Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Econometrics & Mangagement Science

Sample selection models in the presence of outliers - A simulation study of forecast performance.

Alexandru-Sebastian Fursenco (572884)

Supervisor:	Mikhail Zhelonkin
Second assessor:	Eoghan O'Neill
Date final version:	29th July 2024

Abstract

Sample selection bias often arises in practice, as the data available to the researcher is observed based on some selection mechanism. Estimators have been developed in practice to deal with such biases, most building on the works of Heckman (1974) and Heckman (1979). Despite this, literature rarely focuses on out-of-sample performance of sample selection models (SSMs). We treat the issue of forecasting in sample selection models, and investigate how contamination by outliers, which often deteriorates the performance of SSMs, influences the forecasting. Multiple estimators are compared among themselves, including a few density forecasts. We conclude that a few key considerations have to be taken into account when dealing with SSMs in a forecasting setting. Whether mean squared error (MSE) as a metric is preferable for evaluation depends heavily on these considerations. As an alternative, mean absolute error (MAE) can be used, which provides a handy way of directly comparing point forecasts with probabilistic forecasts assessed by the continuous ranked probability score (CRPS). The extensive simulation study indicates based on MSEs that in the absence of outliers, OLS is often preferable to approaches that incorporate sample selection, mainly due the risk of misclassification involved in forecasts created by SSMs. On the other hand, the classical two-step procedure and the robust Mallows-type M-estimator appear to be the best at handling outliers. OLS has an especially deteriorated performance when outliers are present in the forecast sample, albeit this hinges largely on the form of the outliers, and might not always be an issue. The Maximum Likelihood (ML) estimator appears to be less suitable for forecasting data with sample selection, having the worst performance across most scenarios. Finally, when it comes to density forecasts, these are inspired by the theoretical density of the outcome variable, as given by Marchenko and Genton (2012). Despite showing promising results in terms of CRPS, suggesting good sharpness, our attempts at guaranteeing calibration have not led to fruition, therefore such probabilistic forecasts need to be used with extreme caution. The empirical application on ambulatory expenses serves as further confirmation of these insights, and allows us to compare forecast results with earlier studies on in-sample performance.

Contents

Chapter 1

Introduction

Sample selection bias arises when inference or predictions are based on non-randomly selected samples, for example when we only observe the outcome based on some selection rule. This issue along with a possible solution has been first studied by Heckman (1974) and some of his subsequent works, who proposed a so called sample selection model (SSM) to remedy the bias. Heckman originally introduced two estimation methods for his SSM, a one step estimatior that makes use of the joint likelihood of the model, commonly referred to as the maximum likelihood (ML) estimatior, and a two-step estimator referred as such. Some author's have also outlined the link between sample selection models and Tobit models, with Amemiya (1984) classifying Heckman's original model as a Tobit-type 2 model. For a more in-depth, textbook treatment of SSMs one can refer to Cameron and Trivendi (2005).

Although widely used, many authors voiced concerns regarding SSMs in certain situations. For example, Puhani (2000) conducts a survey of model performance and finds that scenarios where selection and outcome errors are heavily correlated, which leads to high selection bias, are precisely the situations where the original sample selection model is the most inefficient. He also mentions an extensive debate about the small sample properties of such models, first sparked by Duan, Manning, Morris and Newhouse (1983). Others such as Marchenko and Genton (2012) or Ogundimu and Hutton (2015) find the performance of the original estimator proposed by Heckman to be lackluster when model errors have a non-normal distribution, which is often encountered in practice. These papers propose solutions to dealing with such errors, the former investigates a model that assumes t-distributed errors, while the later pertains to skew-normally distributed errors. Additionally, it has been established that outliers and other sorts of contamination can lead to problems in sample selection models. To address this robustness issue, alternative estimation procedures have been implemented, e.g. Zhelonkin (2013) and Zhelonkin, Genton and Ronchetti (2016).

Most of the literature is focused on in-sample inference, while the issue of forecasting or out-of-sample predictions is hardly ever treated. The need for elaborating a rigorous approach to forecasting sample selection is self-evident, given forecasting is the end goal of many empirical applications, coupled with the prevalence of sample selection bias in a considerable percentage of such datasets. The main problem of forecasting data sets suffering from sample selection bias lies within some of their features. Namely, the missing data points are all known in-sample, however, out-of-sample we do not know apriori which entries will be observable, therefore additional

problems need to be addressed. This, as well as many other problems plague the forecasts based on non-random sampling, which opens a perfect avenue for research, as in this thesis I plan to investigate the implications of outliers on forecasting performance in SSMs. The main goal of the paper is uncovering the forecasting capabilities of classical SSM estimators, and how these compare to the capabilities of the robust alternatives. By means of a simulation study, I assess the relative performances of the different estimators under various scenarios, including at different sample sizes, number of replications, percentage of outliers, gravity of each outlier. Secondly, forecasting implies a split of the sample into estimation and forecasting periods. In the context of outliers this poses the question of what are the implications of said outliers in either of the two periods. A final consideration to be made is how to deal with the selection rule, as for in-sample inference we have access to this binary variable, whereas forecasts imply the need of predicting it too.

The remainder of this paper is structured as follows: Section 2 discusses the models and estimators employed, along with the necessary theoretical concepts of robustness and forecasting. The set-up and results of the simulation study are given in Section 3, where the main results of the paper are discussed. Section 4 is devoted to an empirical application on ambulatory expenses. The data-set in question is the same as the one used in Zhelonkin et al. (2016), it allows us to compare and put in practice the insights of the simulation. Finally, Section 5 is comprised of concluding remarks, mostly focusing on the main challenges arising in dealing with forecasts of sample selection models.

Chapter 2

Methodology

2.1 Sample Selection Models

The model proposed by Heckman (1979) initially has the following form:

$$\begin{aligned} y_{1i}^* &= x_{1i}^T \beta_1 + e_{1i}, \\ y_{2i}^* &= x_{2i}^T \beta_2 + e_{2i}, \end{aligned} \tag{2.1}$$

Where y_{1i}^* and y_{2i}^* are latent and unobserved by the researcher, x_{1i} and x_{2i} are the regressors and the error terms e_{1i}, $e_{2i} \sim N(0, \Sigma)$ where the variance matrix Σ is of the form:

$$\Sigma = \begin{pmatrix} 1 & \rho\sigma_2 \\ \rho\sigma_2 & \sigma_2^2 \end{pmatrix}, \text{with correlation } \rho \text{ and } \sigma_2^2 \text{ the variance of } \ e_{2i}.$$

The first equation in (2.1) is called the selection equation, while the second one is the equation of interest or outcome equation. In practice one does not observe the true regressands y_{1i}^* and y_{2i}^*, instead we have the observable variables:

$$y_{1i} = \begin{cases} 1, & \text{if } y_{1i}^* > 0, \\ 0, & \text{if } y_{1i}^* \leq .0, \end{cases} \tag{2.2}$$

$$y_{2i} = \begin{cases} y_{2i}^*, & \text{if } y_{1i} = 1, \\ \text{NA}, & \text{if } y_{1i} = 0. \end{cases} \tag{2.3}$$

Note that the original paper switches the order of the two equations in (2.1), while later authors usually preserve the notation that is used in this paper.

In terms of estimation of this model there are a few possibilities available. Firstly, we can maximize the joint likelihood function of the two dependent variables, Heckman (1974). The

log-likelihood in this case can be expressed as:

$$l(\theta) = \sum_{i=1}^{N} y_{1i} * \ln\left(\Phi\left(\frac{x_{1i}^T\beta_1 + \frac{\rho}{\sigma_2}(y_{2i} - x_{2i}^T\beta_2)}{\sqrt{1-\rho^2}}\right)\right) + \sum_{i=1}^{N} y_{1i} * \ln\left(\phi\left(\frac{y_{2i} - x_{2i}^T\beta_2}{\sigma_2}\right)\right) + \sum_{i=1}^{N}(1 - y_{1i}) * \ln\left(\Phi\Big(-x_{1i}^T\beta_1\Big)\right), \quad (2.4)$$

where Φ and ϕ are the cdf and pdf of the normal distribution that is assumed under the classical model. For more detail see Heckman (1974).

The second approach that is widely used is a two-stage estimator. The method makes use of the form of the expectation of the outcome error conditional on the selection error. More specifically it can be proven that $E(e_{2i}|e_{1i} > -x_{1i}^T\beta_1) = \rho\sigma_2\frac{\phi(x_{1i}^T\beta_1)}{\Phi(x_{1i}^T\beta_1)}$. We can use this to rewrite the outcome equation as:

$$y_{2i} = x_{2i}^T\beta_2 + \lambda_i\beta_\lambda + \nu_i, \quad (2.5)$$

where ν is a zero mean error term, $\beta_\lambda = \rho\sigma_2$ and $\lambda_i = \frac{\phi(x_{1i}^T\beta_1)}{\Phi(x_{1i}^T\beta_1)}$ is the inverse Mill's Ratio. We can estimate the selection equation by Maximum likelihood probit, then use the output of this regression, namely the mean of the outcome error conditional on the selection error, as input for the outcome equation, which in turn is estimated by OLS.

The second method is generally more popular due to its simple implementation and interpretation. At the same time the use of the Mill's ratio has raised some concerns, the main one being the possibility of encountering the problem of multicolinearity if the parameters in the selection and outcome equation are very similar. The main problem of the maximum likelihood approach (ML) on the other hand is its sensitivity to the assumption of normality, it being even more unstable than the 2-step estimator. There are a couple other critiques of these estimation methods, generally overlapping among the two estimation routines. For a more comprehensive comparison one can refer to survey studies such as Puhani (2000).

In this paper, both methods are applied and compared in terms of forecasting performance, so that we can assess whether one can significantly outperform the other.

2.2 Robustness concerns

The influence function (IF) as first introduced by Hampel (1974) is a concept that relates to how a functional $T(F)$ changes as we introduce an infinitesimal change to its input distribution F. Formally:

$$IF(z; T, F) = \lim_{\epsilon \to 0}[T((1-\epsilon)F + \epsilon\Delta_z) - T(F)]/\epsilon,$$

where T is the functional/estimator, F the distribution underlying our model and Δ_z is the probability measure that puts mass 1 at point z. It can be shown that the bias introduced by contamination with point mass z, is proportional to the influence function. Therefore, an influence function that is unbounded leads to a non-robust estimator, in other words, the estimator will react arbitrarily to the introduction of an infinitesimal amount of contamination. Zhelonkin et al. (2016) prove that the influence function in the two-stage estimation of the SSM is unbounded, and therefore the estimator can be considered non-robust. This means that even

one outlier, if it is bad enough, can lead to arbitrary results for the two-step estimator. Furthermore, they show how the influence of outlier contamination is not limited to bias but also to the asymptotic variance of the estimator. When it comes to the maximum likelihood approach, it is also well documented that the performance of such estimators deteriorates greatly in situations where the distributional assumptions are not met, or in the presence of outliers. Overviews of two-step maximum likelihood based procedures are given by Murphy and Topel (1985) and Pagan (1986). In general this signifies that the original estimation procedures for sample selection models should be avoided in the presence of outliers.

To solve this issue, Zhelonkin et al. (2016) propose an alternative, robust estimation. They make use of a robust Mallows-type M-estimator for both the first and second stages. For a detailed explanation of the properties of this type of estimator see Cantoni and Ronchetti (2001). The main idea behind the Mallows-type estimators is that they place bounds on the influence of leverage points and vertical outliers separately. In the case of sample selection models, the estimator's associated score function in the first stage is:

$$\Psi_1^R(z_1; S(F)) = \nu(z_1; \mu)\omega_1(x_1)\mu' - \alpha(\beta_1). \tag{2.6}$$

The notation is consistent with Zhelonkin et al. (2016), where the score Ψ is a function of $z_1 = (x_1, y_1)$ and $S(F) = \beta_1$, and is equal to the product between two weight functions ν and ω, minus the α term which is added to ensure the estimator is Fischer Consistent, with $\alpha(\beta_1) = (1/n)\sum_{i=1}^n E(z_1, \mu_i)\omega(x_{1i})\mu_i'$, $\mu_i = \Phi(x_{1i}^T\beta_1)$ and $\mu_i' = \partial\mu_i/\partial\beta_1$. The weight functions are defined as:

$$\nu(z_{1i}; \mu_i) = \psi_{c_1}(r_i)\frac{1}{V^{1/2}(\mu_i)}, \tag{2.7}$$

where $r_i = (y_{1i} - \mu_i)/V^{1/2}(\mu_i)$ are Pearson residuals and ψ_{c_1} is the Huber function given by:

$$\psi_{c_1}(r) = \begin{cases} r, & \text{if } \mid r \mid \leq c_1, \\ c_1 sgn(r) & \text{if } \mid r \mid > c_1. \end{cases} \tag{2.8}$$

The parameter c_1 is a tuning constant that ensures a given level of asymptotic efficiency, typically chosen to equal 1.345, Cantoni and Ronchetti (2001). For the weight function ω, on the other hand, there are a few possible candidates, Zhelonkin et al. (2016) settle on $\omega_{1i} = \sqrt{1 - H_{ii}}$ where H_{ii} is the ith diagonal element of $H = X(X^TX)^{-1}X^T$. The resulting quasi-likelihood equations can be written as:

$$\sum_{i=1}^n \Big\{\psi_{c_1}(r_i)\omega_1(x_{1i}\frac{1}{[\Phi(x_{1i}^T\beta_1)\{1 - \Phi(x_{1i}^T\beta_1 - 1)\}]^{1/2}}\phi(x_{1i}^T\beta_1)x_{1i} - \alpha(\beta_1)\Big\} = 0,$$

and the $E(\psi_{c_1}(r_i))$ in $\alpha(\beta_1)$ is equal to:

$$E\Big[\psi_{c_1}\Big\{\frac{y_{1i} - \mu_i}{V^{1/2}(\mu_i)}\Big\}\Big] = \psi_{c_1}\Big\{\frac{-\mu_i}{V^{1/2}(\mu_i)}\Big\}(1 - \Phi(x_{1i}^T\beta_1)) + \psi_{c_1}\Big\{\frac{1 - \mu_i}{V^{1/2}(\mu_i)}\Big\}\Phi(x_{1i}^T\beta_1).$$

The influence function of the resulting estimator is bounded, which guarantees robustness of the estimator of the selection equation.

In regards to the second stage, the authors propose another Mallows-type M-estimator with

the following Ψ-function:

$$\Psi_2^R(z_2;\lambda,T) = \Psi_{c_2}(y_2 - x_2^T\beta_2 - \lambda\beta_2)\omega(x_2,\lambda)y_1, \tag{2.9}$$

with Ψ_{c_2} being a Huber function defined similarly to (2.8), with a different choice of c_2, and $\omega(x_2,\lambda)$ is a weight function:

$$\omega(x_2,\lambda) = \begin{cases} x_2, & \text{if } d(x_2,\lambda) < c_m, \\ \frac{(x_2 c_m)}{d(x_2,\lambda)}, & \text{if } d(x_2,\lambda) \geq c_m, \end{cases} \tag{2.10}$$

which is based on a robust measure of the Mahalanobis distance $d(x_2,\lambda)$. In our context Mahalanobis distances refer to the distance between a point, in this case x_{2i}, and the centre of the distribution of these points, in other words, points that are too far away from the main body of observations have a bounded influence on the score function. The theoretical measure itself has been introduced by Mahalanobis (1936), while robust estimation can usually be achieved via a Minimum Covariance determinant (MCD) estimator, introduced by Rousseeuw (1985) and later improved upon by Rousseeuw and van Driessen (1999) with the introduction of fast-algorithm for feasible approximation of the estimator. The choice behind the tuning parameters c_m in 2.10 and c_2 in Ψ_{c_2} is made based on the theory of robust linear regression, for more details see the book by Hampel, Ronchetti, Rousseeuw and Stahel (1986). Zhelonkin et al. (2016) stop at values 1.345 for c_2 and c_m is chosen corresponding to a 5% critical level. This proposed estimator is robust to the influence of outliers, however, the authors do point out its possible vulnerability to high degrees in correlation between the variables in the selection and outcome equations. This means that if the exclusion restriction is not met, for example if we use the same variable as input in the first and second stages, the estimator needs to be modified accordingly. Such possible changes are discussed in Zhelonkin et al. (2016), although we leave this scenario beyond the scope of our study.

2.3 Forecast Considerations

Throughout the paper we follow an 80/20 split between the training sample, where the models are estimated, and the testing sample, where we evaluate the performance of the estimators. Say n is the observation such that 80% of the observations within the range: $1, 2, ..., n$ and the remaining 20% are contained in $n+1, n+2, ..., N$. These are our training and testing sets respectively. We first obtain the estimated coefficients in the training set, after which these values are used for constructing predictions within the forecasting sample.

In total, we will review 7 main forecasting approaches, first one being based on simple OLS estimates. Given y_2 our variable of interest, and x_2 our regressor, we specify a relationship of the form:

$$y_{2i} = \beta_0 + \beta_1 * x_{2i} + \epsilon_i, \tag{2.11}$$

if we stack all available observations of y_2 in an n-dimensional column vector, and call this vector

Y_2, while all x_2 can be stacked in a matrix X_2 of the following form:

$$X_2 = \begin{pmatrix} 1 & x_{21} \\ 1 & x_{22} \\ \vdots & \vdots \\ 1 & x_{2n} \end{pmatrix},$$

then the closed-form solution for the minimization problem is $\hat{\beta} = (X^T X)^{-1}(X^T Y)$, where $\beta = (\beta_0 \quad \beta_1)$ is a 2 by 1 vector containing both the intercept and slope coefficients in (2.11). The forecast built on OLS is the simplest, as we merely estimate OLS in-sample, after which the coefficients are used to predict future values of y_2, constructed as follows:

$$\hat{y}_{2i} = \hat{\beta}_0 + \hat{\beta}_1 * x_{2i}, \text{ where } i \text{ takes values } n+1, n+2, ...N. \tag{2.12}$$

The other forecasting procedures are based on variations of sample selection models, and are slightly more involved, just as their associated in-sample estimators. For these estimators, we first predict the selection variable y_1, based on which the prediction for y_2 is constructed. If the first variable y_1 is predicted to be 0, the forecast indicates that the corresponding y_2 will be unobservable, or in other words not available. The forecast $\hat{y}_1$ is set to 0 if we predict a selection probability lower than 0.5, and is set to 1 if said selection probability is higher or equal to 0.5.

To evaluate the forecasting performance we first need to establish how to quantify deviations from reality. Specifically, given we predict y_{1i} to be 0, this leads to a not available (NA) entry for y_{2i}, there is however the risk that y_{2i} will actually be observed, in which case our forecast has wrongfully predicted both y_{1i} and y_{2i}. The opposite situation where the predicted y_2 is in fact unobservable is likewise problematic. There are a few approaches to this: One could remove the observations that have been misclassified, although this hardly seems appropriate, as wrong forecasts are not punished. Alternatively, one could find a substitute value for the missing observations, for example in many applications 0 could be a candidate to replace the NAs in y_2. Although this can be dangerous if the substitute is very far from the actual observations, if the substitute is a suitable approximation, then the biases induced should be relatively mild. Furthermore, according to our observations, the biases tend to preserve the relative standings of the estimators, which means that despite warped absolute values, meaningful conclusions about the comparative performance can still be made. Finally, more complex loss functions can be considered, loss functions that would penalize the forecast for both missclassifying an observation, or in other words the wrong prediction of y_1, and for predicting y_2 with an error. This last approach raises the question of what the appropriate penalty would actually look like, such that there is no significant bias towards either disregarding missclassifications, or giving them too much weight. All 3 methods have significant downsides, and the appropriate choice should be determined based on the application, and whether a suitable substitute value can be found.

In our case, we have decided to focus on the second approach, as an observation of 0 instead of NA is a fairly common occurrence in many applications, such as the original wage example of Heckman (1979) or the empirical application we later review. This method also provides a

natural way for penalizing wrong classifications into observed or unobserved. Therefore, when evaluating the forecast, we calculate the deviation from reality based solely on y_2. In practice this means that the forecast takes a form as shown in equations (2.13) and (2.14). A more detailed discussion of the different approaches and their implications will follow in section 3.

For y_1, we consider two alternatives, classical probit or a robust generalized estimator as described by Cantoni and Ronchetti (2001) with the same link function. We use observation $1, 2, ..., n$, or the testing sample to estimate coefficients of either probit or the robust glm, based on which out-of-sample observations $n+1, n+2, ..., N$ of y_1 are predicted.

Practically, probit specifies the probability of an observation y_{1i} being equal to 1, as the CDF of a standard normal distribution evaluated at $\beta_0 + \beta_1 * x_{1i}$. The probability can be calculated by using the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{P}(y_{1i} = 1|x_{1i}) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 * x_{1i}).$$

We can then estimate the parameters of this model via Maximum Likelihood estimation, by maximizing the joint maximum likelihood, which assuming independent and identically distributed observations, can be written as:

$$L = \Pi_{i=1}^{n} \left(\Phi(\beta_0 + \beta_1 * x_{1i})^{y_1} \left[1 - \Phi(\beta_0 + \beta_1 * x_{1i})\right]^{1-y_{1i}} \right),$$

while the alternative robust approach follows the procedure of the first stage for the robust sample selection model based on the Mallows-type M-estimator, as described in subsection (2.2).

For the outcome variable itself, we make use of the coefficients estimated via three alternative approaches. These are the two-stage estimator and the Maximum Likelihood estimator, as well as a robust alternative. The first two are estimated according to subsection 2.1, and represent the two classical approaches to estimating sample selection data, while the latter follows subsection 2.2, and is based on Mallows-type M-estimators. After obtaining the coefficients, given our previous consideration regarding the substitution of unobserved values by 0, the ML forecast can be constructed as follows:

$$\begin{cases} \hat{y}_{2i} = \hat{\beta}_0 + \hat{\beta}_1 * x_{2i}, & \text{if } \hat{y}_{1i} = 1, \\ \hat{y}_{2i} = 0, & \text{if } \hat{y}_{1i} = 0. \end{cases} \tag{2.13}$$

where the $\hat{\beta}$ coefficients are now obtained via Maximum Likelihood instead. The two-step and robust estimators make use of the inverse Mill's ratio, which also needs to be accounted for in the forecast. Therefore, said forecast takes a slightly different form:

$$\begin{cases} \hat{y}_{2i} = \hat{\beta}_0 + \hat{\beta}_1 * x_{2i} + \hat{\beta}_\lambda * \hat{\lambda}_i, & \text{if } \hat{y}_{1i} = 1, \\ \hat{y}_{2i} = 0, & \text{if } \hat{y}_{1i} = 0. \end{cases} \tag{2.14}$$

where $\hat{\lambda}_i$ is the (estimated) inverse Mill's ratio, and the $\hat{\beta}$ coefficients along with λ are estimated either by the two-step procedure or Mallows-type M.

Given that we have two options for forecasting the selection variable, and three alternatives for the estimation of parameters which are used to compute the outcome variable, in total we have six possible combinations:

Table 2.1: Forecast Alternatives

Approach	two-step	Maximum Likelihood	Robust two-step
Probit	SSM	SSM-ML	SSMR
Robust GLM	R-SSM	R-SSM-ML	R-SSMR

This table indicates the naming system for the different estimation approaches used for later results in the paper. SSM is the most basic approach, which stands for the two-step coefficients with probit for selection. R- before the rest of the name indicates that the robust approach was used in predicting the selection variable instead of probit. -ML indicates that coefficients have been retrieved from a Maximum Likelihood, while SSMR refers to the robust approach obtained via a Mallows-type M-estimator.

When it comes to measures evaluating model performance out-of-sample, there are a few possible candidates. Mean squared error or MSE can be considered for its widespread use both academically but also by practitioners. This measure of forecast accuracy entails taking the square of all the individual errors committed by the forecast and then averaging them over. For the specific variable y_2 which we apply MSE to, it can be expressed as:

$$MSE_{y_2} = \frac{1}{N-n}\Sigma_{i=n+1}^{N}\left(y_{2i} - \hat{y}_{2i}\right)^2,$$

where n is the observation splitting the data into training and testing samples and N is the last observation, while $\hat{y}_{2i}$ is the predicted value of the true observation y_{2i}. Considering the MSE is calculated in the forecasting period, perhaps it would be more precise to call it mean forecast squared error (MSFE), but for simplicity of notation we omit this. The main drawback in our context of MSE is its disproportionately high penalization of bad forecasts. Normally this can be a positive, as we wish to ensure against predictions that are quite far off. In the case of outlier contamination, however, this can lead to our assessment favoring models that are relatively more influenced by the outliers. This is especially undesirable when the outliers are not part of the true DGP, and thus do not present an interest to the researcher.

A measure that might be slightly better suited to deal with such situations is the MAE, or mean absolute error, wherein the forecast error is the mean deviation from the true value in absolute terms, instead of the square in MSE:

$$MAE_{y_2} = \frac{1}{N-n}\Sigma_{i=n+1}^{N} \mid y_{2i} - \hat{y}_{2i} \mid .$$

This measure penalizes big errors in a relatively less strict manner, however, it can still be impacted largely by outliers, although to a lesser extent. At the same time, MAE may penalize more severely errors that are between 0 and 1 in absolute terms, at least compared to MSE, where these errors are squared, thus closer to 0.

In forecasting evaluation literature, there is also the distinct line of thought that the model building stage and that of the forecast assessment need to be consistent with each other. Authors such as Granger and Pesaran (2000) argue for a "closer link between the decision and the forecast evaluation problems". That is, the choice of the forecast evaluation measure needs to be motivated by the decision rule that was used in creating the model in the first place.

This means that MAE, ideally, is applied to a model that is estimated consistently with this procedure, which would lead to a quantile regression. This is motivated by the fact that quantile regression models estimate the conditional median, rather than the conditional mean of the linear regression. This means that if we were to compare a forecast based on a quantile regression and a linear regression, using MSE would disproportionately favor the regression that estimates the mean, while the opposite would happen with MAE, we would prefer the median estimator. However, as we compare only models that estimate the conditional mean, no single model should have an unfair advantage.

Although there are a few options for quantile regressions, most of them are quite problematic. A couple of attempts have been made over the years, perhaps among the most notable being Buchinsky (2001), Huber and Melly (2015) and Arellano and Bonhomme (2017). Their estimation routines vary greatly, Buchinsky (2001) using a polynomial expansion, while Arellano and Bonhomme (2017) make use of a copula based approach. Unfortunately, these are quite sensitive to the assumptions, which if violated, can lead to lack of identification. In light of this, we decide to focus on traditional models instead, leaving the issue of forecasting with quantile regression for future research. Instead, we follow the line of thought of Granger and Pesaran (2000) and Kolassa (2020), which argue for the necessity of going beyond simple point forecasts, rather opting for more complex density forecasting.

2.4 Density Forecasting

As mentioned, many authors in the literature advocate for the use of probabilistic or density forecasting. This entails predicting the whole distribution of a random variable, rather than a point forecast. A review of such forecasts is provided by Gneiting and Katzfuss (2014), where they discuss some of the main theoretical considerations behind probabilistic forecasts, such as sharpness and calibration, and how these are affected by the scoring rule used to evaluate the forecast. Calibration, roughly, pertains to the ability of the forecast to correctly predict the data in expectation, while sharpness refers to the width of the prediction interval that was made. Most often a researcher chooses a desired level of calibration, and maximizes sharpness given the calibration. This translates to setting a target level of compatibility between the forecast and the data, then obtaining the most concentrated possible forecast given this constraint. Depending on the forecast in question, there are various ways of assessing calibration and sharpness. The probability integral transform (PIT) is a tool that can be used for any type of forecast. The PIT is defined as:

$$Z_F = F(Y-) + V(F(Y) - F(Y-)),$$

where V follows a standard uniform distribution, F is the probabilistic forecast, and Y is the observations. The PIT is essentially the value that the CDF takes at the observation, corrected for any points of discontinuity in F (see Gneiting and Katzfuss (2014) for more details). According to Gneiting and Katzfuss (2014), the collection of PITs of a calibrated forecast should follow a standard uniform distribution. In practice, we can plot the histogram of the PITs to examine whether or not these are uniform or not. Statistical tests are also available.

Gneiting and Katzfuss (2014) advocate for the employment of "proper" scoring rules, which roughly translates to using a scoring rule, for which the true distribution of the forecasted value is the best possible forecast. In other words, the expectation of the score is minimized when the forecast density is the same as the true distribution. If F and G are two probabilistic forecasts, and say that G is the true distribution of the forecasted data, then a proper scoring rule satisfies the following equation:

$$E_G[S(G,Y)] \leq E_G[S(F,Y)],$$

where $S(F,Y)$ is the numerical score attributed to forecast F given the data Y, and E_G denotes the expectation given that G is the true distribution. Proper scoring rules assess both sharpness and calibration simultaneously. The necessity of employing such scoring rules is outlined by a few authors, see for example Gneiting and Raftery (2007), who show what may transpire if improper scoring rules are used.

One scoring rule in particular they are in favor of is the Continuous Ranked Probability Score (CRPS). One of the key advantages of this rule, besides being proper, is that it can be used to directly compare point forecasts and probabilistic forecasts. The term has been introduced by Hersbach (2000), which itself was building on previous work of authors such as Matheson and Winkler (1976). It can be described mathematically as:

$$CRPS(F,y) = \int_{-\infty}^{\infty} (F(x) - \mathbb{1}(y \leq x))^2 dx = E_F(Y-y) - \frac{1}{2}E_F(Y-Y'), \tag{2.15}$$

where $F(x)$ is the predicted distribution, and $\mathbb{1}$ is an indicator function that shows whether the actual realization is smaller than the predicted value or not. In the second part of the equation, Y and Y' are independent random variables with cumulative density function (CDF) F and finite moments. The CRPS can be interpreted as the integral of the squared deviation of the forecast from the ideal forecast, which is a point mass prediction of the actual realization. If $F(x)$ itself is a point forecast, then the CRPS reduces to the mean absolute error, which as mentioned, allows us to directly compare the density forecast with point counterparts.

When it comes to a probabilistic forecast of a sample selection model, a few options are available. There are multiple authors implementing Bayesian techniques for estimating SSMs, or notably the semi-parametric approach by Chernozhukov, Fernandez-Val and Luo (2023). Such models often provide good flexibility, however, they can be quite complicated and oriented towards a specific application. In our case, a more parsimonious, yet flexible enough approach is inspired by Marchenko and Genton (2012), where they show that the outcome variable has a skew-normal distribution, if the error terms are jointly normally distributed. Furthermore, the same paper provides an extension to t-distributed errors. We may use said theoretical distribution as our predictive density. The skew-normal distribution of the outcome variable can be written as:

$$f_{ESN} = \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\frac{\Phi(\alpha\frac{y-\mu}{\sigma}+\tau)}{\Phi(\tau/\sqrt{1-\alpha^2})},\ y \in \mathbb{R}, \tag{2.16}$$

here $\mu = x^T\beta$, $\alpha = \rho/\sqrt{1-\rho^2}$ and $\tau = w^T\gamma/\sqrt{1-\rho^2}$, where β is the coefficient of the regressors

in the outcome equation, and γ is the coefficient of the regressors in the selection equation. In other words, $x^T\beta$ would be $x_2^T\beta_2$ and $w^T\gamma$ is $x_1^T\beta_1$ for the notation prior. The variable ρ is the correlation between the errors of the selection and outcome variables and σ is the variance of the outcome variable, while ϕ and Φ stand for the PDF and CDF of the standard normal respectively. For more information about the skew-elliptical distributions, including skew-normal and skew-t, consult Arellano-Valle and Genton (2010).

For each set of regressors in the test sample we compute the predictive density based on (2.16), with parameters such as β, τ or γ being calculated within the estimation sample. The correlation ρ and variance σ. We compute the variance using a standard estimator, while ρ is often found problematic in the context of sample selection, we therefore opt for the maximum likelihood estimator of this quantity which should be consistent, although we have also used a Kendall estimator for contrast. Marchenko and Genton (2012) find the assumption of t-distributed errors to be advantageous and more robust, compared to the classical assumption of normality. Although we focus on the skew-normal, the extension to t-distributed errors can be made relatively easily, therefore allowing for greater flexibility in applications where it is required.

Chapter 3

The Simulation Study

3.1 Set-up

3.1.1 General Considerations

To better understand the forecasting properties of sample selection models, a simulation study is conducted. There are multiple scenarios that we will discuss, namely different types and levels of contamination, different outlier forms, various sample sizes and numbers of replications. There are many estimators that are considered, our forecasting benchmark is a naive OLS estimation that does not account for the selection process. The rest of the estimators are various forms of sample selection models (SSMs) as described in section 2.3.

We split the N data points into a training set, with observations $1, 2..., n$, such that 80% of the observations are contained within the training set, while the rest, specifically $n+1, n+2, ...N$ form the testing set corresponding to 20% of the remaining observations. We are interested in contamination present in various parts of our sample, leading to four possibilities for the estimation. We look at outliers present strictly in the contamination sample, or the first 80%, next, outliers in the forecasting sample, or the last 20% meant for testing, and outliers present in both parts of the dataset. These results are compared to the case where there is no contamination.

Additionally, there are a few problems with forecasting sample selection models. Firstly, we do not have access to the dependent variable in the selection equation, which leads to us having to predict not only the outcome, but also the selection regressand y_1, as we do not know apriori whether or not an observation will be available. In accordance to 2.3, we consider two possibilities for the selection equation, namely a probit regression as well as a robust generalized estimator as described by Cantoni and Ronchetti (2001). Depending on the combination of approaches to predicting y_1 and y_2, we have managed to define a total of 6 methods that we build the forecasts on, in addition to the OLS benchmark. This leads to 7 methods in total, which are described in more detail in subsection 2.3. Furthermore, we start with point forecasts, mostly focusing on MSEs and then discussing other loss functions, before we switch to probabilistic forecasts.

Secondly, missing observations need to be accounted for in some way. This issue arises either when we predict a value to be non-observable when in fact it will be observable, or when we predict a non-observable value as being observed. In practice this means that we either get no forecast for certain observations, or we get a forecast which we cannot compare to the

actual observation as it is not available. The approach chosen to solve this problem, was to set the unobservables to a certain value. A natural candidate seems to be 0, although different applications may require other candidates, after all this value is arbitrary, and entirely at the discretion of the researcher. We demonstrate the implications of alternative approaches later on.

This decision has ramifications not only for the forecasting sample, but also for the estimation, particularly of OLS. As mentioned in section 2, Heckman (1979) makes a point of how sample selection bias can be handled. Specifically, OLS can be estimated using only the subset which is observed, or alternatively the whole sample. The first approach leads to the classical definition of sample selection bias, while the second induces bias because of the large point mass at the value at which the unobserved realizations are set. In other words, we may again use the unobserved values by setting them to 0, or we can decide to exclude the missing data from our estimation completely. Both options have been tested, and it appears, at least in our case, that OLS performs much better when the whole data set is used. This is not to say that this result is universal, far from it, in reality which approach is more appropriate once again depends on the distribution of the data, and the particular interests within the application. For our DGP, setting the missing values to 0 leads to the estimates being pulled towards the origin, although the solution is still pointing in the correct general direction. The other approach leads to coefficients which are vastly more erroneous, and therefore forecasts that are very far from the truth.

3.1.2 Data Generating Process

The data in the simulation study is obtained according to the standard selection model described in equations 2.1:

$$y_{1i}^* = x_{1i}^T \beta_1 + e_{1i},$$
$$y_{2i}^* = x_{2i}^T \beta_2 + e_{2i},$$

with $\beta_1 = \beta_2 = 1$. That is, the true selection variable y_1^* and the true outcome y_2^* are a sum of their respective exogenous regressors x_1 and x_2 and some errors e_1 and e_2. In our particular case, the exogenous regressors follow a standard normal distribution and are independent of each other, while $e_1, e_2 \sim N(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \text{ with correlation } \rho \text{ of the error terms.}$$

For most of the paper the correlation ρ is set to 0.7 to ensure a relatively strong selection mechanism that is nonetheless not extreme. In the parts where this changes we notify the reader.

As stipulated by equations (2.2) and (2.3), the models that we use only have access to the values of y_2 if the corresponding value of y_1 is non-zero.

3.2 Results - Point Forecasts

3.2.1 The baseline

Before following with the results, let us remind the reader of the nomenclature established in subsection 2.3. There are 7 estimators, which are: OLS, two-step (SSM), Maximum Likelihood (SSM-ML), and Mallows-type (SSMR), and each one except OLS has two variations, the normal version with probit in the first forecasting step, and a robust version, marked by an $R-$ before the model name, as in R-SSM-ML, stands for robust sample selection model estimated by maximum likelihood.

There are a few quantities of interest in the simulation set-up, namely we are interested in the effect of the percentage of contamination ϵ, the severity of the introduced contamination, and the correlation among the error terms in the two variables of the SSM. We also test whether results change under different sample sizes and numbers of replications as an additional robustness check. Let us first establish our baseline case that we compare other scenarios to. We have stopped on $M = 1000$, $N = 5000$, $\epsilon = 0.01$ and outliers of the form: $(x_1, x_2, y_1, y_2) = (-3, -3, 1, 3)$. As mentioned earlier, the generated sample is split in 80% estimation and 20% forecasting. Results of this scenario are shown in table 3.1.

Table 3.1: Point forecast results for the Baseline denoted in MSEs

Contamination Type	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
No contamination	0.7111	0.7358	0.7358	0.7507	0.7507	0.735	0.735
Estimation sample	0.7346	0.7422	0.7384	0.7746	0.7748	0.7344	0.7315
Forecast sample	1.6011	1.1488	1.1488	1.163	1.163	1.148	1.1481
Estimation & Forecast	0.8721	0.8204	0.8178	0.8506	0.8507	0.8133	0.8116

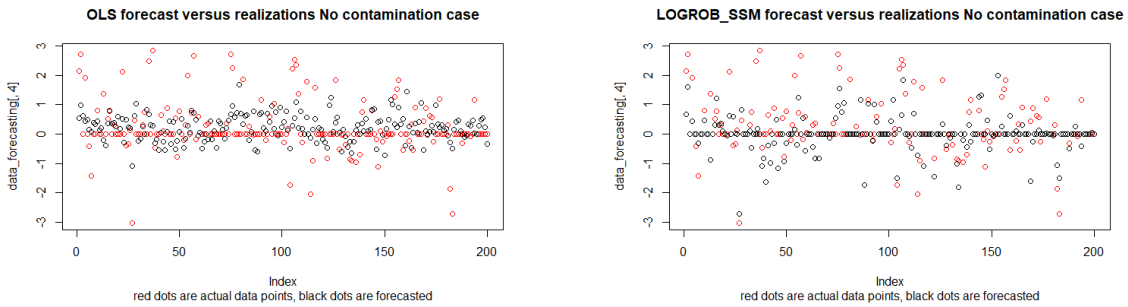
Note: *Baseline scenario M=1000, N = 5000,* $\epsilon = 0.01$ *and* $(x_1, x_2, y_1, y_2) = (-3, -3, 1, 3)$. *Row 1 is reserved for the name of the approach, 2-5 correspond to contamination in different parts of the sample (either estimation or testing period). Model names can be deciphered as follows: R - robust version of probit, SSM is the classical two step estimator while SSM-ML is the maximum likelihood version. SSMR is the robustified two-step procedure based on the Mallows-type M estimators.*

Firstly, row two displays the performance in the absence of contamination. As we can see, most models have similar performances, with the notable exception of the MSE of OLS being the smallest, however, not by a large margin. The OLS is unique in the way we compute the forecast for it, as it completely ignores the selection mechanism, while the others take a two phase approach, first predicting the selection variable, and only then building the outcome forecast based on the first one. Moving on, when having outliers in the estimation sample OLS still is best performing, although the gap between OLS and SSMs becomes smaller. It seems that the robust estimator becomes the best in the presence of estimation outliers. Secondly, robustifying the forecast of the selection variable comes at almost no cost in the absence of outliers, while allowing for a mild increase in performance in the presence of outliers. We will see later how this increase can be more drastic as the outliers get more severe. Finally, it seems OLS is the one forecast most influenced by the introduction of outliers in the testing sample, as shown in row four of the table.

For now let us focus on the first point, namely why OLS outperforms the SSMs in the absence

of outliers. It is useful to have a look at the plots of the actual realizations of the data, and the forecasts that each model makes. Left side of Figure 3.1 plots these values for OLS, while on the right we have them for one of the sample selection models. Note also that the plots for all SSMs are similar, so that what is discussed next pertains to all of them at the same time, and not just the one in the figure.

Figure 3.1: Forecast versus Realization (OLS and RP-SSM)



Note: Red - realizations; Black - Forecasted Values

The OLS estimates and the way the forecast is computed, makes it so that the model always predicts a value, which is rarely extremely accurate, however, it is mostly somewhere in the vicinity of the true realization. OLS makes plenty of small errors, but it does not run the risk of not predicting a value which is non-zero. The same cannot be said about SSMs, as they are designed to set some of the values to 0, they can sometimes missclasify an observation. The resulting forecast is much more accurate most of the times, as for most observations it correctly can predict them as zeroes. In spite of this, sometimes the sample selection model can make a mistake, which usually means quite a large error. As MSE is much harsher on a small number of large errors compared to large numbers of small errors, this forecast evaluation measure is very favorable to OLS in this setting. This is in contrast to what we see for example with the mean absolute deviation, which, as we discussed in the evaluation measures part of section 2, leads to much higher penalties for smaller errors, and milder penalties for larger errors. To test this further, we look into an alternative method of forecasting. Namely, instead of substituting the unobserved y_{2i} values by 0, we simply remove the observations, as they are unobservable and cannot be evaluated otherwise. We display these results in blocks 2 and 3 of Table 3.2. In block 2, we exclude all unobservable y_{2i} from the forecasting sample, while block 3 takes it one step further, and removes both unobserved y_{2i}, but also any observation that our models predict as unobserved (even if in fact it might be observable).

Results OLS with 2 alternative approaches (No observations omitted from forecasting sample):

Contamination Type	$OLS_{alt.}$	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
No contamination	1.539	0.7111	0.7358	0.7358	0.7507	0.7507	0.735	0.735
Estimation sample	1.848	0.7346	0.7422	0.7384	0.7746	0.7748	0.7344	0.7315
Forecast sample	1.567	1.6011	1.1488	1.1488	1.163	1.163	1.148	1.1481
Estimation & Forecast	1.792	0.8721	0.8204	0.8178	0.8506	0.8507	0.8133	0.8116

Results with omission of unobserved y_{2i} in the forecasting set:

Contamination Type	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
No contamination	0.566	0.591	0.626	0.591	0.591	0.591	0.626
Estimation sample	0.641	0.592	0.684	0.587	0.605	0.611	0.694
Forecast sample	1.463	1.011	1.044	1.011	1.011	1.085	1.044
Estimation & Forecast	0.637	0.595	0.677	0.589	0.604	0.610	0.685

Results with omission of unobserved and predicted as unobserved from the forecasting sample:

Contamination Type	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
No contamination	0.404	0.301	0.336	0.301	0.302	0.302	0.336
Estimation sample	0.484	0.334	0.426	0.329	0.347	0.353	0.436
Forecast sample	0.384	0.286	0.319	0.286	0.287	0.310	0.319
Estimation & Forecast	0.463	0.324	0.405	0.317	0.332	0.339	0.414

Table 3.2: Table Exclusion of unobservables and OLS comparison

Note: Block 1 - Results for OLS with 2 alternative procedures. OLS displays the standard approach where we replace unobserved values of y_{2i} in the estimation sample by 0, while OLS_{alt} reflects the alternative routine where we exclude all the observations which are incomplete. In block 1, no observations are excluded in the forecasting set. Block 2 - Results where y_{2i} is excluded from the forecasting set if they are unobserved. Block 3 - Results where we exclude observations in the forecasting set if y_{2i} is not available or if we predict $\hat{y_{2i}}$ to be unobservable. We consider only the baseline case of $M = 1000$, $N = 5000$, $\epsilon = 0.01$ and standard outlier form $(x_1,x_2,y_1,y_2) = (-3,-3,1,3)$.

When it comes to the problem of misclassifying, it appears to be the driver of the disparity between OLS and SSM, although notably, it is a specific kind of misclassification that presents the main issue. It seems that removing realizations that are unobservable in the testing sample does not change the relative standings (block 2), instead, it is the values that SSMs predict as 0 but are actually non-zero that are the main driving force behind the higher MSEs of models other than OLS (block 3). Once again it is unclear how to approach this, evidently, not predicting an observed value is a mishap of SSMs and should be penalized in some way, however, the prediction itself is given in terms of a binary variable denoting whether or not we have access to the next value. Quantifying the effects of misclassification will inevitably depend on the specifics of the situation, and must be decided by the researcher. As postulated before, we proceeded with our standard approach of treating the prediction of unobserved as predicting a 0. Not considering these observations in Block 3 of Table 3.2 leads to us preferring SSMs over OLS even without contamination.

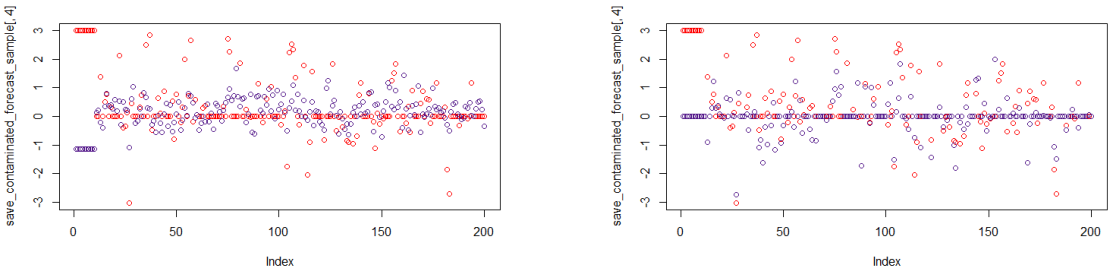
Additionally, let us compare the OLS procedure used so far, with another method, instead of keeping all the the values in the estimation set, and replacing values of y_{2i} where missing with 0,

one could remove these altogether, and estimate OLS only on the complete set. This results in a more classical sample selection bias arising, instead of bias introduced by the inclusion of a large number of 0 observations. Block 1 of Table 3.2 displays a comparison of the two. The alternative approach to OLS is clearly inferior, having a much worse performance across the board, except in the case of forecasting outliers, where its mostly unaffected due to the outlier form. This is holds true when most observations are centered around 0. While in other cases it might be problematic to find a good value to set the unobserved as, however, in our particular case it is not. We leave the decision about the choice between the two approaches to the practitioner, like many other decisions regarding forecasting sample selection models, this is very much dependent on the application, and a universal solution is very hard to imagine.

We make a final remark about the choice of the sample size and number of replications we have used this far. For the sake of completeness we provide Appendix A, which holds information about scenarios at different values of M and N. The general result of this appendix may be summarized as the fact that increasing number and dimension of samples leads to more accurate results, while confirming that the choice of these two values in the main text is appropriate, as other variations of the two lead to results which are very close.

The very lackluster performance of OLS when the testing sample is contaminated, can be explained easily if we once again look at plots of the forecast of OLS versus SSMs in the presence of contamination in the testing sample. The OLS forecast is always much more off given this particular type of contamination than those of the other models. It turns out, as we will see later in the subsection dedicated to outlier form, that OLS displays a similar behaviour under other types of outliers too, although how severe the deterioration depends on outlier form, that is, under certain circumstances, OLS can be competitive in contamination in the forecast sample. An illustration of the forecasted values and the true values is provided in Figure 3.2.

Figure 3.2: Forecast versus Realization - outliers in forecasting sample (OLS and R-SSM)



Note: Red - Realizations (contaminated); Blue - Forecasted Values

Additional informative boxplots illustrating the MSEs of different models across the samples are provided to the interested reader, Appendix B.

It is also of use to have a look at the forecast performance given the true realizations, by looking at counterfactuals. That is, evaluating the forecasts not only in comparison to the observed data, which is subject to the selection rule, but also assessing it given the true realizations, unaffected by said selection. This means that we leave the estimation procedure untouched, but in the forecasting sample, we are observing y_1^* and y_2^* in equation (2.1), rather

than y_1 and y_2 from (2.2) and (2.3) that we mostly work with. That is:

$$y_{2counter} = y_2 \text{ for } i \text{ in } 1, ...n \text{ and } y_{2counter} = y_2^* \text{ for } i \text{ in } n+1, n+2, ...N.$$

This means that we estimate the same coefficients, and leave the forecasting procedure the same, but now we can compare our forecasts with the real values, even if in practice they would be unobserved. This also means that we should forecast a value regardless of whether or not we believe that an observation will be available or not. Furthermore, given that we no longer only predict observables, we should drop the $\beta_\lambda * \lambda$ correction from the two-step and robust forecasts. Consequently, we have the following forecast for all 3 SSM procedures:

$$\hat{y}_{2i} = \beta_0 + \beta 1 * x_{2i}, \text{ where i takes values } n+1, n+2, ...N,$$

where the β values are estimates form either ML, two-step or Mallows-type M. Note also, that if we predict all values indiscriminately, there is no longer the need to distinguish between values that will be observed or not. One could attempt to do so, and derive conditional expectations given that we believe a value will be available or not, after which a different correction would be applied to observed and unobserved values. Still, our attempts to do so have resulted in worse results than indiscriminate forecasting. We report results in Table 3.3, only for the case of no contamination, as the other cases are in line with other Tables in the paper. Furthermore, since for the counterfactual case there is no distinction between unobserved and observed, the robust cases are dropped, as the forecasts are identical.

Table 3.3: Forecast comparison given access to counterfactual data

Estimator	OLS	SSM	SSM-ML	SSMR
MSEs	1.289	1.0024	1.0020	1.0026

Note: *The table shows the MSEs of the different models when we compare the forecasts not to the observed sample, but instead to the true sample, which is unaffected by selection mechanism.*

These results are very similar across the different SSMs, with OLS being significantly behind. Just like the case where the unobserved values are omitted, this points to the fact that SSM coefficients are generally more precise. Still, empirically there is no access to the counterfactual, therefore miss-classification poses a risk, and SSMs are not as clearly favored.

3.2.2 Impact of percentage of contamination

Let us go back to the main approach, where the NA values are substituted with 0s, and we perform a two stage forecast, as described in 2.3, so that we can delve deeper into the implications of outliers. Perhaps unsurprisingly, increasing the percentage of contamination leads to worse results, what is of interest, however, is the relative deterioration of the different models. When it comes to outliers in the estimation sample, the best result is displayed by the OLS, at least while outliers do not overwhelm the model fully. Under 1% of total contamination or less, OLS is still best performing. Higher percentages of outliers lead to a steeper worsening of OLS and ML, while the two-step and Mallows-type M are generally more robust. The robust version of the first step is preferred to the non-robust probit, as it does not lead to any significant losses

in efficiency in the absence of outliers, but makes a difference when outliers are present. Before discussing the likely cause of these results, an overview of said results is given by Table 3.4.

No contamination

M	N	ϵ	(x_1,x_2,y_1,y_2)	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
1000	5000	-	-	0.711	0.735	0.735	0.750	0.750	0.735	0.735

Contamination in the estimation sample

M	N	ϵ	(x_1,x_2,y_1,y_2)	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
1000	5000	0.004	(-3,-3,1,3)	0.715	0.731	0.730	0.756	0.756	0.729	0.729
1000	5000	0.01	(-3,-3,1,3)	0.734	0.742	0.738	0.774	0.774	0.734	0.731
1000	5000	0.02	(-3,-3,1,3)	0.788	0.772	0.755	0.810	0.809	0.757	0.742
1000	5000	0.05	(-3,-3,1,3)	1.007	0.949	0.840	3.029	2.712	0.938	0.831

Contamination in the forecast sample

M	N	ϵ	(x_1,x_2,y_1,y_2)	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
1000	5000	0.004	(-3,-3,1,3)	1.067	0.900	0.900	0.915	0.915	0.900	0.900
1000	5000	0.01	(-3,-3,1,3)	1.601	1.148	1.148	1.163	1.163	1.148	1.148
1000	5000	0.02	(-3,-3,1,3)	2.491	1.562	1.562	1.576	1.576	1.561	1.561
1000	5000	0.05	(-3,-3,1,3)	5.161	2.801	2.801	2.813	2.813	2.801	2.801

Contamination in the estimation and forecast samples

M	N	ϵ	(x_1,x_2,y_1,y_2)	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
1000	5000	0.004	(-3,-3,1,3)	0.779	0.763	0.763	0.787	0.787	0.763	0.763
1000	5000	0.01	(-3,-3,1,3)	0.872	0.820	0.817	0.850	0.850	0.813	0.811
1000	5000	0.02	(-3,-3,1,3)	1.004	0.923	0.913	0.960	0.959	0.913	0.904
1000	5000	0.05	(-3,-3,1,3)	1.281	1.275	1.206	2.165	2.010	1.257	1.190

Table 3.4: Point forecast results in MSEs (impact of percentage of contamination)
Note: *MSEs under different scenarios. The values are split into set-up parameters, followed by the MSE figures with either no contamination, or outliers in estimation/forecast/estimation & forecast. R− before the model name stands for the robust version of probit. SSM is the classical two step estimator while SSM-ML is the maximum likelihood version. SSMR is the robustified two-step procedure based on the Mallows-type M-estimators.*

Firstly, regarding the relative robustness of the two-step coefficients. It appears that the estimated intercept and slope both become downwards biased under the particular contamination of Table 3.4, on the other hand the inverse mill's ratio has an opposite upwards bias. The resulting estimates themselves can be quite biased, however, these two opposing forces cancel each other to a large extent when computing forecasts. Therefore, the two-step estimators do not suffer as much. Once again, the presence of outliers in the estimation sample leads to a preference to the robust probit. Finally, it appears that the Maximum Likelihood is affected by estimation sample outliers on par with OLS under most levels of contamination, while being most affected under 5% contamination. Outlier in both estimation and forecast samples lead to a mixture, an average between what is witnessed in when they are present either in estimation or forecasting separately.

One more thing to consider is the actual number of outliers, as the values in the table indicate a percentage of the total sample, which in practice can be very impactful. For example, given our

sample size of 5000 observations, 1% of outliers from the total sample leads to 50 outliers in total. Given that there are only 4000 observations for building the model, this leads to 1.25% outliers in the estimation sample. This might be a modest increase, however, we should also consider the fact that almost half of all observations are not used in the case of SSMs, as they only make use of non-zero, or observed realizations. In that context, 1% of the total sample contaminated can have a much higher influence than under normal circumstances. This can particularly affect results in the forecast sample, where a much higher percentage of outliers is introduced than the nominal level. We do note that this only changes the magnitude of the results, and not the general direction. To show the impact of this mechanism we also look into contamination where the nominal level is strictly followed in the respective sample. Furthermore, so far the outliers have been introduced indiscriminately. It is also of interest to see what happens if we have a situation where unobservable realizations only are replaced by observables, or vice versa. We illustrate this in Table 3.5 along with the indiscriminate case where the nominal level of contamination is respected.

Indiscriminate Contamination:							
Contamination Type	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
No contamination	0.711	0.735	0.735	0.750	0.750	0.735	0.735
Estimation sample	0.729	0.740	0.738	0.770	0.770	0.733	0.732
Forecast sample	0.891	0.821	0.821	0.835	0.835	0.820	0.820
Estimation & Forecast	0.874	0.823	0.821	0.852	0.852	0.816	0.814
Only contaminate 0s into 1s:							
Contamination Type	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
No contamination	0.711	0.735	0.735	0.750	0.750	0.735	0.735
Estimation sample	0.727	0.738	0.736	0.766	0.766	0.731	0.729
Forecast sample	0.894	0.823	0.823	0.839	0.839	0.822	0.823
Estimation & Forecast	0.894	0.823	0.823	0.839	0.839	0.822	0.823
Only contaminate 1s into 0s:							
Contamination Type	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
No contamination	0.711	0.735	0.735	0.750	0.750	0.735	0.735
Estimation sample	0.712	0.737	0.736	0.775	0.776	0.735	0.734
Forecast sample	0.705	0.762	0.763	0.776	0.776	0.761	0.761
Estimation & Forecast	0.705	0.761	0.760	0.798	0.798	0.762	0.761

Table 3.5: Results discriminating contamination and respected nominal levels

Note: The outlier types are the same for the first 2 scenarios (blocks 1 and 2), but differ in the last block, where from the normal outlier form of (-3,-3,1,3) we move to another form of contamination, namely (3,3,0,1). This change is motivated by the expected values of the predictors given that an values are observed or not. Respected nominal level refers to the fact strictly ϵ % of the respective sample will be contaminated. Other scenario parameters are exactly the same as the baseline, M= 1000, N=1000, ϵ = 1%

Respecting the nominal level of contamination is especially important when it comes to the forecast sample, as we have the least number of observations there, which means the same number of outliers in absolute terms can have a higher influence. Still, the patterns we have seen previously are preserved. Moving on to the other two cases, it appears that swapping unobservables for realized values is especially devastating, at least with the particular form that was tested. Interestingly, doing the opposite has a much milder effect, especially on OLS, which becomes relatively unaffected. This is likely due to the fact that the effective estimation

samples are already quite limited for SSMs. Targeting specifically those observations that would normally contain information means even less observations for SSMs. At the same time the relation between the predictor of the outcome x_2 and the outcome y_2 is somewhat in line with what the OLS predicts. Therefore, for OLS we can actually witness a slight decrease in MSE for the last case, where observed values are swapped for non-observed ones.

3.2.3 Impact of outlier form

We have already seen in the last example how outliers can have a significantly different impact on the different approaches. We will now investigate in more detail what happens under various forms of outliers. We are interested in the implications of changing the point mass contamination. Results are shown in Table 3.6.

No contamination (different scenarios/outliers forms lead to the same outcome)

M	N	ϵ	(x_1,x_2,y_1,y_2)	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
1000	5000	-	-	0.711	0.735	0.735	0.750	0.750	0.735	0.735

Contamination in the estimation sample

M	N	ϵ	(x_1,x_2,y_1,y_2)	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
1000	5000	0.01	(-3,-3,1,3)	0.734	0.742	0.738	0.774	0.774	0.734	0.731
1000	5000	0.01	(-3,-3,0,3)	0.734	0.735	0.735	0.750	0.750	0.735	0.735
1000	5000	0.01	(-2,-2,1,0)	0.711	0.735	0.735	0.763	0.762	0.733	0.733

Contamination in the forecast sample

M	N	ϵ	(x_1,x_2,y_1,y_2)	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
1000	5000	0.01	(-3,-3,1,3)	1.601	1.148	1.148	1.163	1.163	1.148	1.148
1000	5000	0.01	(-3,-3,0,3)	1.601	1.148	1.148	1.163	1.163	1.148	1.148
1000	5000	0.01	(-2,-2,1,0)	0.707	0.698	0.698	0.713	0.713	0.698	0.698

Contamination in the estimation and forecast samples

M	N	ϵ	(x_1,x_2,y_1,y_2)	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
1000	5000	0.01	(-3,-3,1,3)	0.872	0.820	0.817	0.850	0.850	0.813	0.811
1000	5000	0.01	(-3,-3,0,3)	0.872	0.818	0.818	0.833	0.833	0.817	0.817
1000	5000	0.01	(-2,-2,1,0)	0.710	0.727	0.727	0.753	0.752	0.726	0.726

Table 3.6: Point forecast results in MSEs (impact of outlier form)

Perhaps two things can be outlined from the last table. First that the value of the selection variable y_1 has an impact on whether the outlier will affect the SSMs when contamination is present in the estimation sample, but y_1 it is inconsequential to the performance of estimators when the testing sample is contaminated. Secondly, we can see how under a slightly modified outlier form, specifically row 3 of each block, the general patterns we have seen so far are repeated. This indicates that despite some variation in results depending on outlier form, we are likely to observe something very similar in Tables 3.4 and 3.6. That is OLS to be leading under no contamination, while being relatively more affected under outliers, especially the ones in the forecast samples. This seems to be overwhelmingly the case when comparing two-step procedures and OLS, while with ML exceptions can occur, specifically under extreme outliers. That is, occasionally the ML based estimates can become so erroneous under extreme contamination, that the forecasting performance falls below that of OLS.

3.2.4 Other loss functions

Until now we have discussed at length the dynamics of forecasting under the MSE loss function. In what follows, we evaluate the forecasts by means of mean absolute errors instead. Table 3.7 shows the MAEs of the models under a few interesting scenarios.

	M	N	ϵ	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR	CRPS
No	1000	5000	-	0.638	0.527	0.527	0.534	0.534	0.527	0.527	0.484
Est.	1000	5000	0.01	0.625	0.539	0.530	0.550	0.542	0.535	0.527	0.494
For.	1000	5000	0.01	0.822	0.651	0.651	0.657	0.657	0.650	0.650	0.618
Both	1000	5000	0.01	0.659	0.560	0.553	0.571	0.565	0.565	0.550	0.514
Est.	1000	5000	0.02	0.634	0.561	0.538	0.568	0.554	0.554	0.532	0.514
For.	1000	5000	0.02	1.005	0.775	0.775	0.780	0.780	0.774	0.774	0.752
Both	1000	5000	0.02	0.686	0.600	0.584	0.609	0.598	0.595	0.580	0.545
Est.	1000	5000	0.05	0.731	0.670	0.597	1.251	1.130	0.665	0.593	0.591
For.	1000	5000	0.05	1.554	1.145	1.145	1.150	1.150	1.145	1.145	1.153
Both	1000	5000	0.05	0.804	0.742	0.686	1.000	0.928	0.734	0.679	0.625

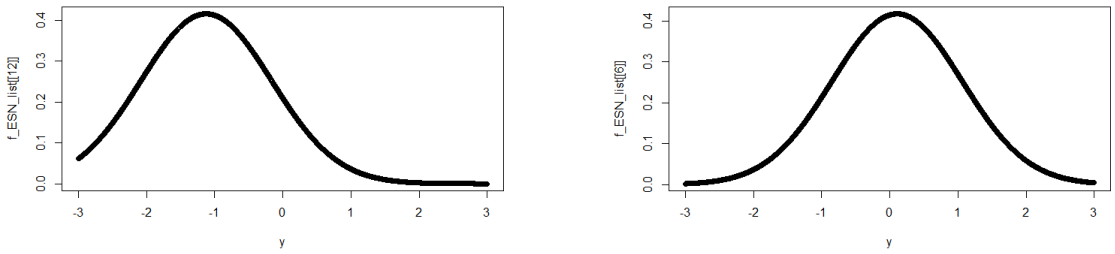
Table 3.7: Mean Absolute Error or MAE (1-7) and CRPS of the density forecast (8). *Note: No - no outliers, Est. - estimation outliers, For - forecast outliers, Both. - estimation and forecast outliers. Varying outlier form leads to similar results as for the MSEs. For $\epsilon = 0.2$ and 0.5, the no contamination case is omitted, as it is identical to $\epsilon = 0.1$ due to the lack of outliers in either case.*

As you may see, the more forgiving nature towards large errors of MAE leads to OLS underperforming compared SSMs. Apart from that, the patterns we are familiar with from the previous sections are mostly present. OLS is still the most susceptible to forecast outliers, while ML is generally influenced the most by estimation outliers. The two-step and robust procedures still dominate the contamination scenarios, while now also being preferred under no contamination. Finally, just like before robustifying the first step improves the forecast while coming at no costs. These results reinforce the theory that it is miss-classification that drives the poorer performance of SSMs, which tend to be quite accurate in the vast majority of observations, but can make some significant errors in a small, but influential minority of observations.

3.3 Results - Probabilistic forecasts

An alternative approach to forecasting in the presence of sample selection would be to resort to a probabilistic forecast. We make use of the theoretical distribution of data affected by sample selection, namely the density of a skew-normal. We resort to the same parametrization as described in section 2.4. We can estimate parameters γ and ρ, and their functions α and τ from the data, and subsequently use these to compute the pdf in (2.16. This is done by allowing y in the same equation to vary from -3 to 3, yielding a pdf on this range. A few examples of how such distributions/forecasts look for a set of realizations and their regressors are shown in Figure 3.3.

Figure 3.3: Density forecasts for 2 realizations

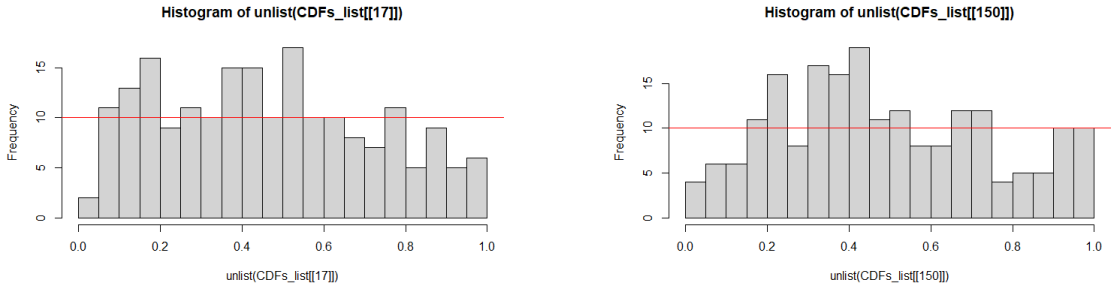


Note: First realization is -2.93, second observation is 2.12

As we can see, the forecast generally correctly predicts observations closer to the correct realization. The skew can vary a lot from one forecast to another, as displayed in the figure, as some observations can appear to be predicted as almost fully normal, while others can be heavily skewed to either the right or the left. The unobserved observations are most often predicted as slightly skewed, but centered around 0.

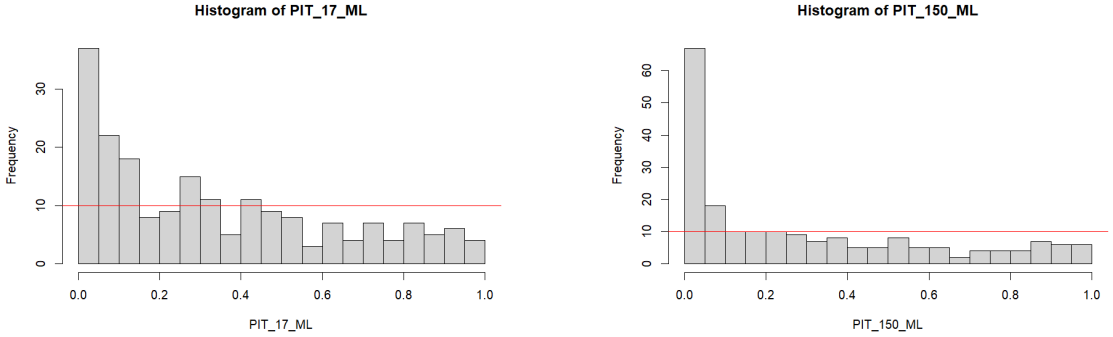
The forecast evaluation is done by means of a continuous ranked probability score (CRPS) described in section 2.4. When it comes to parameter estimation, the correlation coefficient of the errors, ρ, was estimated with two alternative approaches, firstly, the maximum likelihood estimator, which ensures consistency, as well as a a Kendall correlation coefficient. The first one is much more precise and leads to consistent estimates of ρ, while the second clearly holds a downward bias. On the other hand, it appears that the second approach leads to much lower CRPS values, which are lower than the MAEs of any of the original 7 models, while the consistent routine leads to CRPS values that are generally higher than any other model. These values are reported in Table 3.7. These CRPS values, though lower than the previous MAEs, might be misleading. Upon further inspection of the probability integral transforms (PIT), defined in section 2.4, it appears that the PITs are far from a uniform distribution. In fact using the Kendall coefficient leads a forecast which is quite overdispersed. On the other hand the maximum likelihood approach suffers from the opposite problem, it is highly underdispersed. What this effectively means is that we cannot guarantee calibration, in other words we may have inconsistencies between the density forecast and the observation. For the maximum likelihood approach, it would seem this is largely caused by observations close to 0. We display the PITs for the forecast based on both ML and Kendall ρ in Figures 3.4 and 3.5.

Figure 3.4: PIT histograms for Kendall based ρ



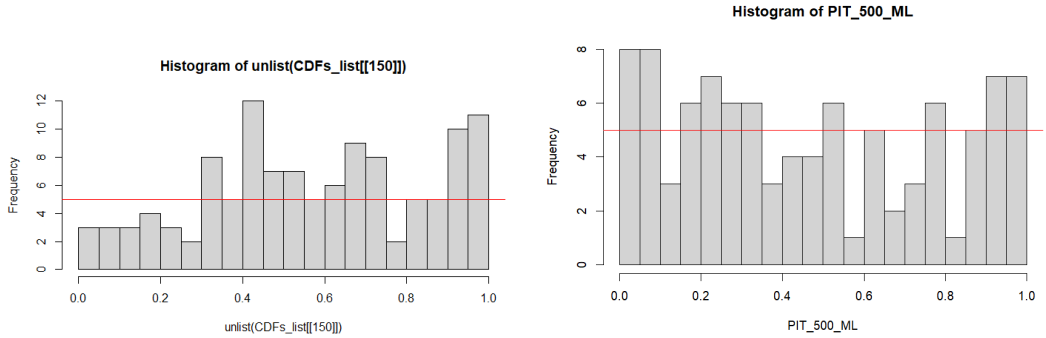
Note: The red line indicates the expected number of observations in each bin given a uniform distribution.

Figure 3.5: PIT histograms for Kendall based ρ



A calibrated forecast would result in a uniform distribution, while our forecasts clearly display some asymmetries. We have checked whether the removal of unobservable realizations remedies the situation (Figure 3.6), and while it does appear to improve the situation somewhat, it does not fix the problem altogether. The results suggest that the Kendall based forecast has good sharpness but is overdispersed, while the ML one is underdispersed and the sharpness is also far from ideal. The failure to secure calibration indicates that these forecasts cannot be trusted, as they do not guarantee consistency, and their predictions may not be reliable.

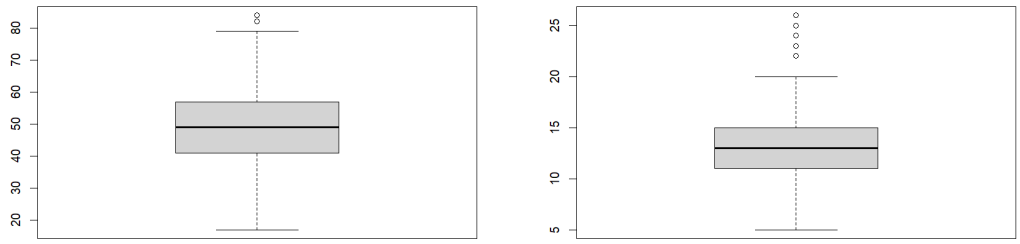
Figure 3.6: PIT histograms Kendall (with exclusion of unobserved values)



Note: First graph shows the histogram of PIT corresponding to non-zero realizations for the ML ρ, while the second graph shows the same for the Kendall ρ. The scales on the left differ.

To visualize the calibration problem at the scale of all samples at the same time, we create boxplots of the number of observations in each bin in every sample. Normally, if a uniform distribution would be respected, the median value should be centered around the expected value, which equals the number of observations divided by the number of bins. In our case, with $N = 200$ observations and 20 bins, this is the same as the one corresponding to the red line, namely 10 observations. This unfortunately is not the case, at least not for all boxplots. For the ML version, we see large discrepancies in boxplots of the first few bins, while for Kendall the issue arises in the middle bins. A demonstration is given by Figures 3.7, while the boxplots for all the bins are provided in Appendix C. Generally bins 1 to 3 hold many more values than expected for the ML version, while for Kendall the anomalous bins are 6-11, which are over-represented. These increased number of PIT values within a bin come at the expense of the later bins for ML, and at the expense of tail bins for Kendall.

Figure 3.7: Boxplots of bins with number of PIT values. left - ML bin 1 , right - Kendall bin 9



Note:Bins reflect number of PIT values that fall within the particular bin. The figure is meant to demonstrate the problematic bins, which hold too many observations for the density forecasts. Please note that for a clearer picture, the boxplots for all the bins need to be considered, these are provided within the appendix.

Chapter 4

Empirical Application

An additional display of the forecasting capabilities of the different methods is given by an empirical application. We consider data from the 2001 Medical Expenditure Panel survey. This choice is motivated by the data set being quite well studied in the literature, see for example the analysis by Cameron and Trivendi (2009). It is also the data set used by Zhelonkin et al. (2016), who have conducted an in-sample comparison of the classical and robust methods. This leads to a good understanding of what to expect in terms of the differences, and any deviations from this expectation must be due to the out-of-sample nature of the analysis. Coupled with the earlier knowledge gained from the simulation, it will lead to significant insights into where the in-sample and forecasting performances diverge.

The data set itself is characterized by 3328 observations of ambulatory expenses supplemented by independent variables such as: age, a gender dummy (female), income, number of years of education (educ) , whether insurance is present or not (ins), number of chronic diseases (totchr), an ethnicity dummy (black), and others which will not be included in the estimation so that results are more comparable to Zhelonkin et al. (2016). Of the 3328, 526 entries are zeroes, constituting almost 16% of the data. The extreme values that some of the ambulatory expenditure takes, coupled with the many unobserved, or zero observations makes this data set ideal for studying sample selection models in a robust setting. A visualization of the ambulatory expenses is provided by Figure 4.1. A considerable number of observations are at or around 0, while some observations are far away in absolute terms. The maximum is at almost 50000, making it considerably higher than the median expenditure at 534.5.

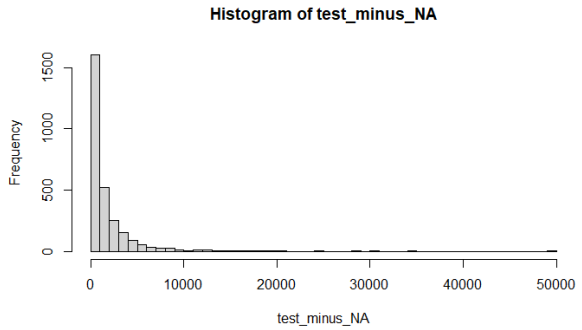


Figure 4.1: Ambulatory Expenditures Histogram

The forecasts are evaluated in terms of MSE and CRPS/MAE, and the 80/20 split into estimation and test samples from the simulation is kept. Once more, we do not evaluate the in-sample performance or significance of the estimates, as a comprehensive assessment is already done in Zhelonkin et al. (2016), and instead focus on the out-of-sample metrics. Table 4.1 is a summary of the results of the empirical application.

	OLS	SSM	R-SSM	SSML-ML	R-SSM-ML	SSMR	R-SSMR	$CRPS^1$	$CRPS^2$
MSE	6.137	7.271	7.271	7.215	7.215	7.434	7.434	-	-
MAE	1.914	1.770	1.770	1.767	1.767	1.779	1.779	1.384	1.434

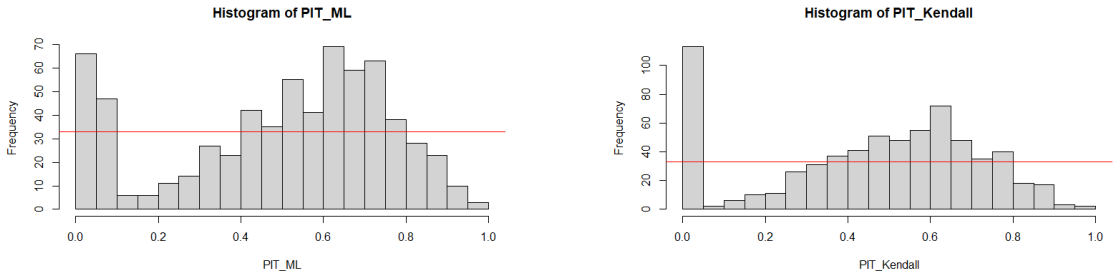
Table 4.1: Table Empirical Application

Note: The performance of the first 7 models is in terms of MAE and MSE, while for the density forecast we report CRPS for both routines, directly comparable to MAE (as for point forecasts CRPS reduces to MAE). $CRPS^1$ *- CRPS of density forecast with ML* ρ *and* $CRPS^2$ *- a similar figure for density with Kendall correlation*

This result is very similar to what we have seen earlier in the simulation, with MSE giving the preference to OLS, while MAE leads to us preferring some form of SSM. It is also noteworthy that despite having around 15% of unobserved entries in the forecasting sample, the probit forecast is almost always a 1, in other words we predict that we will have access to the observation in 97-98 % of the cases. The CRPS values appear to be highly misleading, since once more we cannot guarantee calibration when we analyze the PIT histograms, please see Figure 4.2. Apart from that, density forecasts for individual realizations follow the same trends we have seen earlier. We display these in Figure 4.3.

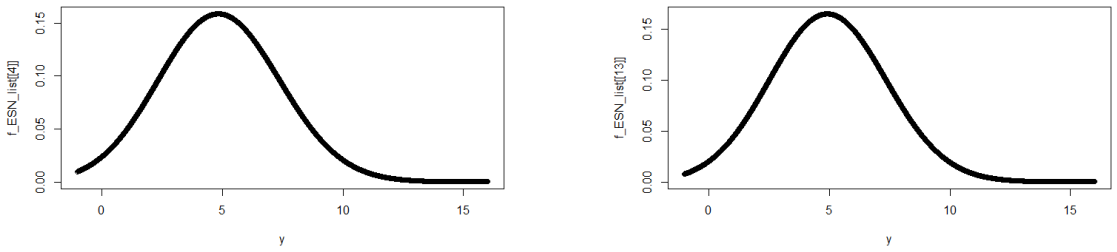
There are perhaps two points of contention between the simulation and the empirical result. Firstly the Maximum Likelihood point forecast is lower than the other two SSM models. The difference, however, is minimal, and when we judge it by MAE, almost non-existent. It is therefore likely to be a statistical fluke. Secondly, CRPS values for the density obtained via the ML ρ appears to be lower than Kendall, which is in contrast to what we have seen before. Again, however, the edge of the ML is negligible. Furthermore, the calibration problem witnessed earlier did not disappear, in fact the PIT histograms look even more un-uniform. Additional trials with the removal of the 0 observations to try to check for the driving force behind the unreliable nature of the density forecast have been made, but none have remedied the issue.

Figure 4.2: Density forecasts for 2 realizations



Note: The red line denotes the point at which each box reaches 5% of the total number of observations. This is the expected number for uniformly distributed PITs, and given 666 forecast observations translates to roughly 33 observations per box. Scales differe across the plots.

Figure 4.3: Density forecasts for 2 realizations



Note: True realizations are both 0 to demonstrate skewness. Non-zero observations generally lead to forecasts that are less skewed towards 0, looking almost like the density of a normal distribution.

In general, the simulation results are confirmed. OLS does a worse job at fitting the data, on the other hand SSMs run the risk of misclassifying which leads to higher errors. This results in OLS being preferred under MSE, while the SSMs are the preferred choice under MAE. Furthermore, the density forecasts are rather unreliable still. Despite minor divergences, the insights gained during the simulation are largely validated.

Chapter 5

Conclusion

Through the extensive simulation study as well as the empirical application, a few things have become much clearer about sample selection models in a forecasting context. Firstly, it is now apparent that forecasting in the presence of non-randomly missing data requires the researcher to take a few fateful decisions. It must be decided how to deal with unobserved values, as well as with forecasts that wrongly predict a value as unobserved. In the case OLS is used, the researcher needs to decide whether the OLS will only be used on observations that are complete, i.e. he has access to all variables within that observation, or whether he must substitute the incomplete values in some way. These decisions are highly dependent on the application, and apriori it might be hard to say which approach is optimal.

Given that the rules for forecasting have been established, a second question arises promptly. Namely, evaluating the forecast becomes an issue of preference and objectives. If the researcher aims to limit cases in which the prediction is vastly wrong, which often arise with sample selection models, then MSE can be used as the loss function. This will, however, down-weight many small errors, which in turn will likely favor OLS heavily. Such forecasts will hardly ever be extremely precise, but they will be reliable enough. On the other hand, if precision with most predictions is desired, while extreme or tail cases that lead to large losses are not as problematic, then one of the sample selection model estimators can be employed, while forecast evaluation should be done by MAE. They are much more precise, but run the risk of miss-classifying an observation, which in many applications can lead to rare but significant losses.

When choosing which SSM to use specifically, it appears that the robust Mallows-type M-estimator has the upper hand under most scenarios, especially when there are outliers suspected in the estimation sample and the number of these outliers is substantial. Finally, the nature of predictions with sample selection models is such, that we likely need to do it in 2 steps. First predict the selection variable, and next predict the outcome only where the selection forecast is non-zero. We have investigated two alternatives for the first step, and have concluded that there is almost no cost to using the robust estimator instead of the classical probit. Additionally, the former also leads to improvements when outliers are present.

In general the addition of outliers leads to deterioration of the forecasts, although the two-step procedures appear to be better suited to dealing with this. In the presence of mild to severe outliers they generally start outperforming the OLS, despite the advantage of not having the risk of misclassification. When it comes to forecasting outliers, our results indicate that OLS

might be more susceptible to such contamination, although it must be noted that this largely depends on the outlier form.

This paper has attempted to employ a density forecast based on the theoretical distribution of the outcome variable. Although the loss function of such forecasts indicates they are reasonably sharp, all the attempts at guaranteeing calibration did not lead to success. It appears that obtaining a well-dispersed forecast when sample selection bias is present is highly problematic.

Overall, OLS should be considered as a safe bet for predicting data with sample selection, while the robust Mallows-type M-estimator with robustified first step prediction can be used for better accuracy but with a slight risk of misclasification. In the presence of outliers the Mallows-type M is preferred.

References

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics*, *24*, 3-61.

Arellano, M. & Bonhomme, S. (2017). Quantile selection models with an application to understanding changes in wage inequality. *Econometrica*, *85*, 1-28.

Arellano-Valle, R. B. & Genton, M. G. (2010). Multivariate extended skew-t distributions and related families. *METRON*, *68*, 201-234.

Buchinsky, M. (2001). Quantile regression with sample selection: Estimating women's return to education in the u.s. *Empirical Economics*, 87-113.

Cameron, A. C. & Trivendi, P. K. (2005). *Microeconometrics: Methods and applications.* Cambridge University Press.

Cameron, A. C. & Trivendi, P. K. (2009). *Microeconometrics using Stata.* Stata Press.

Cantoni, E. & Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, *96*, 1022-1030.

Chernozhukov, V., Fernandez-Val, I. & Luo, S. (2023). Distribution regression with sample selection, with an application to wage decompositions in the uk. *Journal of Political Economy*.

Duan, N., Manning, W. G., Morris, C. N. & Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics*, *1*, 115-126.

Gneiting, T. & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*, 125-151.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359-378.

Granger, C. W. & Pesaran, M. H. (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, *19*, 537-560.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*, 383-393.

Hampel, F. R., Ronchetti, E., Rousseeuw, P. J. & Stahel, W. A. (1986). *Robust statistics—the approach based on influence functions.* John Wiley & Sons.

Heckman, J. J. (1974). Shadow prices, market wages, and labor supply. *Econometrica*, *42*, 679-694.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153-161.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*, 559-570.

Huber, M. & Melly, B. (2015). A test of the conditional independence assumption in sample selection models. *Journal of Applied Econometrics*, *30*, 1144-1168.

Kolassa, S. (2020). Why the "best" point forecast depends on the error or accuracy measure. *International Journal of Forecasting*, *36*, 208-211.

Mahalanobis, C. P. (1936). On the generalized distance in statistics. *The Indian Journal of Statistics*, *80*, S1-S7.

Marchenko, Y. V. & Genton, M. G. (2012). A Heckman selection-t model. *Journal of the American Statistical Association*, *107*(497), 304-317.

Matheson, J. E. & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*, 1087-1096.

Murphy, K. M. & Topel, R. H. (1985). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, *3*, 370.

Ogundimu, E. O. & Hutton, J. L. (2015). A sample selection model with skew-normal distribution. *Scandinavian Journal of Statistics*, *43*, 172-190.

Pagan, A. (1986). Two stage and related estimators and their applications. *The Review of Economic Studies*, *53*, 517.

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, *14*, 53-68.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, 283-297.

Rousseeuw, P. J. & van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*, 212-223.

Zhelonkin, M. (2013). *Robustness in sample selection models* (Doctoral Dissertation). University of Geneva.

Zhelonkin, M., Genton, M. G. & Ronchetti, E. (2016). Robust inference in sample selection models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 805-827.

Appendix A

Impact of Sample Size and Number of Samples

No contamination								
M	N	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
1000	5000	0.711	0.736	0.736	0.751	0.751	0.735	0.735
500	5000	0.713	0.738	0.737	0.753	0.753	0.737	0.737
1000	1000	0.715	0.742	0.743	0.757	0.757	0.742	0.743
1000	10000	0.712	0.736	0.736	0.750	0.750	0.736	0.735
10000	5000	0.711	0.736	0.736	0.750	0.750	0.735	0.735
10000	10000	0.711	0.735	0.735	0.750	0.750	0.734	0.734
Estimation sample contamination								
M	N	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
1000	5000	0.735	0.742	0.738	0.775	0.775	0.734	0.732
500	5000	0.737	0.744	0.740	0.778	0.778	0.737	0.733
1000	1000	0.738	0.750	0.747	0.781	0.782	0.742	0.740
1000	10000	0.735	0.744	0.740	0.774	0.774	0.736	0.732
10000	5000	0.734	0.742	0.738	0.774	0.774	0.734	0.732
10000	10000	0.734	0.742	0.738	0.774	0.774	0.734	0.731
Forecast sample contamination								
M	N	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
1000	5000	1.601	1.149	1.149	1.163	1.163	1.148	1.148
500	5000	1.603	1.151	1.150	1.165	1.165	1.150	1.150
1000	1000	1.606	1.154	1.155	1.168	1.168	1.153	1.154
1000	10000	1.601	1.150	1.150	1.163	1.163	1.149	1.149
10000	5000	1.601	1.149	1.149	1.162	1.162	1.148	1.148
10000	10000	1.601	1.149	1.149	1.162	1.162	1.148	1.148
Estimation and Forecast sample contamination								
M	N	OLS	SSM	R-SSM	SSM-ML	R-SSM-ML	SSMR	R-SSMR
1000	5000	0.872	0.820	0.817	0.850	0.850	0.813	0.811
500	5000	0.874	0.822	0.820	0.853	0.854	0.815	0.814
1000	1000	0.876	0.828	0.826	0.857	0.858	0.821	0.820
1000	10000	0.872	0.821	0.819	0.850	0.850	0.814	0.813
10000	5000	0.872	0.820	0.818	0.850	0.850	0.813	0.812
10000	10000	0.872	0.820	0.818	0.849	0.849	0.813	0.811

Table A.1: Point forecast results in MSEs (impact of the sample size and number of replications)

Appendix B

Boxplots MSEs

Figure B.1: Boxplots MSEs (no contamination)

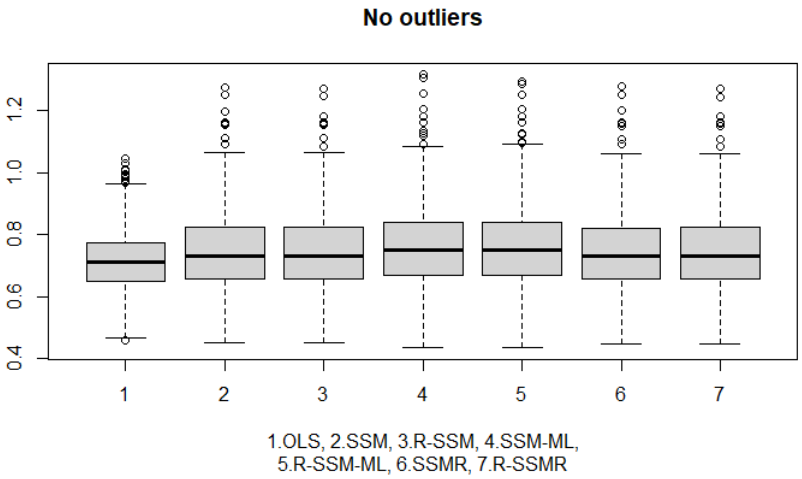


Figure B.2: Boxplots MSEs (estimation)

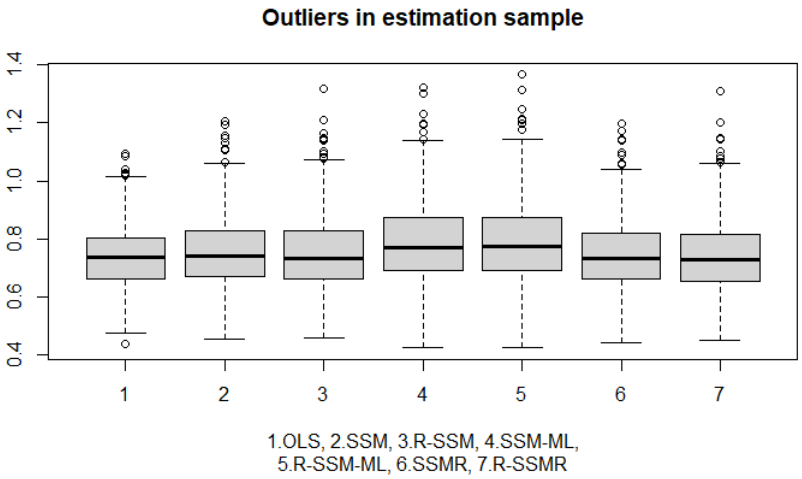


Figure B.3: Boxplots MSEs (forecasting)

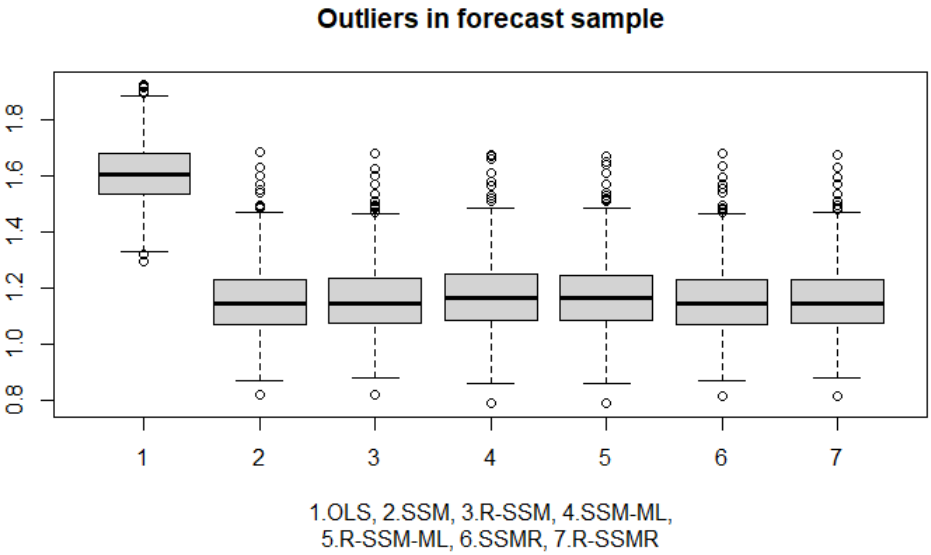
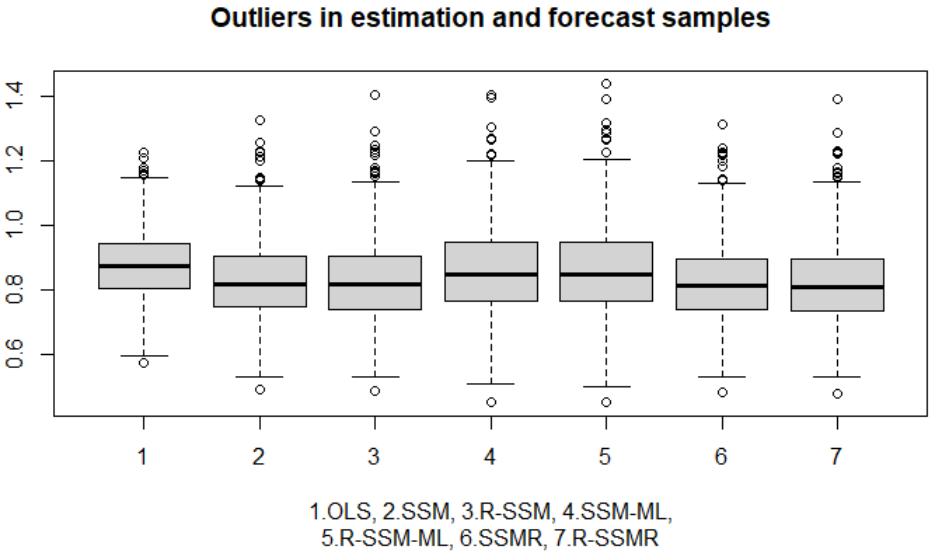


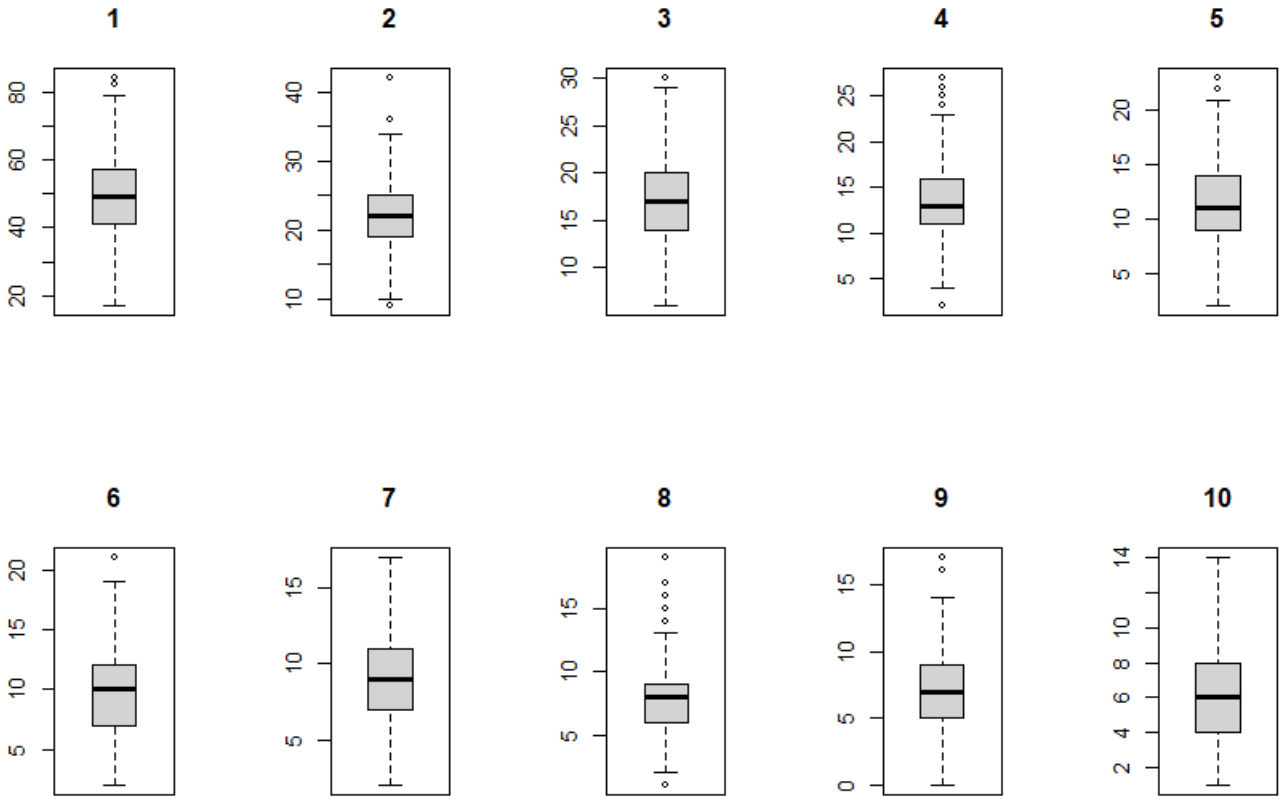
Figure B.4: Boxplots MSEs (estimation and forecasting)



Appendix C

Boxplots for the number of PIT histograms in each bin

Figure C.1: Bins for the ML based procedure (1-10)



Note: We expect the median to be centered roughly at 10 observations for each bin. Significant deviations from this value indicate non-uniformity of the PITs, and therefore are a sign of alarm. This signifies problems with the calibration of the density forecast.

Figure C.2: Bins for the ML based procedure (11-20)

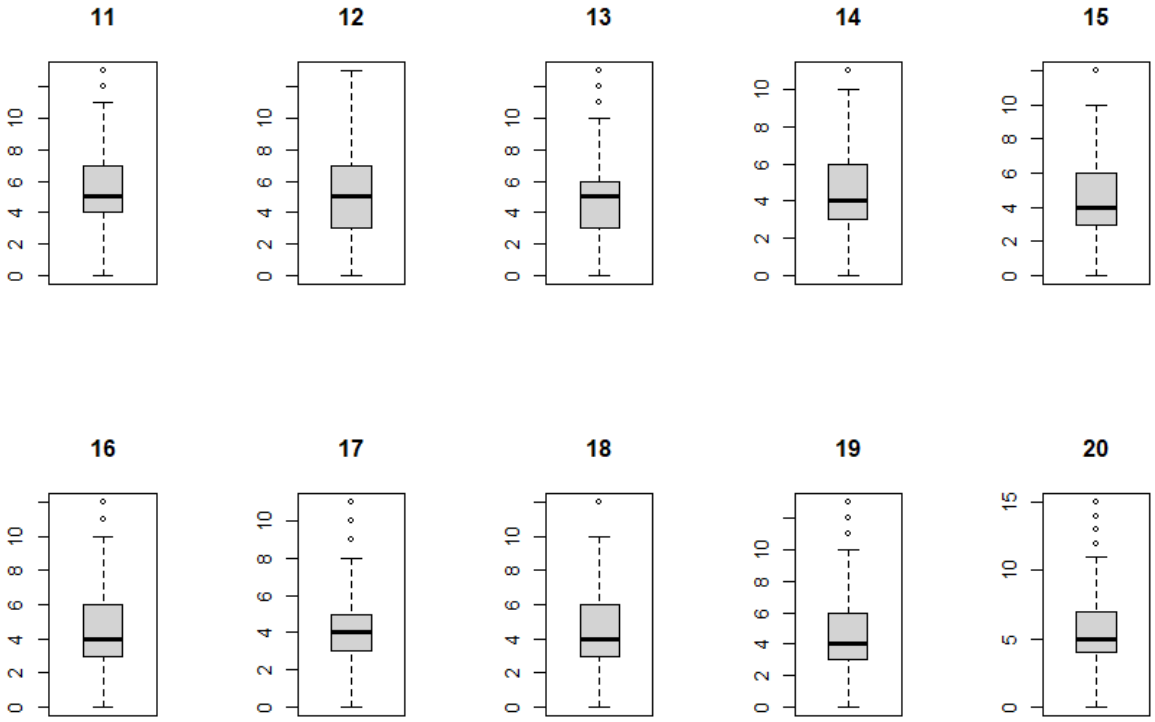


Figure C.3: Bins for the Kendall based procedure (1-10)

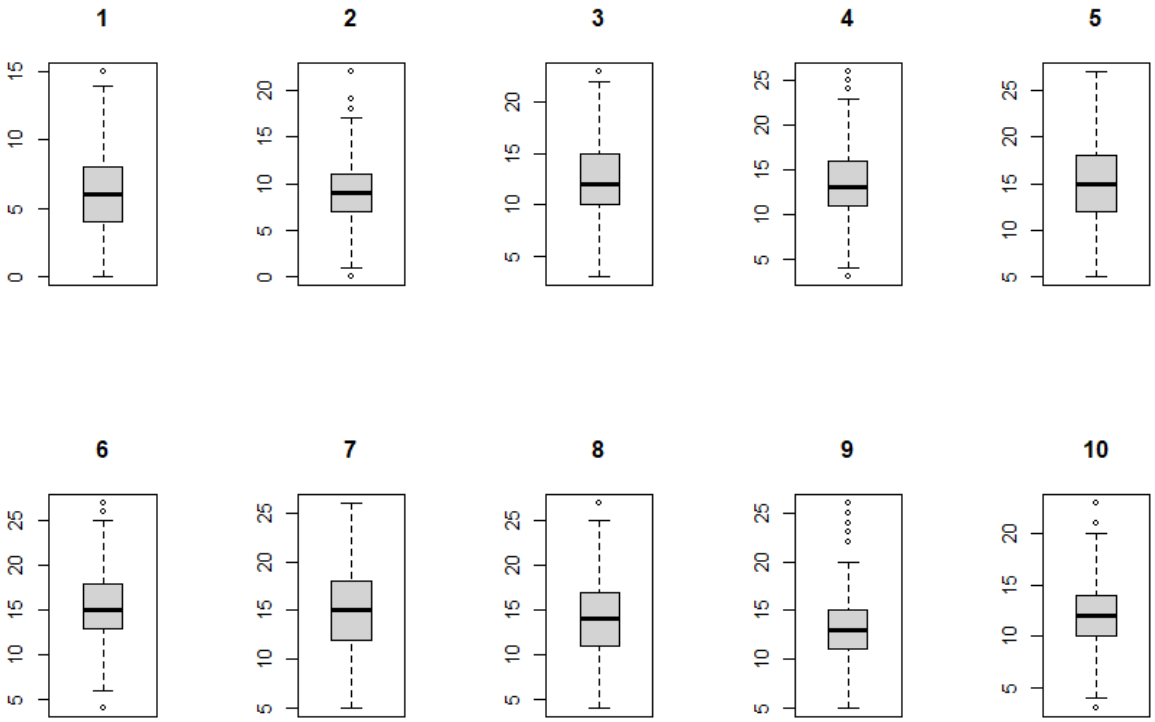
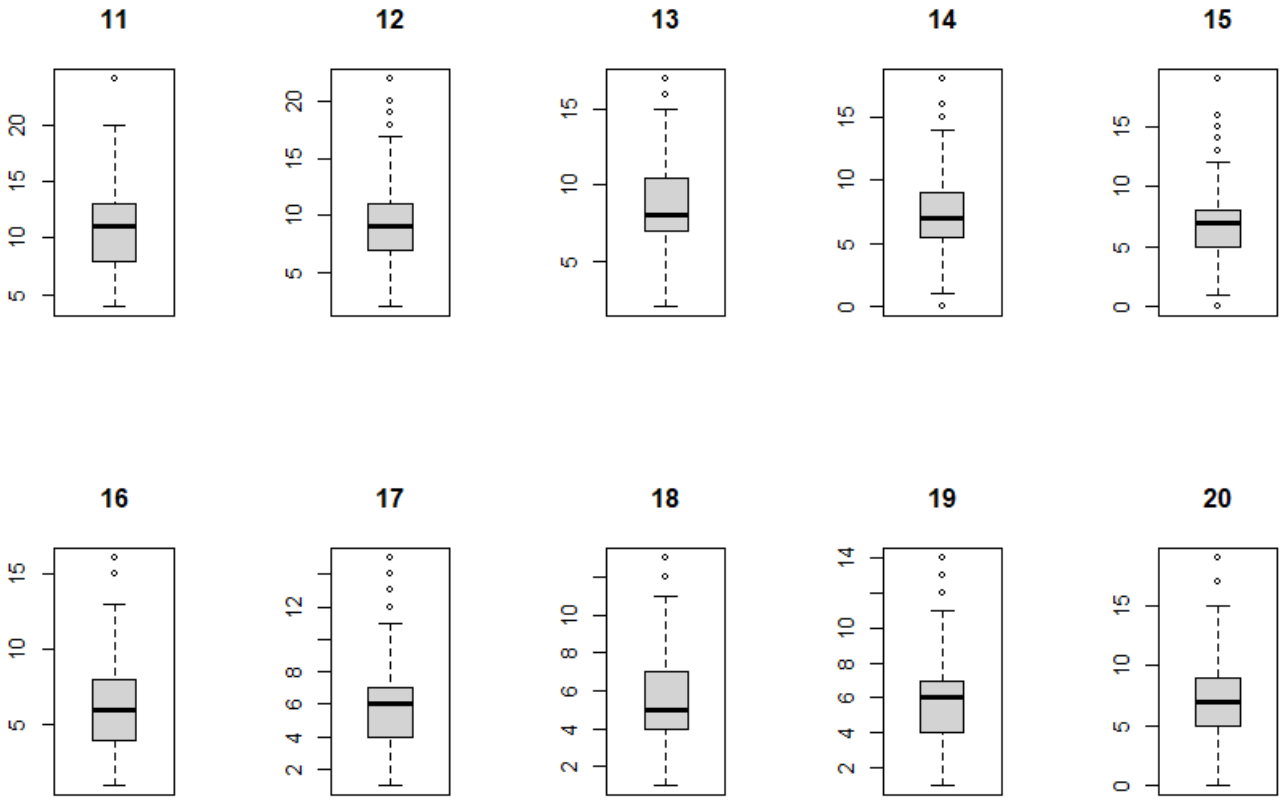


Figure C.4: Bins for the Kendall based procedure (11-20)



Note: Expected value of median of each bin - 10

Appendix D

R code