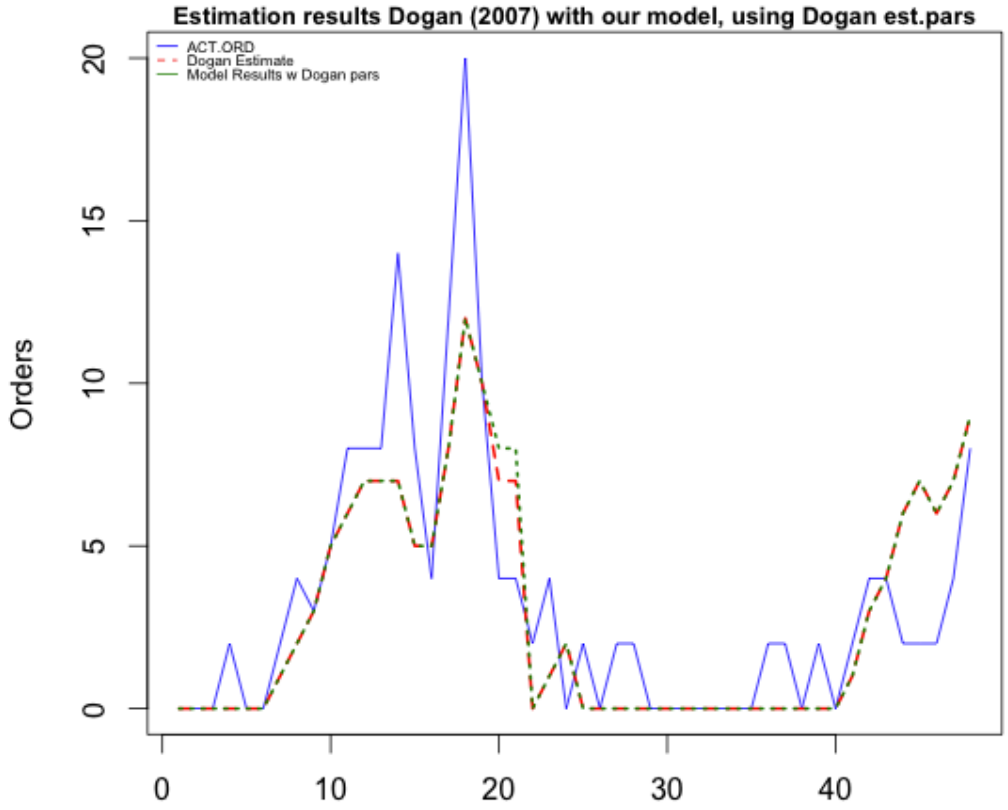

Figure CS.1.1 – Model results using Dogan (2007) parameter estimates

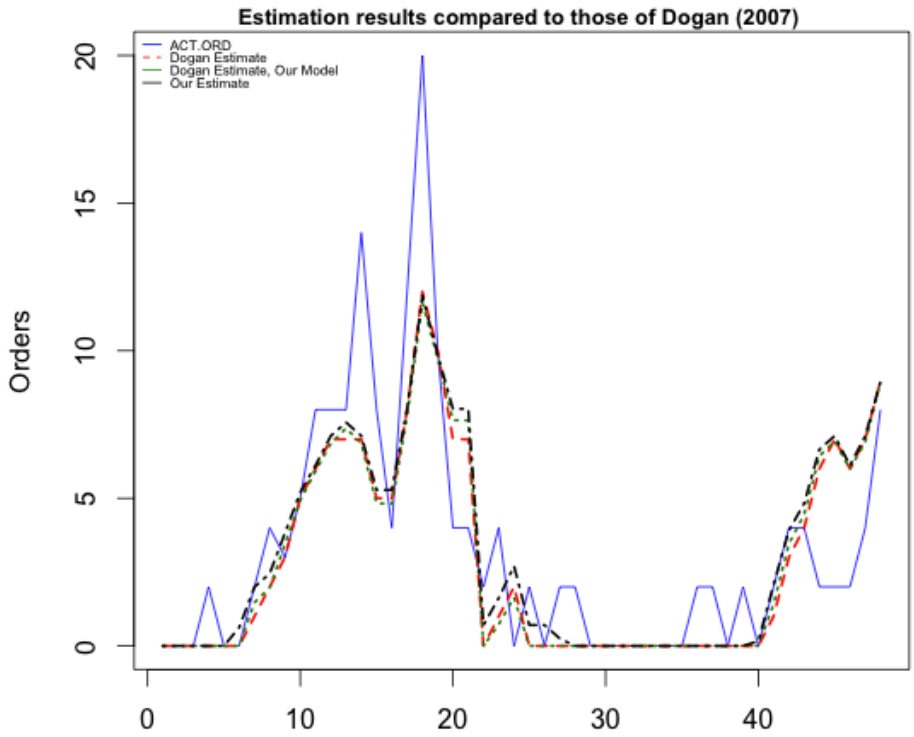

Figure CS.2.1 Actual orders compared to estimated orders and to those in Dogan (2007).

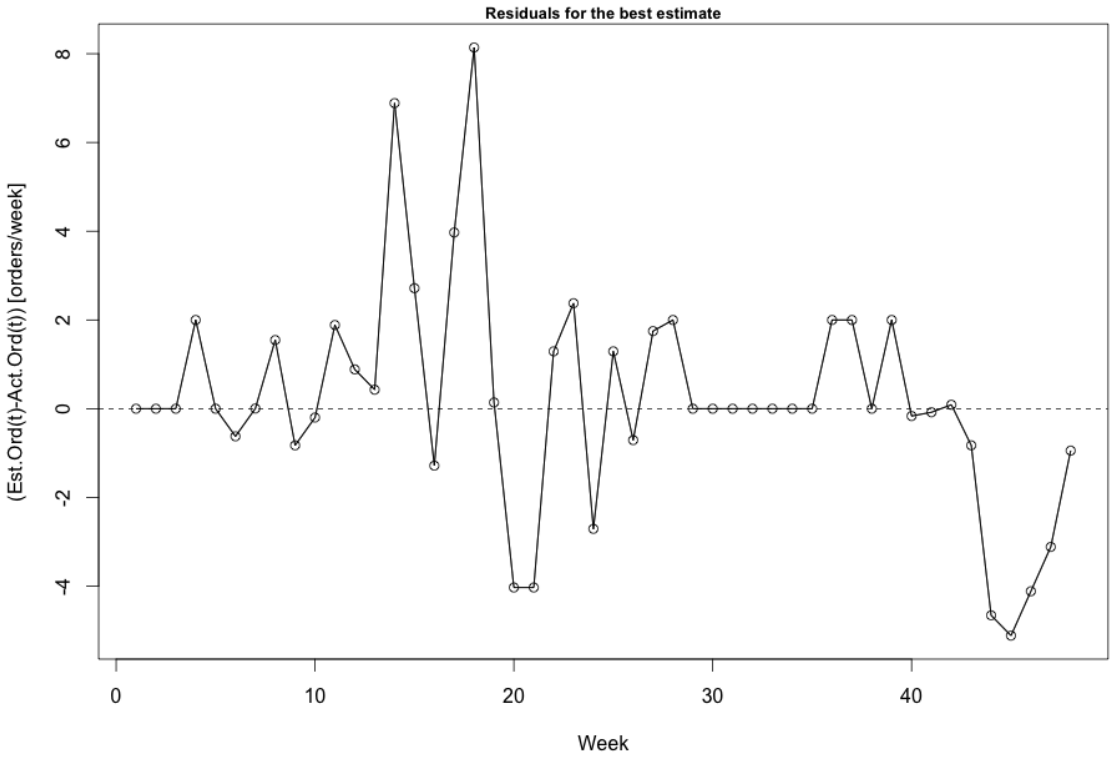

Figure CS.2.2 Residuals for our best estimate using nls.

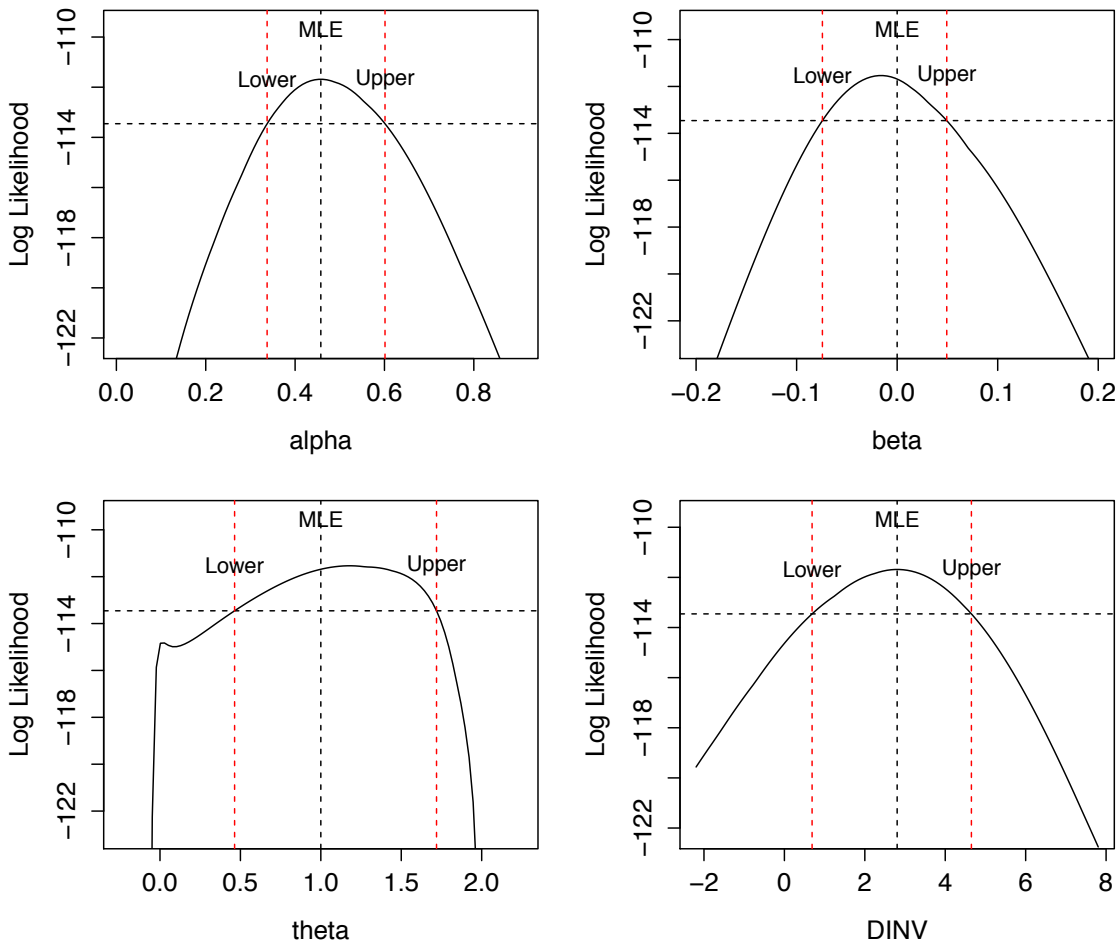

Figure CS.3.1. Log likelihood function for the four parameters (holding others at MLE) and 95% univariate confidence level for the beer game example.[2]

[2] Where the MLE differs from the peak (as in beta and theta) the estimates were constrained to fall within the natural region of operation (being 0 and 1 for all parameters except desired inventory, which had an infinite upper-limit. While the MLE is constrained, the likelihood profile is not so.

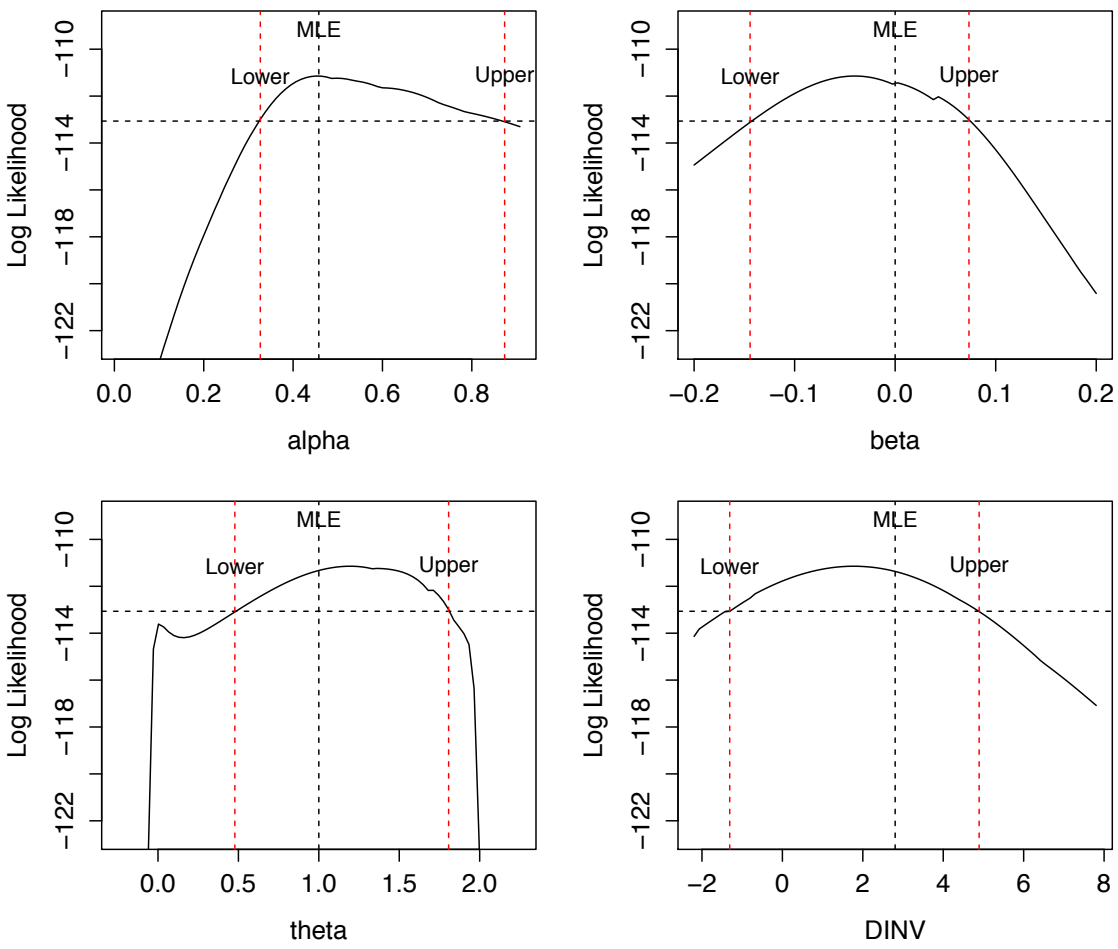

Figure CS.3.2. Profile likelihood for the four parameters and 95% confidence level for the beer game example.

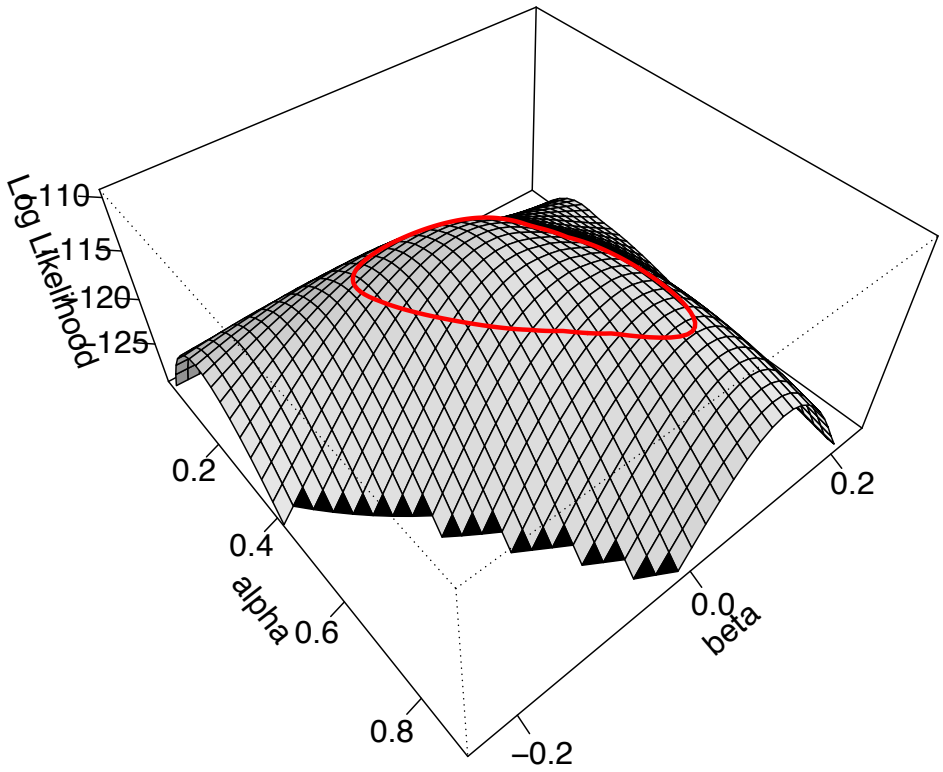

Figure CS.3.3. Bivariate confidence interval for α and β. (Other parameters remain at their estimated value)

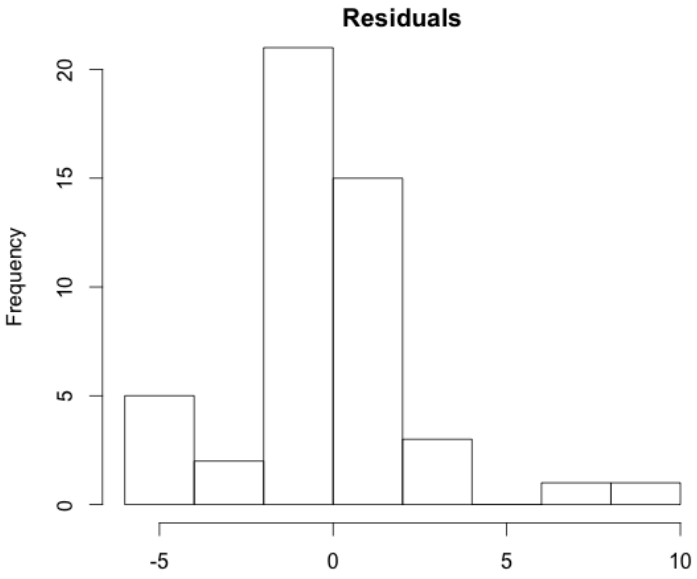

Figure CS.4.1. Quantile distribution of residuals

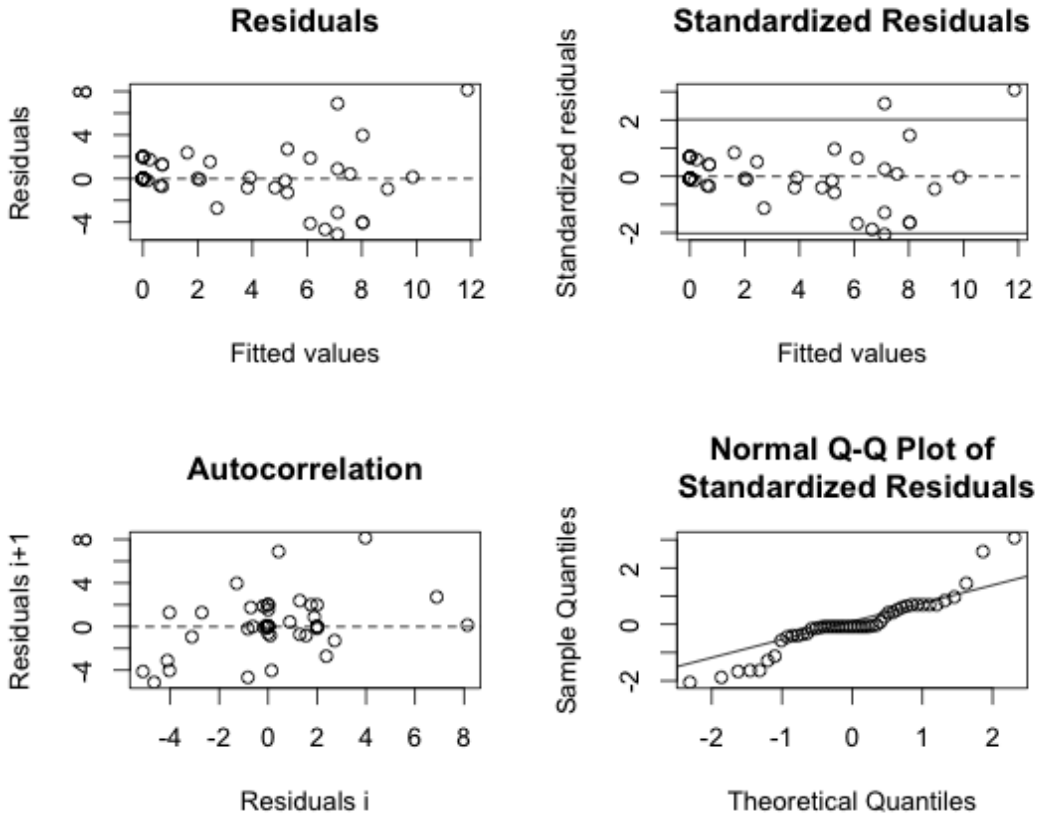

Figure CS.4.2 Standard plots of residuals using *nlsResiduals*

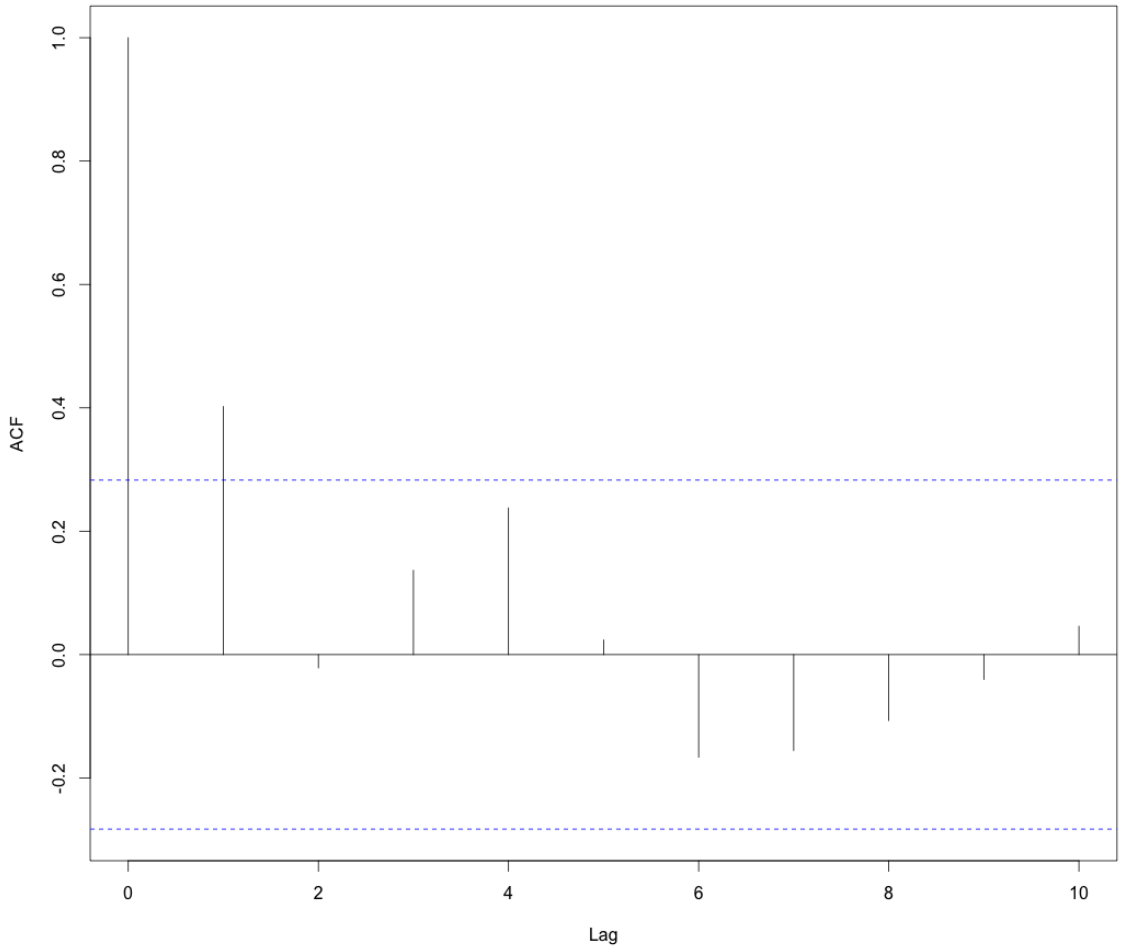

Figure CS.4.3. Autocorrelation, by lag

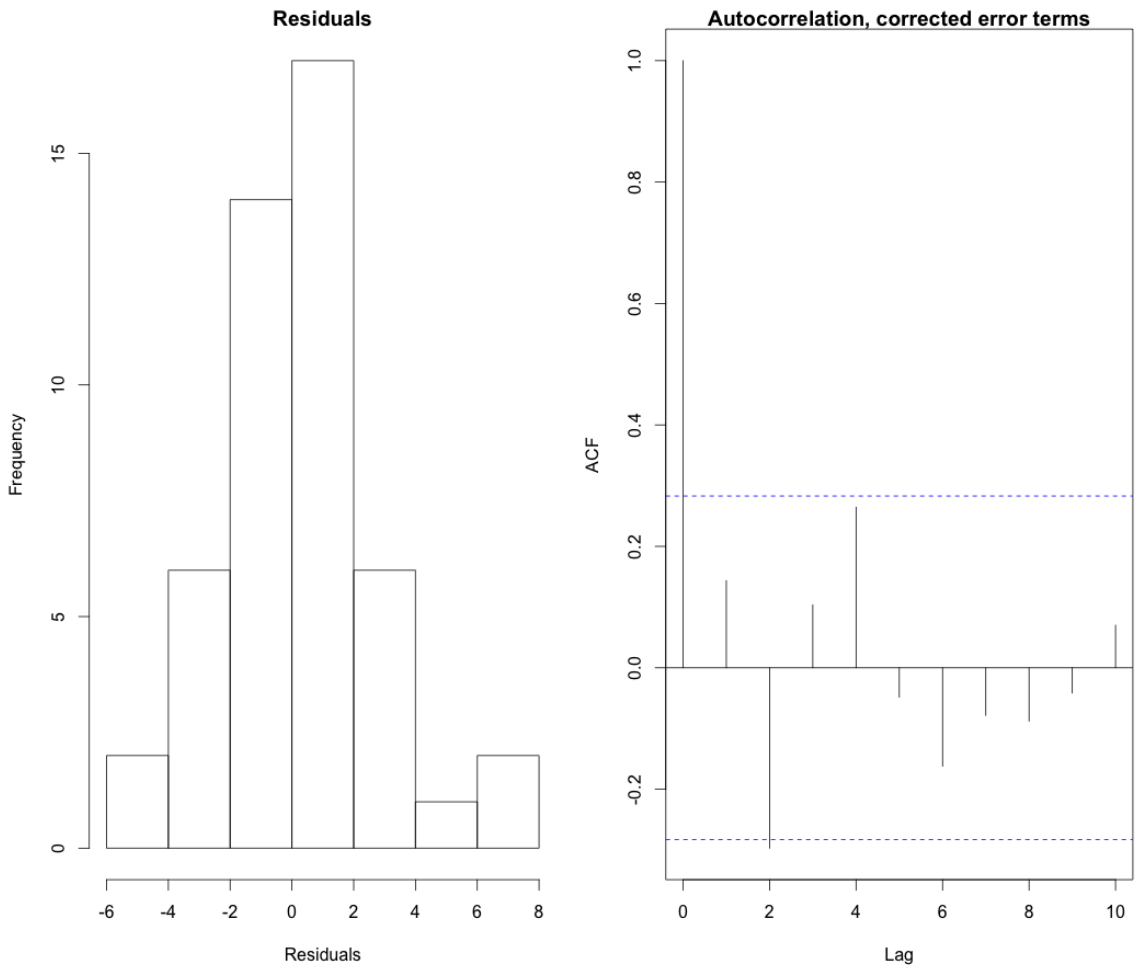

Figure CS.4.4. Quantile distribution of error terms and autocorrelation after AR(1) correction

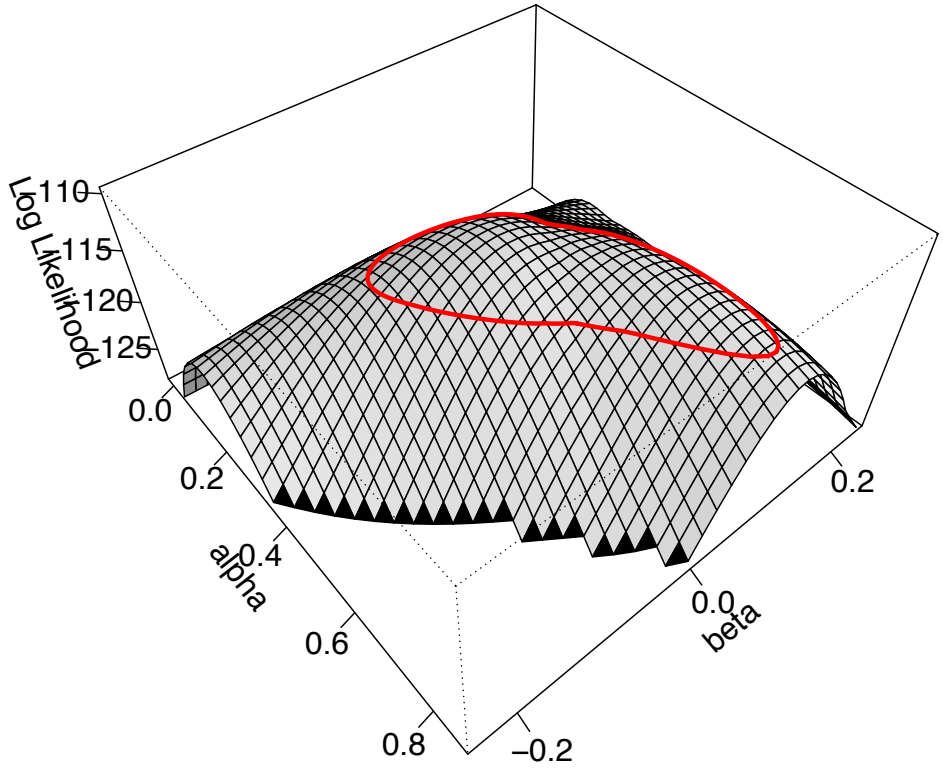

Figure CS.4.5 Bivariate confidence interval for α and β for autocorrelation-corrected data (Other parameters remain at their estimated value)

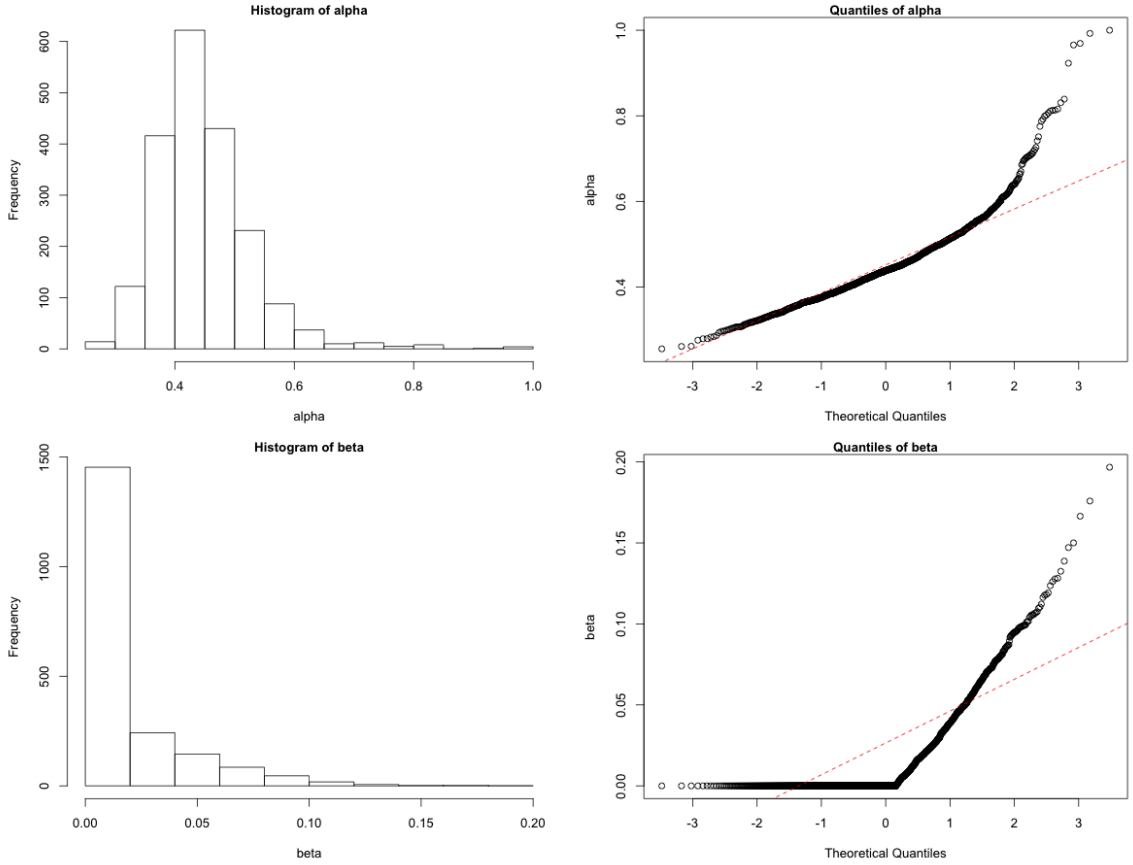

Figure CS.5.1. Histogram and quantile distribution of α and β, for AR(1) corrected data.

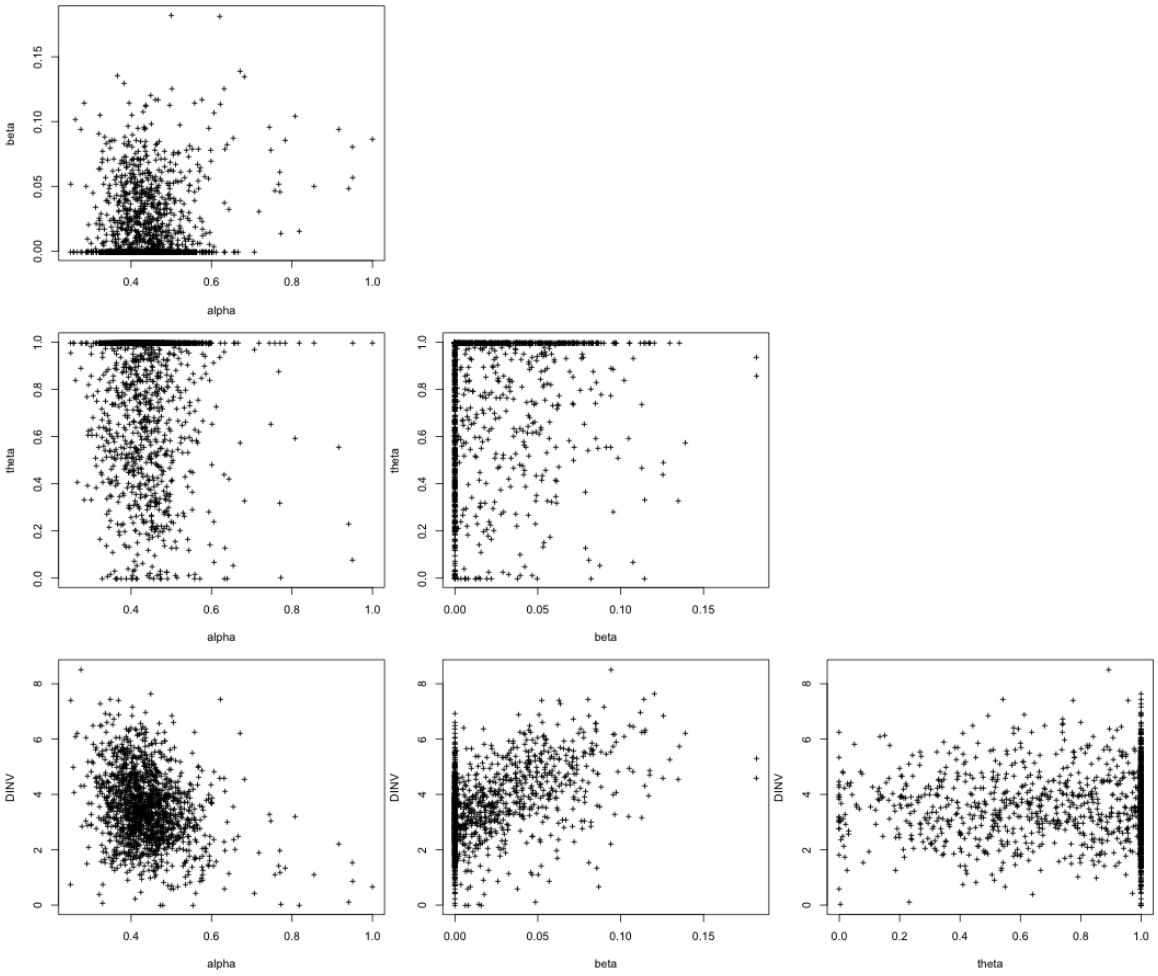

Figure CS.5.2 Pairwise plotting for the mean-centered *plot(nlsBoot)* bootstrapped parameter estimates of the autocorrelation-corrected data