Overlapping Observations in Credit Risk Models

Dobromił Serwa[*][♣]

[♣]SGH Warsaw School of Economics, Poland

Submitted: December 29, 2022 • Accepted: December 31, 2022

ABSTRACT: Parameters in logistic regression models for probability of default are typically estimated using the maximum likelihood method. The aim of this paper is to verify whether the use of overlapping observations improves precision or causes deterioration of estimation results in these models. Our Monte Carlo simulations demonstrate that the difference between parameter estimates using all overlapping observations in a sample and only non-overlapping observations in a reduced sample is not statistically significant, but the variance of parameter estimates is reduced when overlapping observations are used.

JEL classification: C13, C23, C25, E51.

Keywords: PD Model, Credit Risk, Overlapping Observations, Logistic Regression.

1 Introduction

In this paper, we analyze the impact of overlapping observations on parameter estimates in credit risk models. Overlapping observations are different observations that contain similar (overlapping) information. It is well known that overlapping observations are correlated and in general violate the assumption of independent observations required by many estimation methods.

In statistical models predicting credit risk, observations of the target variable often overlap by construction. The target variable identifies an event of default in a loan contract

[*]Corresponding Author. E-mail: dserwa@sgh.waw.pl

DOI: 10.2478/erfin-2022-0007

over the annual horizon. At the same time, the model uses monthly observations to predict such an event, which means that consecutive months provide similar information regarding the risk of default. Consecutive observations of the target variable overlap and are correlated.[1]

Numerous statistical techniques have been proposed to deal with overlapping data in estimation, prediction, and testing (e.g., Hansen and Hodrick, 1980; Boudoukh et al., 2020; Zwetsloot and Woodall, 2021; Xu, 2021). Previous research focuses on macroeconomic and financial market variables applied in linear regression models. For instance, demand for goods, foreign exchange rates, interest rates, stock returns are modelled and predicted using the overlapping observations (Chinn, 2006; Darvas, 2008; Boylan and Babai, 2016; Lioui and Poncet, 2019).

This research reveals that the use of overlapping observations instead of the non-overlapping observations helps reduce the variance of parameter estimates (e.g., Richardson and Smith, 1991; Britten-Jones et al., 2011) and takes into account complex cyclical patterns of high-frequency observations (e.g., Rossana and Seater, 1995; Mamingi, 2017).

In contrast to the studies of financial returns, very few analyses related to overlapping observations have been performed with nonlinear credit risk models (e.g., Kiesel et al., 2001; Cipollini and Fiordelisi, 2012; European Banking Authority, 2017; Frankland et al., 2019). We attempt to fill this gap by verifying the hypothesis that different parameter estimates in terms of mean and variance are generated using the non-overlapping and overlapping observations in logistic regression models.

The Monte Carlo simulations performed show that the difference between parameter estimates using all overlapping observations in a sample and only non-overlapping observations in a reduced sample is not statistically significant, but the variance of parameter estimates is reduced when overlapping observations are used.

2 Problem of overlapping observations

Consider a portfolio of bank loans which consists of a number of individual risky exposures. Let $D_{n,t}$ denote the event where exposure n defaults at time t. In turn, $ND_{n,t}$ denotes the event where the exposure does not default at time t. When annual data are applied then t denotes a specific year and an appropriate statistical model can be used to predict a possible default occurring next year or over longer time horizons. Such predictions can also be used to assess annual default rates of an aggregate portfolio, an important measure of loan portfolio quality.

When higher frequency data are used to measure credit risk, e.g. quarterly or monthly observations, the problem of overlapping observations arises. For instance, monthly obser-

[1] There is also a branch of literature devoted to another type of data overlap in a logistic regression, where no specific parameter vector exists for a given dataset that correctly allocates all observations to their group (e.g., Albert and Anderson, 1984, and the references therein).

vations are frequently used to predict defaults happening over the forthcoming 12 months. This means that predictions of such defaults do not change significantly between consecutive months, i.e. observations are correlated between each other.

Let $D_{n,t+1}^{t+12}$ denote the event of at least one default taking place in the period between $t+1$ and $t+12$ for exposure n and $ND_{n,t+1}^{t+12}$ denote the event of no default of exposure n between $t+1$ and $t+12$. If it is additionally assumed that only one default can take place in a 12-month period then any default in the period between $t+1$ and $t+12$ is a union of 12 mutually exclusive (disjoint) events defined as "exposure going into default at time $t+i$", where $i = 1, 2, ..., 12$. Therefore, the sum of probabilities of these disjoint events gives

$$Pr(D_{n,t+1}^{t+12}) = Pr(D_{n,t+1}) + Pr(D_{n,t+2}) + \ldots + Pr(D_{n,t+12}). \tag{1}$$

In some loan portfolios, e.g. mortgage portfolios, multiple defaults of a customer in a 12-month period are unlikely, but possible. In such portfolios, the following inequality holds

$$Pr(D_{n,t+1}^{t+12}) \leq Pr(D_{n,t+1}) + Pr(D_{n,t+2}) + \ldots + Pr(D_{n,t+12}). \tag{2}$$

In the empirical analysis presented in the next section, multiple defaults in a 12-month period are allowed. However, it is more convenient to assume disjoint default events in the theoretical derivations presented here. Moreover, the formula (1) is a good approximation for most credit exposures in practice.

Both D_{t+i} and ND_{t+i} are conditional on a vector of explanatory variables (risk drivers) $\mathbf{X}_t$ measured at time t and some parameter vector θ. The 12-month probability of default, conditional on $\mathbf{X}_t$ and the parameter vector θ, calculated at time t equals

$$Pr(D_{n,t+1}^{t+12}|\mathbf{X}_t, \theta) = 1 - Pr(ND_{n,t+1}^{t+12}|\mathbf{X}_t, \theta). \tag{3}$$

The mutually exclusive defaults $D_{n,t+i}$ and $D_{n,t+j}$ for some $i \neq j$, conditional on some risk drivers $\mathbf{X}_t$, are still disjoint events. Therefore, the following formula is valid as well,

$$Pr(D_{n,t+1}^{t+12}|\mathbf{X}_t) = Pr(D_{n,t+1}|\mathbf{X}_{n,t}) + Pr(D_{n,t+2}|\mathbf{X}_{n,t}) + \ldots + Pr(D_{n,t+12}|\mathbf{X}_{n,t}). \tag{4}$$

For a given exposure n, two chronologically consecutive observations $D_{n,t}$ and $D_{n,t+1}$ are disjoint events, but the events $D_{n,t}^{t+11}$ and $D_{n,t+1}^{t+12}$ are not disjoint but rather strongly correlated. The latter events are usually correlated because there is an overlap of 11 periods (e.g. months) between them. This can be easily observed when one looks at the sums of probabilities for the events $D_{n,t}^{t+11}$ and $D_{n,t+1}^{t+12}$ in formula (1). In practice, clustered observations of $D_{n,t+1}^{t+12} = D_{n,t+2}^{t+13} = \ldots = D_{n,t+12}^{t+23} = 1$ are observed for the consecutive periods preceding the default time $t+12$ in the estimation sample. Thus, observations are clustered in groups of zeros and ones.

This correlation of consecutive observations may in general cause estimation problems when the models used to predict $D_{n,t+1}^{t+12}$ assume independent observations. For example, the maximum likelihood method used to estimate parameters of a logistic regression explaining $D_{n,t+1}^{t+12}$ typically assumes that observations of $D_{n,t+1}^{t+12}$, conditional on the set of risk drivers $\mathbf{X}_t$, are independent.

One commonly applied approach to deal with overlapping observations is to use only non-overlapping observations in the estimation sample. For instance, one can use every twelfth monthly observation (e.g. only observations from December each year) for each exposure in the estimation sample. The drawback of this approach might be that the precision of estimates may be low in a reduced sample.

In this paper, we investigate if the differences between estimates from the reduced sample are much different from those in the full sample in the datasets of the size typically observed in banking loan portfolios.

3 Simulating datasets with overlapping observations

We assume that the target variable $D_{n,t+1}^{t+12}$ is explained by a set of three independent variables, namely one “aplication” variable X^A, one “behavioral” variable X^B, and one “macroeconomic” variable X^M. The application variable does not change its value in time but has a different value for different exposures. The behavioral variable changes its value over time and for different exposures. Finally, the macroeconomic variable takes on common values for all exposures but changes its value over time.

The values of explanatory variables are generated from the standard normal distribution truncated from above at the level of 2. We have decided to truncate upper tails of these distributions in order to limit predicted values of $Pr(D_{n,t+i}|\mathbf{X}_{n,t}, \theta)$. For instance, in the special case when there are equal probabilities that a default happens during any one of the next 12 months, conditional on the current information, $Pr(D_{n,t+1}|\mathbf{X}_{n,t}) = Pr(D_{n,t+2}|\mathbf{X}_{n,t}) = \ldots = Pr(D_{n,t+12}|\mathbf{X}_{n,t})$, it can be concluded that $Pr(D_{n,t+1}^{t+12}|\mathbf{X}_t) = 12 \times Pr(D_{n,t+1}|\mathbf{X}_{n,t})$. In this case $Pr(D_{n,t+i}|\mathbf{X}_{n,t})$ should not exceed $1/12$ for any $i = 1, 2, ..., 12$. The choice of a normal distribution from which the variables are randomly generated is not particularly important. We assume that similar results can be obtained when other distributions, e.g. distributions with fat tails, are considered. The analysis of such distributions would be complementary to this experiment.

Next, we construct a shock variable $s_{n,t}$ that can take two values. First, $s_{n,t} = S$ if a shock affecting exposure n takes place at time t and potentially causes its default within the next 12 months, i.e. between $t + 1$ and $t + 12$. Second, $s_{n,t}$ equals zero otherwise. The values of S are observable in clusters of up to 12 observations (i.e. 12 months or less) before a potential default. The values of a shock variable affect risk drivers X^A, X^B, and X^M in the way that we explain below.

The shock can be of a macroeconomic type or a microeconomic type. During a macroeconomic shock the value of the variable X^M is increased by S, but only for a single randomly selected exposure. Only selected exposures affected by macroeconomic shocks is a strong assumption, but it is done for two reasons. The first rationale is to limit the number of defaults due to macroeconomic reasons in the simulation sample. The second reason is that different exposures may be more or less robust to macroeconomic shocks and not all exposures need to default due to such shocks.

During a microeconomic shock the value of the variable X^B is increased by S, but only for a single randomly selected exposure. In addition, the value of the application variable X^A is also increased by S for the latter exposure.

For simplicity, equal probabilities (0.5) are set for the selection of micro and macroeconomic shocks. The reason for introducing special shocks in the model is to increase values of respective variables in such a way that these variables significantly affect the risk of default.

This way of simulating the values of explanatory variables addresses the potential problem of risk drivers capturing characteristics of defaulted obligors and not being relevant for the whole portfolio. The data are simulated from homogeneous distributions and the values of respective variables are particularly high when defaults happen in each of the simulations.

Since the shocks are time independent and random, the volatility of specific variables does not depend on time. The purpose of such an approach is to avoid investigating the impact of changing volatility of variables on parameter estimates here. Naturally, the changing volatility would affect precision of estimates and would add an unnecessary dimension to the analysis.

The multicollinearity of explanatory variables in this simulation depends on the number time-series observations and the frequency of shocks in the sample, because (a) the truncated normal distributions, from which the respective variables are generated, are independent, (b) X^A and X^B are correlated only due to the fact that default shocks affect the same customers, (c) correlation between X^A and X^B becomes smaller when the number of time-series observations is increased. Nevertheless, the VIF statistics computed for 3 risk drivers were always close to 1 in our simulations. In practice, multicollinearity of variables may affect precision of parameter estimates, but we have decided to leave this research area for future investigations.

We assume that the model explaining $D_{n,t+1}^{t+12}$ is a logistic regression

$$Pr(D_{n,t+1}^{t+12}|\mathbf{X}_t) = \frac{exp(\theta_0 + \theta_1 \cdot X_{n,t}^A + \theta_2 \cdot X_{n,t}^B + \theta_3 \cdot X_{n,t}^M)}{1 + exp(\theta_0 + \theta_1 \cdot X_{n,t}^A + \theta_2 \cdot X_{n,t}^B + \theta_3 \cdot X_{n,t}^M)}. \tag{5}$$

In statistical experiments, the defaults are typically randomly selected by checking if $\theta_0 + \theta_1 \cdot X_{n,t}^A + \theta_2 \cdot X_{n,t}^B + \theta_3 \cdot X_{n,t}^M + u_{n,t} > 0$, where $u_{n,t}$ is a random independent variable drawn from the logistic distribution. However, by doing so, it would be impossible to generate clustered observations of $D_{n,t+1}^{t+12} = 1$ for consecutive periods t.

Instead, shock events are randomly and independently drawn from the Bernoulli distribution with a given success parameter (denoting average probability of default by a respective exposure in a given period) for each observation. Then, dates of default events are set as the respective dates of the shock events. The consequence is that $D_{n,t+1}^{t+12} = 1$ if $s_{n,t} = S$ and $D_{n,t+1}^{t+12} = 0$ if $s_{n,t} = 0$, and the observations of $D_{n,t+1}^{t+12} = 1$ are clustered in the groups of 12 observations (with some exceptions when defaults happen in the first 12 months of the sample or in the last 12 months of the sample).

In the next step, the estimation sample is reduced by removing the last 12 months of observations for each exposure. The last year is not representative to the rest of the sample, because it does not take into account potential defaults that may happen in the next year after the sample end date.

Finally, the parameters θ_0, θ_1, θ_2, and θ_3 are estimated using all the simulated observations and compared with the estimates of the same parameters obtained in the sample reduced by selecting only non-overlapping observations. In the latter reduced sample, there is only every twelfth observation for each exposure from the full sample.

4 Empirical results

In the main simulation, we assume that there are 10000 exposures in a loan portfolio and the number of investigated months for each exposure is equal 120 (10 years). The frequency of shocks (and consequently defaults) is set to 0.01, which is a relatively high number corresponding to the 12% annual default rate. In such a portfolio shocks are frequent and many clusters of observations with $D_{n,t+1}^{t+12} = 1$ are observable. This is clear evidence of correlated observations in the sample.

The size S of each shock is set to 2 which allows to generate estimated models with the AUC statistic values around 85%, typically observed in credit risk models of homogeneous loan portfolios. The simulations are replicated 1000 times for each sample and main estimation outcomes are presented in Figures 1 and 2, and Tables 1 and 2. Different simulations contain different shock and default dates, and different values of explanatory variables. The two samples compared are: (1) the sample containing all observations, including overlapping observations, and (2) the sample containing only non-overlapping observations, i.e. every twelfth observation in each time series in the sample.

The results presented in Figure 1 and Table 1 suggest that reducing the sample increases uncertainty regarding the predictive power of a logistic regression. Surprisingly, the models based on annual data generate on average higher in-sample AUC statistics than the models based on the full sample data. The higher AUC may be observed due to a lower number of observations of the target variable equal 1 in the reduced sample.

The constant term is somewhat lower in the reduced-sample estimates in comparison to the full-sample estimates, while parameters of all three explanatory variables are higher in the reduced sample than in the full sample. In line with expectations, the reduced-sample estimates are less precise than the full-sample estimates of all parameters.

Figure 1: AUC statistics from models estimated using simulated data

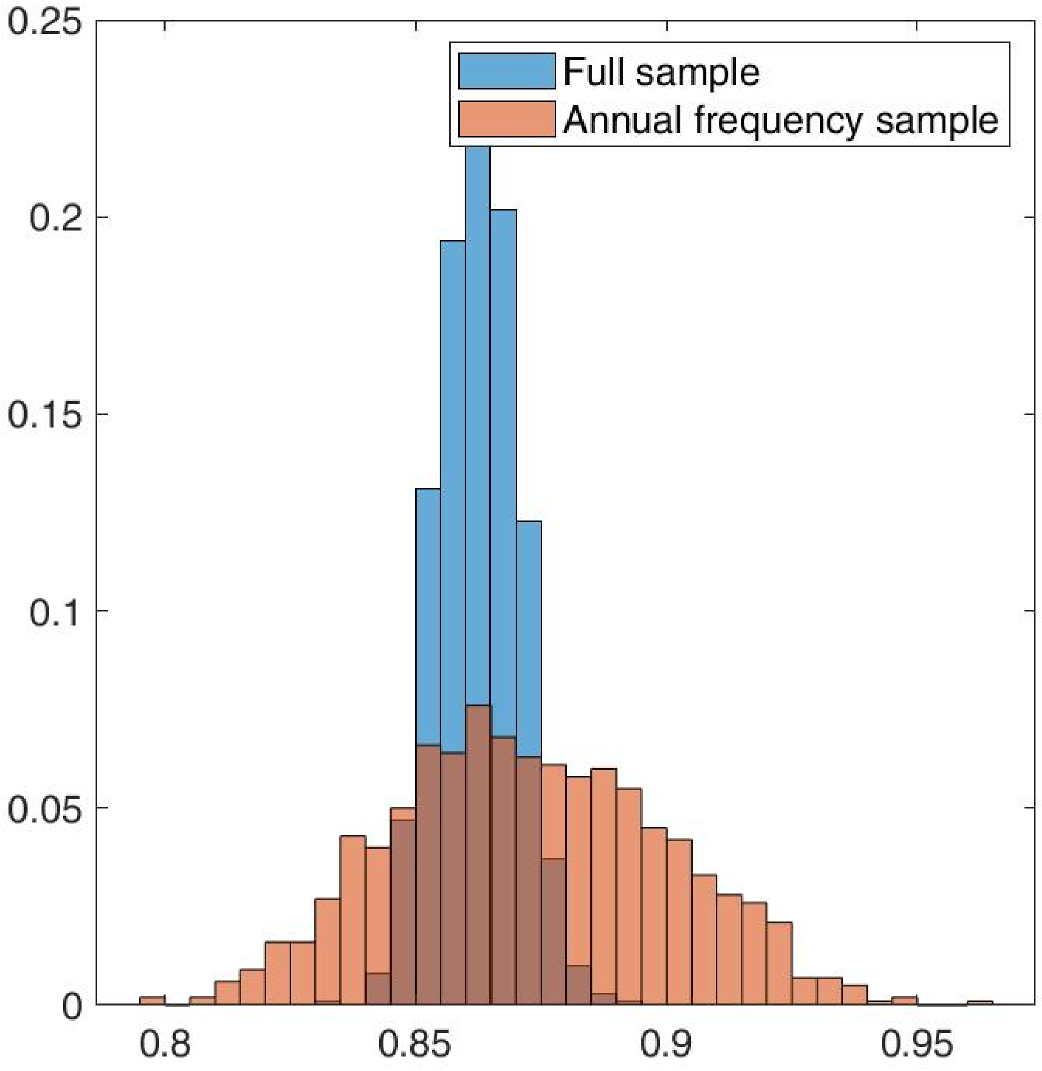
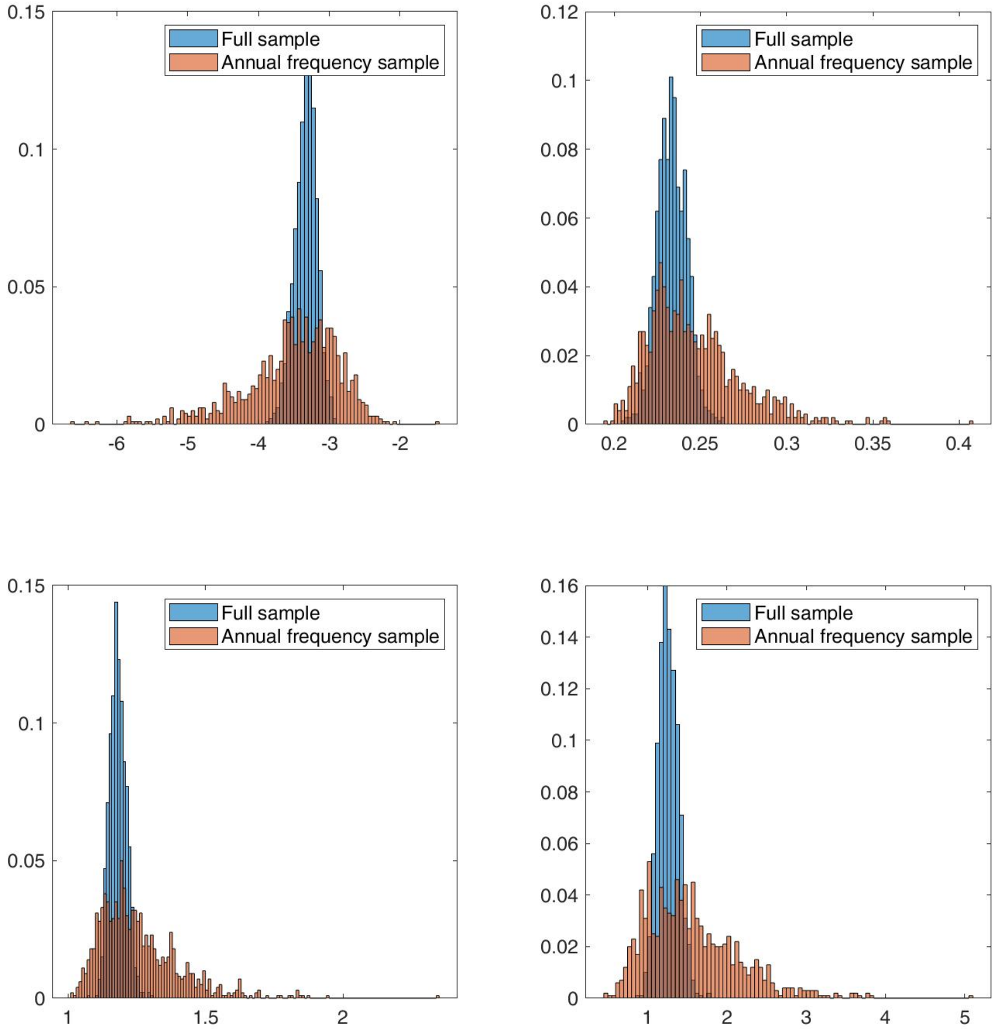


Table 1: AUC statistics from models with shocks of size 2

	Mean	Standard deviation
Full sample	0.862	0.008
Annual frequency	0.873	0.027

When the size S of a shock is reduced to 1.5, the AUC statistics become smaller in line with expectations, because the size of noise in each variable is larger relative to the size

Figure 2: Parameter estimates from models estimated using simulated data



Note: The estimated parameters are θ_0, θ_1, θ_2, and θ_3, respectively. Histograms are shown separately for each parameter.

of shocks. Nevertheless, the models estimated with annual data generate higher variance of predictive power than the models based on full sample estimates (Figure 3). Again, the constant term is slightly lower in the reduced-sample estimates in comparison to the full-sample estimates, while parameters of all three explanatory variables are higher in the reduced sample than in the full sample (Figure 4).

When the size of shocks is set to 3, the AUC statistics increase to values around 95% (Figure 5), but the main results tend to be robust, i.e. a slightly lower constant term and higher variable parameters in models estimated using reduced samples (Figure 6). The differences between parameter estimates are never statistically significant and may result

Table 2: Parameter estimates from models with shocks of size 2

	Constant term	Application variable	Behavioral variable	Macroeconomic variable
Full sample - mean	-3.330	0.233	1.183	1.261
Full sample - st.dev.	0.154	0.009	0.030	0.126
Annual data - mean	-3.520	0.246	1.257	1.558
Annual data - st.dev.	0.669	0.027	0.148	0.588
Difference - mean	0.189	-0.013	-0.074	-0.297
Difference - st.dev.	0.642	0.025	0.145	0.573

Note: The difference is computed between the full sample estimates and the annual data estimates. Mean and standard deviation statistics are based on 1000 simulation replications.

from increased uncertainty of the reduced-sample estimates.

In addition, the samples where the average monthly default rate is reduced to 0.2% or increased to implausible 5% are investigated. Changing the frequency of shocks and defaults does not affect the general conclusions, because the estimates in both overlapping and non-overlapping samples have similar values, but the latter estimates are more volatile than the former.

Moreover, halving the length of time series observations from 120 to 60 does not produce different conclusions either (Figures 7 and 8). We have expected that reducing the time series dimension in relation to the cross-sectional dimension may have some effect on the results, but again the conclusions remain unchanged.

Finally, we have introduced strong serial correlation to explanatory variables by incorporating the autoregressive process of order one with the parameter equal 0.7 to each time series of the respective behavioral and macroeconomic variable. Figures 9 and 10 confirm earlier results with one additional note that correlated observations of explanatory variables reduce the gain of using overlapping observations, namely the volatility of estimates is less reduced in these simulations.

5 Conclusions

The difference between parameter estimates using all overlapping observations in a sample and only non-overlapping observations in a reduced sample is not statistically significant, but the variance of parameter estimates is reduced when overlapping observations are used. This reduction of volatility is less visible when time-series observations of explanatory variables are strongly correlated.

Although the differences in mean parameter estimates are not statistically significant, some patterns are clearly observable. The constant term tends to be lower and parameters of explanatory variables tend to be higher in the reduced-sample estimates (with non-overlapping observations) than in the full-sample estimates (with overlapping observations). Further research may investigate the sources of these differences and, even more interestingly, the potential estimation methods that make use of information contained in correlated observations in credit risk models.

Acknowledgements

This research has been financed by the National Centre of Science (grant no. OPUS 2020/37/B/HS4/01893). We would like to thank Paweł Kopciuszewski for valuable comments on the early draft of this paper.

References

Albert, A. and Anderson, J. A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 71(1):1–10.

Boudoukh, J., Israel, R., and Richardson, M. P. (2020). Biases in Long-Horizon Predictive Regressions. Working Paper 27410, National Bureau of Economic Research, Inc.

Boylan, J. E. and Babai, M. Z. (2016). On the Performance of Overlapping and Non-overlapping Temporal Demand Aggregation Approaches. *International Journal of Production Economics*, 181:136–144.

Britten-Jones, M., Neuberger, A., and Nolte, I. (2011). Improved Inference in Regression with Overlapping Observations. *Journal of Business Finance & Accounting*, 38(5-6):657–683.

Chinn, M. D. (2006). The (Partial) Rehabilitation of Interest Rate Parity in the Floating Rate Era: Longer Horizons, Alternative Expectations, and Emerging Markets. *Journal of International Money and Finance*, 25(1):7–21.

Cipollini, A. and Fiordelisi, F. (2012). Economic Value, Competition and Financial Distress in the European Banking System. *Journal of Banking & Finance*, 36(11):3101–3109.

Darvas, Z. (2008). Estimation Bias and Inference in Overlapping Autoregressions: Implications for the Target-Zone Literature. *Oxford Bulletin of Economics and Statistics*, 70(1):1–22.

European Banking Authority (2017). Guidelines on PD Estimation, LGD Estimation and the Treatment of Defaulted Exposures.

Frankland, R., Smith, A. D., Sharpe, J., Bhatia, R., Jarvis, S., Jakhria, P., and Mehta, G. (2019). Calibration of VaR Models with Overlapping Data. *British Actuarial Journal*, 24.

Hansen, L. P. and Hodrick, R. J. (1980). Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis. *The Journal of Political Economy*, 88(5):829–853.

Kiesel, R., Perraudin, W., and Taylor, A. (2001). The Structure of Credit Risk: Spread Volatility and Ratings Transitions. Working Paper 131, Bank of England.

Lioui, A. and Poncet, P. (2019). Long Horizon Predictability: An Asset Allocation Perspective. *European Journal of Operational Research*, 278(3):961–975.

Mamingi, N. (2017). Beauty and Ugliness of Aggregation Over Time: A Survey. *Review of Economics*, 68(3):205–227.

Richardson, M. and Smith, T. (1991). Tests of Financial Models in the Presence of Overlapping Observations. *The Review of Financial Studies*, 4(2):227–254.

Rossana, R. and Seater, J. (1995). Temporal Aggregation and Economic Time Series. *Journal of Business & Economic Statistics*, 13(4):441–51.

Xu, K.-L. (2021). On the Serial Correlation in Multi-Horizon Predictive Quantile Regression. *Economics Letters*, 200(C).

Zwetsloot, I. M. and Woodall, W. H. (2021). A Review of Some Sampling and Aggregation Strategies for Basic Statistical Process Monitoring. *Journal of Quality Technology*, 53(1):1–16.

Figure 3: AUC statistics from models with shocks of size 1.5

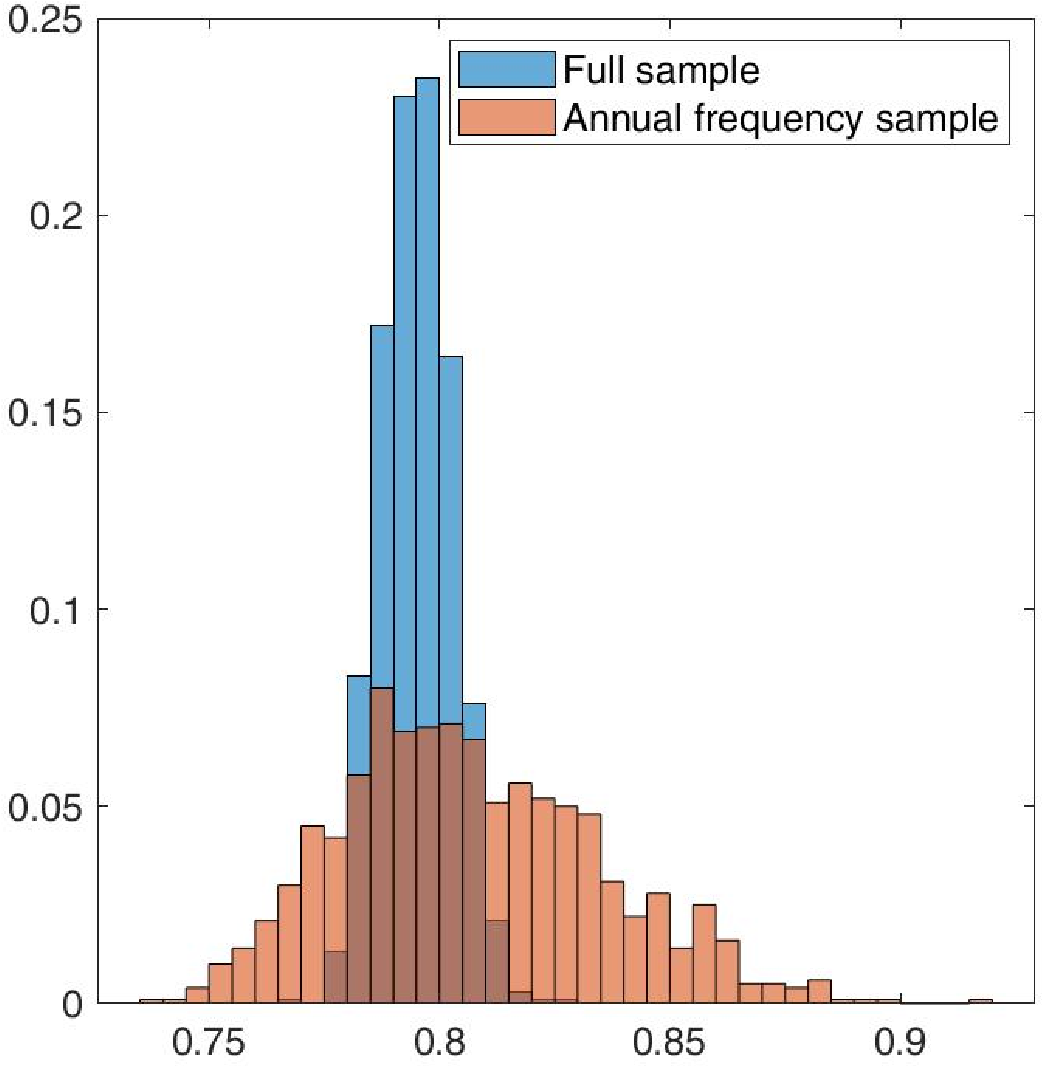
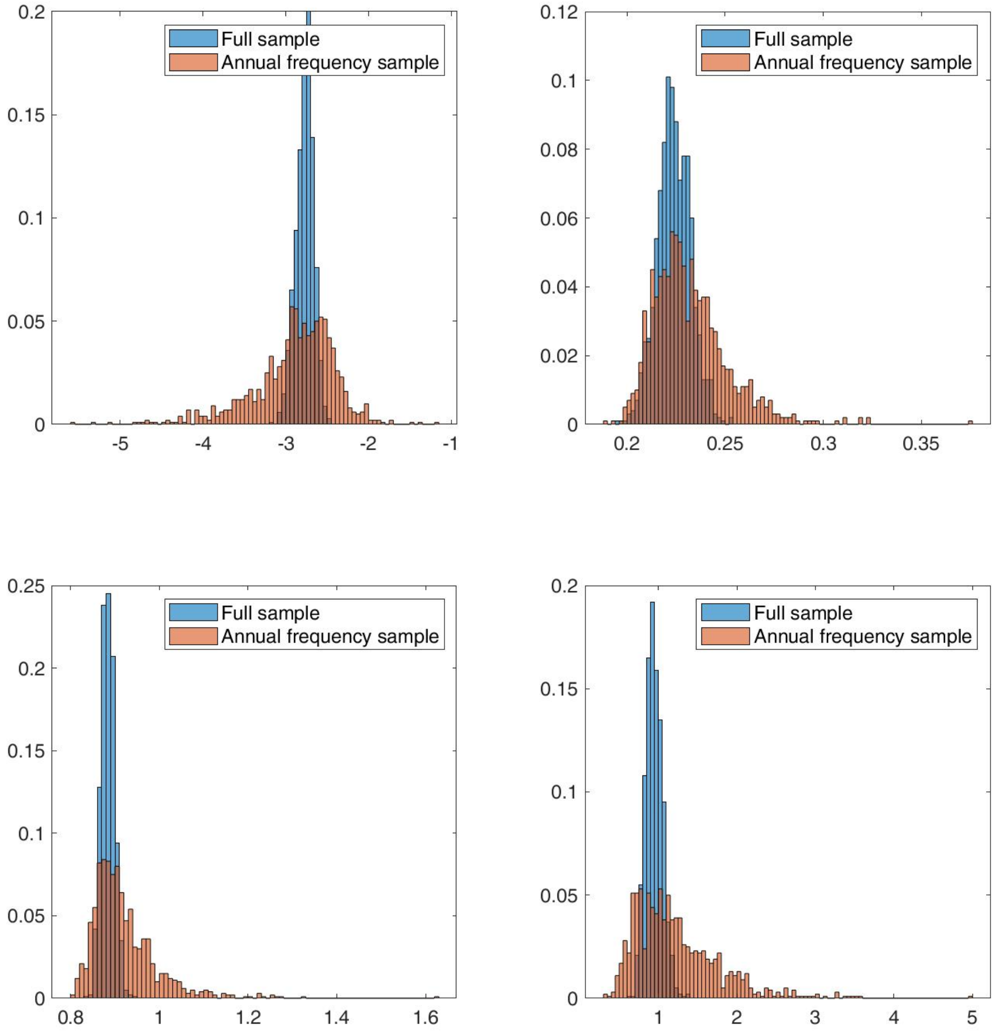


Figure 4: Parameter estimates from models with shocks of size 1.5



Note: The estimated parameters are θ_0, θ_1, θ_2, and θ_3, respectively. Histograms are shown separately for each parameter.

Figure 5: AUC statistics from models with shocks of size 3

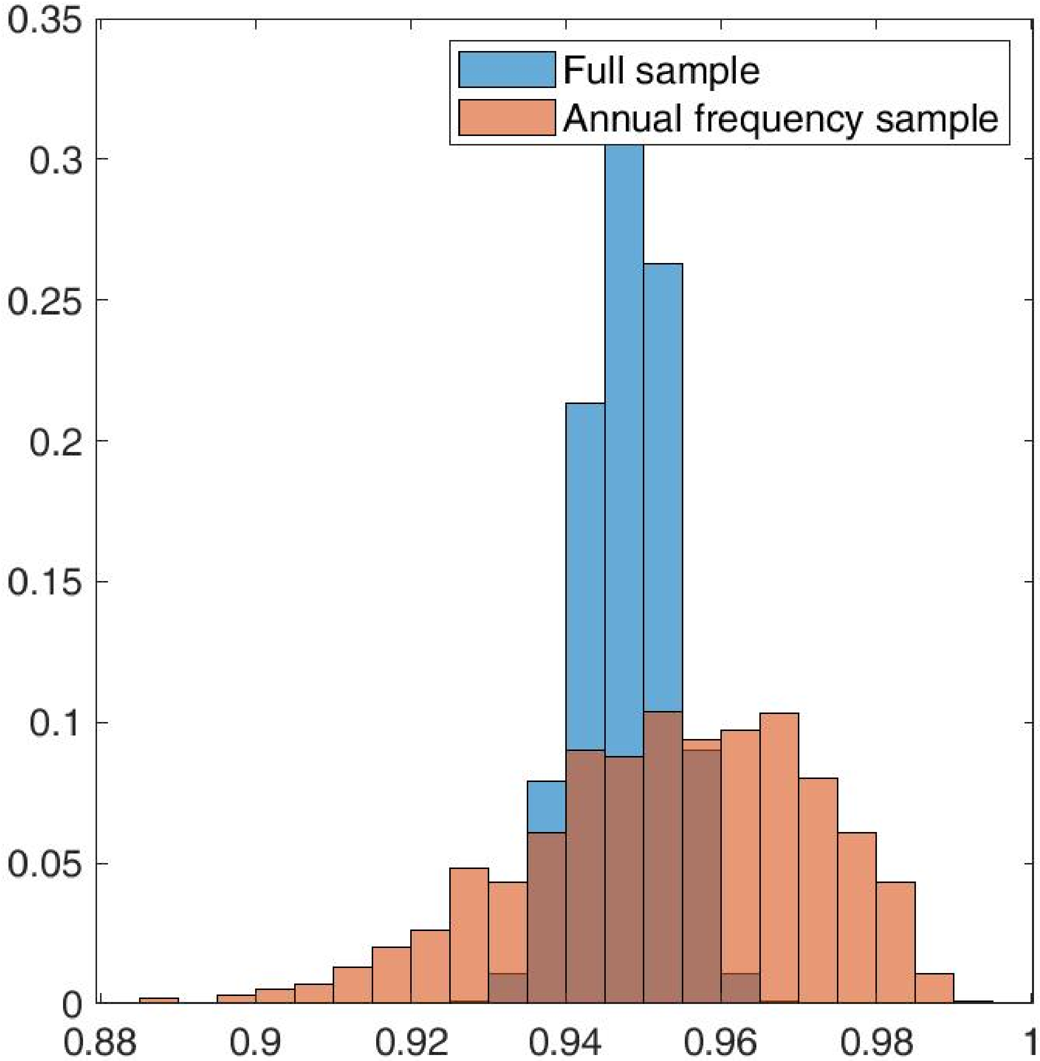
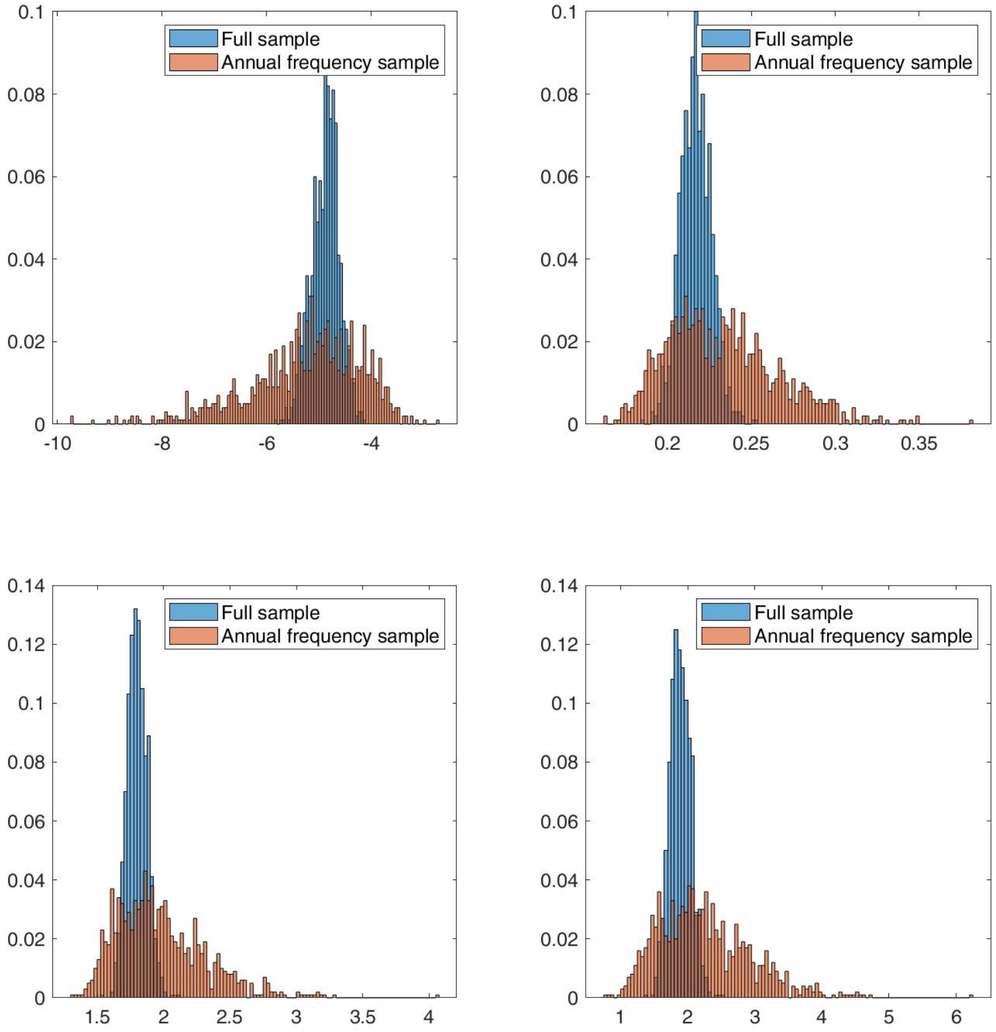


Figure 6: Parameter estimates from models with shocks of size 3



Note: The estimated parameters are θ_0, θ_1, θ_2, and θ_3, respectively. Histograms are shown separately for each parameter.

Figure 7: AUC statistics from models with the reduced time-series dimension

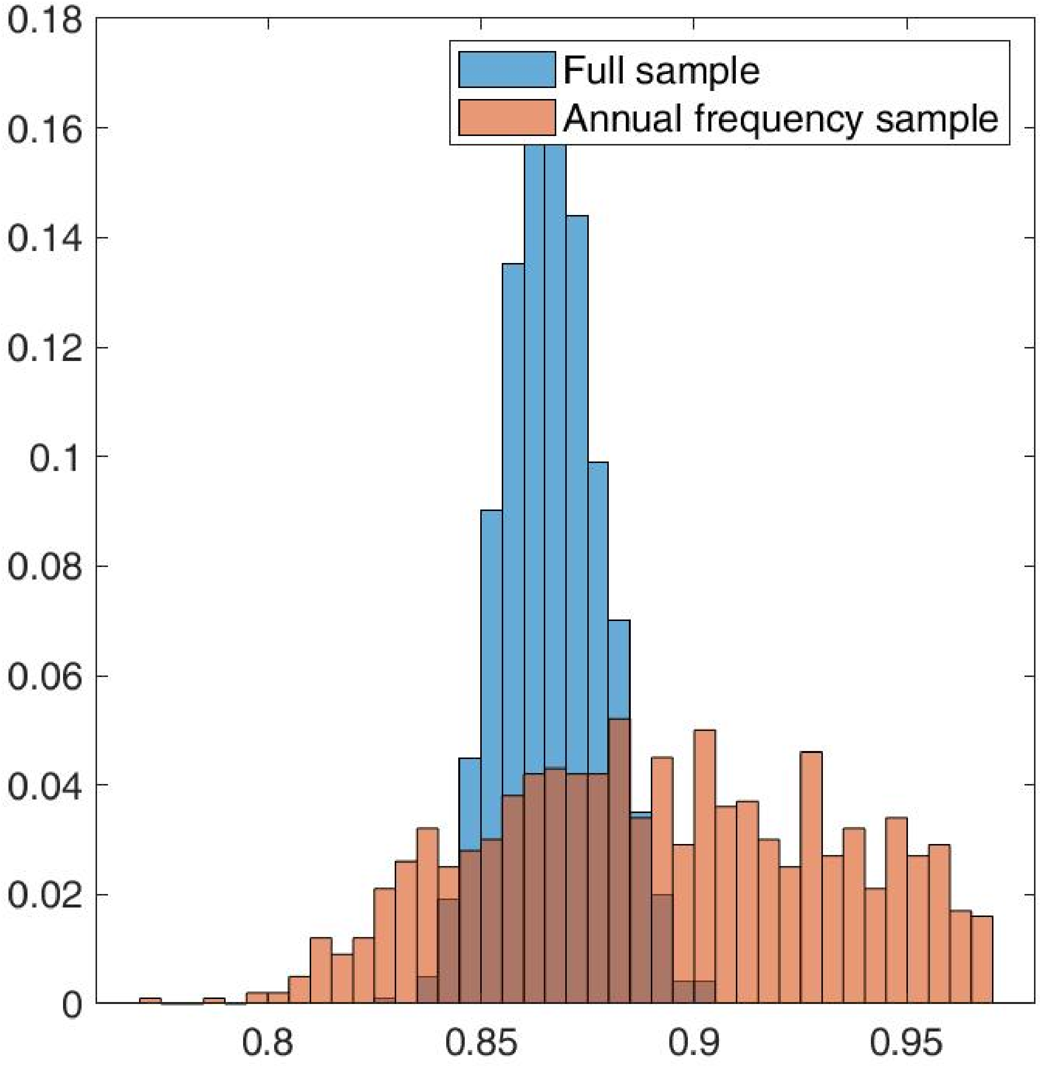
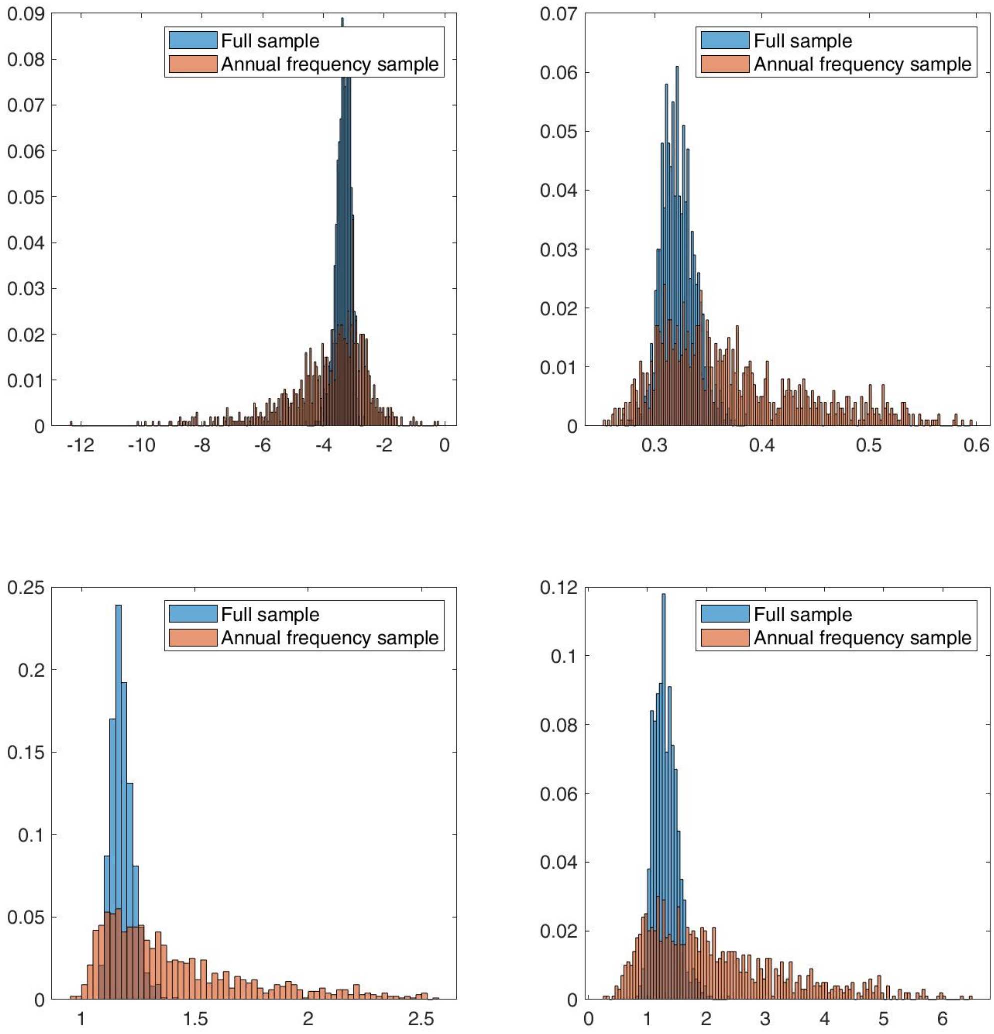


Figure 8: Parameter estimates from models with the reduced time-series dimension



Note: The estimated parameters are θ_0, θ_1, θ_2, and θ_3, respectively. Histograms are shown separately for each parameter.

Figure 9: AUC statistics from models with serial correlation of explanatory variables

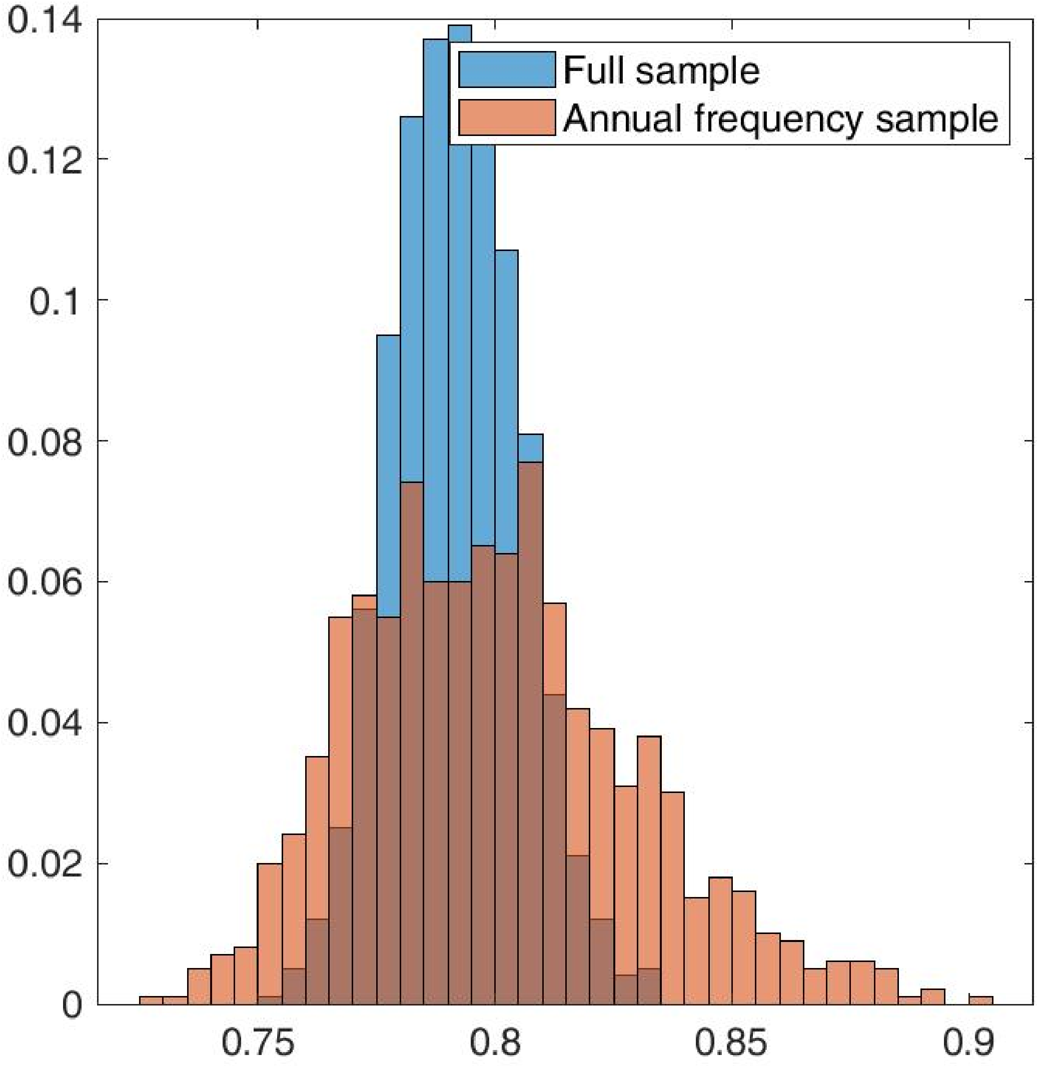
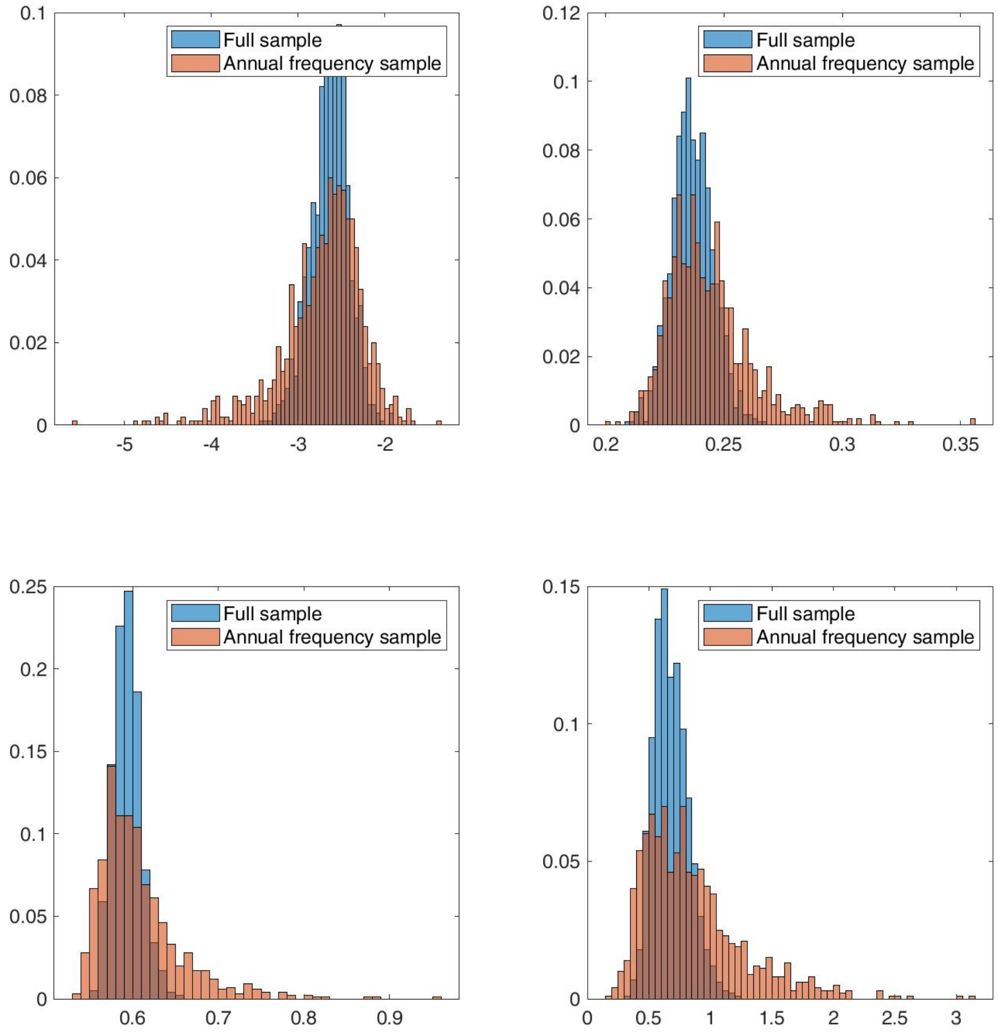


Figure 10: Parameter estimates from models with serial correlation of explanatory variables



Note: The estimated parameters are θ_0, θ_1, θ_2, and θ_3, respectively. Histograms are shown separately for each parameter.