ESTIMATING ECONOMETRIC MODEL OF AVERAGE TOTAL MILK COST: A SUPPORT VECTOR MACHINE REGRESSION APPROACH

Reet Põldaru, Jüri Roots, Ants-Hannes Viira

Institute of Informatics, Estonian Agricultural University, Estonia

This paper gives an overview of the basic ideas underlying support vector machines (SVM) for regression and function estimation. A summary of currently used algorithms for training SVM is be presented. Application of SVM regression for estimating parameters of econometric model of average total milk cost in Estonian farms is considered and possibilities of application of SVM regression in rural areas are discussed. Studies on implementation of SVM regression methods (algorithms) and software packages in agricultural research and business must be extended.

Key words: econometric models, data mining, support vector machine regression, average total milk cost.

Introduction

In recent years the long-term prospects for the agricultural world markets have been subject of intensive discussions mainly for two reasons:

• The rising concern is given to the food security situation in a number of developing countries.
• The high level of support for the agricultural sectors and large production surpluses in many developed countries, which could be exported only with the extensive use of export subsidies, result in the need for international efforts to liberalise the markets.

Through last decades, the use of economic models in relation to agricultural policy issues has increased substantially, and a big number of literature sources on these issues is available. A number of different modelling approaches have been applied.

Estonia is one of the new members of the European Union. The EU enlargement brings for East European countries a lot of changes in their agriculture. These are changes at political, economic and technical level. This means that information systems on agriculture (databases, models, etc.) have to move along with those changes. Consequently, the economic models in Estonia have to be created, developed and renewed, and must be harmonised with the European requirements. Hopefully, new information technology can be used to lead such evolutions.

A variation in the behavioural characteristics of agricultural production systems over time as well as between countries is recognised. The diverse nature of agricultural production systems and agri-food markets across the EU poses a challenge to anyone seeking to develop a model that can be used to analyse policy at the EU and its member state level.

The guiding principle in constructing the national level commodity models is that the models are first and foremost economic models and such economic theory is our first guide in specifying the models. Economic relationships in the national commodity models are based, in so far as is practicable, on time series econometric estimates of these relationships. Theory and expert judgement are also used in the verification and, if necessary, adjustment of econometrically estimated equations, particularly when used to generate projections for further periods.

Improving the competitiveness of Estonian agriculture is a priority objective of Estonian agricultural policy. The outcome and impacts of those policy actions will strongly depend on development of agricultural world markets.

Dairy sector is the most competitive branch of Estonian agriculture. Consequently, the need to make Estonian dairy farms more competitive is obvious.

New data analysis procedures provided by current data mining (DM) (Andriaans and Zantinge 2003, Dunham 2003, Fayyad et al. 1996, Friedman 1997) have substantially changed the situation in the field of data processing (DP). The situation in data mining is the most challenging one. Data mining, often called knowledge discovery in databases (KDD), started to depart from the statistics and machine learning ghettos and move into the mainstream almost 10 years ago.

Data mining is the process of discovery of useful information from large collections of data. It has common frontiers with several fields including Data Base Management (DBM), Artificial Intelligence (AI), Machine Learning (ML), Pattern Recognition (PR), and Data Visualisation (DV).

The researchers of the Institute of Informatics of the EAU have investigated the possibilities of using some new DM methods and also have some experience in implementing algorithms used in DM packages (Bayesian statistical methods (Põldaru and Roots 2001b, 2001c, 2003b), neural networks (Põldaru and Roots 2002a, 2003a), principal components method (Põldaru and Roots 2001a), decision trees and rules (e.g. CART – classification
and regression trees) (Põldaru et al. 2003a, 2003d), association rules discovery (Põldaru et al. 2003b), fuzzy regression methods (Põldaru et al. 2004b) and support vector machine regression (Põldaru et al. 2004a, 2004b)). At Estonian Agricultural University, some experience has been gained in teaching the new data analysis procedures (principal component analysis, Bayesian methods, neural networks) for postgraduate students (Põldaru and Roots 2001b, 2003b, Põldaru et al. 2003c, Põldaru et al. 2004a). The results of the research are published in many papers and conference theses (see references by Põldaru, Roots, Ruus and Viira).

Considering the reasons outlined above, there is a continuing need for the application of more accurate and informative estimation techniques to econometric analysis.

The objective of this study is estimation of parameters of econometric model of average total milk cost and analysis of the results. The paper provides an overview about the support vector machines regression (SVMR), describes the potential implementation of it in rural areas and discusses the implementation of this method for analysing the dairy sector in Estonia (estimating econometric model of average total milk cost). The data used is an unbalanced panel of milk producers drawn from the FADN (Farm Accountancy Data Network) database of Estonian milk producers. The parameters are estimated on the basis of alternative models of SVM regression using non-linear model. The results are compared mutually and with results of ordinary linear regression. For model (parameter) estimation the SVM module of Programming Environment R is used. R is an integrated suite of software facilities for data manipulation, calculation and graphical display.

### Methods of investigation

Support vector machines have been successfully applied to a number of applications ranging from particle identification, face identification and text categorisation to engine knock detection, bioinformatics and database marketing. The approach is systematic and properly motivated by statistical learning theory (Vapnik 1998). Training (model parameter estimation) involves optimisation of a convex cost function: there is no false local minimum to complicate the estimation process. The approach has many other benefits, for example, the model constructed has an explicit dependence on the most informative patterns in the data (the support vectors), hence interpretation is straightforward and data cleaning could be implemented to improve performance. SVMs are the best known from the class of algorithms, which use the idea of kernel substitution and which we will broadly refer to as kernel methods.

Suppose being given statistical data \((x_1, y_1), \ldots, (x_n, y_n)\). The goal in SVM regression is to find the function \(f(x)\) that has at most \(\varepsilon\) deviation from the actually obtained targets \(y_i\) for all the (training) data, and at the same time, is as flat as possible. SVM regression uses the \(\varepsilon\)-insensitive loss function shown in Figure 1. If the deviation between the actual and predicted value is less than \(\varepsilon\), then the regression function is not considered to be in error.

![Figure 1. A piecewise linear \(\varepsilon\)-insensitive loss function and plot of \(f(x) = a \cdot x + b\) versus \(y\) with \(\varepsilon\)-insensitive tube. Points outside tube are errors](image)

Thus, mathematically it looks like \(-\varepsilon \leq a \cdot x_i + b - y_i \leq \varepsilon\). Geometrically, this can be visualized as a band or tube of size 2\(\varepsilon\) around the hypothesis function \(f(x)\) and any points outside this tube can be viewed as errors (Figure 1). All training (data) points \((x_i, y_i)\) for which \(\left|f(x_i) - y_i\right| > \varepsilon\) are known as support vectors; it is only these points that determine the parameters of \(f(x)\). In other words, errors are not considered as long as they are less than \(\varepsilon\), but any deviation larger than this will not be accepted. To begin the case of simple linear functions \(f(x)\), taking the following form is described:

\[
f(x) = a \cdot x + b
\]

(1).

For estimating the parameters of model (1) this problem can be written as a convex optimization problem (Vapnik 1998):

\[
\text{minimise } \|\theta\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)
\]
subject to \[
\begin{align*}
&y_i - a \cdot x_i - b \leq \varepsilon + \xi_i^i, \\
&a \cdot x_i + b - y_i \leq \varepsilon + \xi_i^i, \\
&\xi_i^i, \xi_i^* \geq 0
\end{align*}
\] (2),

where \(\xi_i^i, \xi_i^*\) are slack variables to cope with otherwise infeasible constraints of the optimization problem (2). The constant \(C > 0\) is specified beforehand. \(C\) is a regularization parameter that controls the trade-off between the flatness of \(f(x)\) and minimizing the training error. If \(C\) is too small then insufficient stress will be placed on fitting the training data. If \(C\) is too large then the algorithm will overfit the training data. The formulation above corresponds to dealing with a so called \(\varepsilon\)-insensitive loss function \(\|e\|\_\varepsilon\) described by

\[
\|e\|_\varepsilon = \begin{cases} 
0 & \text{if } |\mathbf{e}| \leq \varepsilon \\
|\mathbf{e}| - \varepsilon & \text{otherwise}
\end{cases}
\] (3).

It turns out that the optimization problem (3) can be solved more easily in its dual formulation. Hence a standard dual method utilizing Lagrange multipliers will be used. In the case of the Lagrangian dual (supporting) optimisation problem needs to be optimised:

maximise

\[
\left\{ -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (a_i - a_i^*) (a_j - a_j^*) x_i \cdot x_j - \varepsilon \sum_{i=1}^{n} (a_i + a_i^*) + \sum_{i=1}^{n} y_i (a_i - a_i^*) \right\}
\]

subject to

\[
\sum_{i=1}^{n} (a_i - a_i^*) = 0
\]

\[
0 \leq a_i, a_i^* \leq C
\] (4),

where \(a_i\) and \(a_i^*\) are Lagrangian multiplier.

The value of regression parameter \(a\) and predicted value \(f(x)\) can be calculated as follows:

\[
a = \sum_{i=1}^{n} (a_i - a_i^*) x_i
\] (5)

and

\[
f(x) = \sum_{i=1}^{n} (a_i - a_i^*) x_i \cdot x + b
\] (6).

This is the so-called support vector expansion, i.e. the regression coefficient \(a\) can be completely described as a linear combination of the training patterns \(x_i\). In a sense, the complexity of a function’s representation by SVs is independent of the dimensionality of the input space \(X\), and depends only on the number of SVs. Moreover, the complete algorithm can be described in terms of dot products between the data. Even when evaluating \(f(x)\) explicit computing of \(a\) is not needed (although this may be computationally more efficient in the linear setting). These observations will come handy for the formulation of a non-linear extension.

The next step is to make the SVM algorithm non-linear. This, for instance, could be achieved by simply preprocessing the training patterns \(x_i\) by a map into a feature space \(F\), as described in (Vapnik 1998) and then applying the standard SVM regression algorithm.

Firstly, a mapping must be defined from the space \(X\) of regressors to the possibly infinite dimensional hypothesis space \(H\), in which an inner product < , > is defined. This map is formally described as

\[
\Phi: X \rightarrow H \text{ or } x \mapsto \Phi(x)
\] (7).

The choice of regression function \(f(x)\) is limited to the class of functions which can be expressed as inner products in \(H\), taken between some weight vector \(a\) and the mapped regressor \(\Phi(x)\):

\[
f(x) = \langle a, \Phi(x) \rangle + b
\] (8).

The regression function in the hypothesis space is consequently linear, and thus the non-linear regression problem of estimating \(f(x)\) has become a linear regression problem in the hypothesis space \(H\). Note that the mapping \(\Phi(\cdot)\) need never be computed explicitly; instead, the fact that if \(H\) is the reproducing kernel Hilbert space induced by \(k(\cdot, \cdot)\), then writing \(\Phi(x) = k(x, \cdot)\) is used. This gives

\[
\langle \Phi(x_i), \Phi(x_j) \rangle = k(x_i, x_j)
\] (9).

The latter requirement is met for kernels fulfilling the Mercer conditions (Vapnik 1998). These conditions are satisfied for a wide range of kernels, including Gaussian radial basis function (RBF)

\[
k(x_i, x_j) = \exp\left\{-\gamma \cdot (x_i - x_j)^2\right\}
\] (10),

and polynomial function

\[
k(x_i, x_j) = (-\gamma \cdot x_i \cdot x_j + g)^d
\] (11).

It is emphasised that the feature space need never be defined explicitly, since only the kernel is used in SVM regression algorithms. Indeed, it is possible for multiple feature spaces to be included by a single kernel.

Consequently, this allows to rewrite the SV algorithm (formulas (4)…(6)) as follows:

\[
25
The difference to the linear case is that $a$ is no longer explicitly given. However due to the theorem of Fischer-Riesz (see e.g. (Riesz and Nagy, 1955)) it is already uniquely defined in the weak sense by the dot products $\langle a, \phi(x) \rangle$. Also note that in the non-linear setting, the optimization problem corresponds to finding the flattest function in feature space, not in input space.

**Results and discussion**

Next the potential implementation of SVM regression in rural areas is considered and the implementation of this method for estimating an econometric model of the average total milk cost (average total milk cost per kg output) is discussed.

The econometric model is defined by

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_4 \cdot x_4 + b_5 \cdot x_5 + b_6 \cdot x_6 + b_7 \cdot x_7$$  \hspace{1cm} (15),

where $y$ represents the average total milk cost per unit of output (Estonian kroons per kg output), $x_1$ represents the average milk yield per cow (kg), $x_2$ represents the farm total labor input (hours per hectare), $x_3$ represents the manufactured (purchased) milk price (Estonian kroons per 100 kg milk), $x_4$ represents the total labor input per 100 kg of milk (hours), $x_5$ represents the wage per hour (Estonian kroons), $x_6$ represents the total costs of feed per cow (Estonian kroons), $x_7$ represents the invested capital per hectare (Estonian kroons).

The data is an unbalanced panel of milk producers drawn from the FADN (Farm Accountancy Data Network) database of Estonian milk producers. Some previous studies (Põldaru et al. 2004a, 2004b) have also based on the FADN database. The characteristics of the data are reported in Table 1.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>$Y$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$x_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.11</td>
<td>5128</td>
<td>127.7</td>
<td>264.7</td>
<td>5.1</td>
<td>10.4</td>
<td>6889.5</td>
<td>16184.2</td>
</tr>
<tr>
<td>Median</td>
<td>1.99</td>
<td>5013.8</td>
<td>87.2</td>
<td>274.6</td>
<td>4.7</td>
<td>8.1</td>
<td>6122</td>
<td>10189.1</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.68</td>
<td>1154.6</td>
<td>115.7</td>
<td>53.8</td>
<td>2.6</td>
<td>9.3</td>
<td>3486.6</td>
<td>17180.8</td>
</tr>
<tr>
<td>Min</td>
<td>0.81</td>
<td>2447.3</td>
<td>13.2</td>
<td>95.3</td>
<td>0.7</td>
<td>6.0</td>
<td>613.8</td>
<td>425.2</td>
</tr>
<tr>
<td>Max</td>
<td>3.99</td>
<td>7906.7</td>
<td>769.8</td>
<td>409.4</td>
<td>16</td>
<td>44.4</td>
<td>22243.7</td>
<td>115657.9</td>
</tr>
</tbody>
</table>

Table 1. Data summary statistics

The total number of observations $n = 436$.

In the case of the average total milk cost the non-linear functions are the most acceptable and must exhibit characteristics stated in the law of diminishing returns. According to the law of diminishing returns when one or more variable inputs are added to one or more fixed inputs the extra production obtained will, after a point, decline.

For linear and non-linear model (parameter) estimation the SVM module (Meyer 2003) of Programming Environment R (Venables et al. 2003) is used. R is an integrated suite of software facilities for data manipulation, calculation and graphical display.

When using SVM module for any given task, it is always necessary to specify a set of parameters (the parameters must be chosen in advance). Normally the architecture of the SVM is specified in advance and weights and biases are estimated by supervised learning. These parameters include such indexes as whether one is interested in regression estimation or pattern recognition, what kernel is used, what scaling is to be done on the data, etc.

Previous studies (Põldaru et al., 2004a, 2004b) show that the most influential parameters are kernel type, parameter gamma and parameter epsilon. The some non-linear SVM models are sensitive to “overfitting” and
polynomial kernel function gave more acceptable results. Previous experience has been considered for selecting parameter sets in Table 2.

The summary of (given) parameter set for various variants of the non-linear model is reported in Table 2. Two other parameters (parameter $g$ and $d$ in formula (11)) have constant values for all alternatives, whereas $g = 5$, and $d = 3$.

The SVM models in Table 2 are compared with the results of neural network models (varnn1 and varnn2) and ordinary linear regression (OLS). In the case of alternative “varn1” there is one hidden node and in the case of alternative “varn2” – two hidden nodes.

The study shows that the parameter sets in Table 2 give the most acceptable results.

Table 2 presents also result summaries of the results of various model alternatives. Summary characteristics for various alternatives are number of support vectors and coefficient of determination $R^2$.

### Table 2. Parameter set for various alternatives and summary characteristics of the models

<table>
<thead>
<tr>
<th>Variant</th>
<th>Specified set of parameters</th>
<th>Summary characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kernel type</td>
<td>Epsilon</td>
</tr>
<tr>
<td>var1</td>
<td>Polynomial</td>
<td>0.5</td>
</tr>
<tr>
<td>var2</td>
<td>Polynomial</td>
<td>0.4</td>
</tr>
<tr>
<td>var3</td>
<td>Polynomial</td>
<td>0.3</td>
</tr>
<tr>
<td>var4</td>
<td>Radial</td>
<td>0.3</td>
</tr>
<tr>
<td>var5</td>
<td>Linear</td>
<td>0.3</td>
</tr>
<tr>
<td>varnn1</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>varnn2</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>OLS</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

* parameters epsilon, gamma and C are specified for standardised data

Further the summary characteristics in Table 2 are discussed. For different alternatives the number of support vectors are different. The number of support vector depends mainly on the value of parameter epsilon ($\varepsilon$). The SVM models have, in general, offered greater accuracy than have their statistical forebears (OLS and neural network models). The values of the coefficient of determination $R^2$ are higher than in linear model case (OLS and var4 in Table 2). The minimal value in Table 2 (0.886) for non-linear SVM models is higher than maximal value for linear models (0.806). Consequently, the non-linear SVM regression models work well.

For econometric models the values of the coefficient of determination $R^2$ are not the only characteristic for estimating the models generalizing capacity. Next exhibiting of the characteristics, stated in the law of diminishing returns, by selected models is analysed. That may be done by calculating the rules for partial derivatives of the models.

Next the partial derivatives for average milk yield per cow ($x_1$) for considered alternatives are computed and analyzed. The values of partial derivatives are computed from predicted values numerically. The graph of derivatives with respect to independent variables for average milk yield per cow ($x_1$) is shown in Figure 2. The dot line on the graphs presents OLS regression coefficient.

From the economic point of view the derivative for average milk yield per cow, as for a production factor or resource, must be negative (increasing the milk yield the milk cost decreases) and increase (increasing the milk yield the cost decrease diminish).

![Figure 2. Graphs of partial derivatives with respect to value of independent variable for average milk yield per cow](image-url)
From Figure 2, it can be seen that when the Gaussian radial basis kernel is used (var4), the graph of the partial derivative is essentially non-linear (can not be founded from economic point of view), differ substantially from other graphs on Figure 2 and varies within [-0.73, 0.33] (has positive values). Consequently, the case of “overfitting” is observed. Although, in this case, (of the alternative var4) the value of the coefficient of determination is maximal (0.921) that alternative is not acceptable from economic point of view. Therefore, for estimating the parameters of econometric model the Gaussian radial basis kernel can not be recommended.

From Figure 2, it can be seen that when the neural network models are used (varnn1 and varnn2), values of derivatives are negative and the graphs of the partial derivative are concave (can not be founded from economic point of view).

From the economic point of view the alternative var3 is the most acceptable one.

The study shows that the graphs of derivatives with respect to the other independent variables behave analogously. Consequently, the most acceptable kernel from economic point of view is polynomial kernel, and in the following discussion the potential implementation of alternative var1, var2 and var3 (based on polynomial kernel) is considered.

Next the partial derivative for the most essential independent variables is computed and analyzed: average milk yield per cow ($x_1$), total labor input per 100 kg of milk ($x_4$), the wage per hour ($x_5$), and the total costs of feed per cow ($x_6$) for considered alternatives.

The graph of derivatives with respect to independent variables for average milk yield per cow is shown in Figure 3.

![Graph of partial derivatives with respect to average milk yield per cow](image)

Figure 3. Graphs of partial derivatives with respect to value of independent variable for average milk per cow

Analysis of the graphs may bring to the following conclusions:

- The graphs of partial derivatives are analogous (moderately non-linear and convex).
- The non-linear SVM regression models with lower value of epsilon are more flexible (the value of derivative varies more (see var3).
- From the economic point of view the values of partial derivative for average milk yield per cow, as for a production factor, are negative (increasing the milk yield the milk cost decreases) and increase (increasing the milk yield the cost decrease diminishes). Consequently, every relation is acceptable and the considered alternatives can be recommended for practical use.

Next the partial derivative with respect to independent variable for total labor input per 100 kg of milk ($x_4$) is computed and analyzed.

The graph of the partial derivative with respect to the independent variables for total labor input per 100 kg of milk is shown in Figure 4.

The graph shows that the relation for alternative var2 and var3 is essentially non-linear. At the same time for alternatives var1 graph is moderately non-linear and decreasing (Figure 4). Consequently, all considered alternatives are acceptable from economic point of view.

Figure 5 shows the graphs of partial derivatives with respect to wage per hour ($x_5$).

The last graph shows that all relations are moderately non-linear and convex. The graphs of partial derivatives for alternative var2 and var3 have a tendency to increase. At the same time in the case of alternative var1 the graph is decreasing and, consequently, that variant is acceptable from economic point of view.

The graph of derivatives with respect to independent variables for total costs of feed per cow ($x_6$) is shown in Figure 6.
Analysis of these graphs may bring to the following conclusions:

- The graphs of partial derivatives are analogous and moderately non-linear.
From the economic point of view the values of partial derivative for total costs of feed per cow, as for a production factor, are positive (increasing the total costs of feed per cow, the milk cost increases) and decrease (increasing the total costs of feed per cow the cost decrease diminish). Consequently, every relation is acceptable and the considered alternatives can be recommended for practical use.

Figures 3, 4, 5 and 6 and Table 2 show that the considered models are acceptable from economic point of view and the models have high generalising capacity ($R^2$ varies within $[0.886, 0.909]$). In comparison with the other methods, the SVM regression gives better results.

From considered alternative models the most suitable for practical use are the alternatives var1 and var2 when there is polynomial kernel, $\varepsilon = 0.5$ (var1) and $\varepsilon = 0.4$ (var2) and gamma $= 0.08$.

Conclusions

SVM regression provides a new approach to the problem of parameter estimation of linear and especially non-linear econometric models. In this paper a brief exposition of SVM regression and their flexibility in handling economic data is given. Different SVM regression models are used for estimation of the econometric model of average total milk cost in Estonian farms. The results are compared mutually and with results of ordinary linear regression. The discussion may be summarised in the following conclusions:

1. Application of the SVM classification and regression in many fields of science and engineering (including econometrics) is rapidly increasing.
2. The SVM regression models may be used for estimating the parameters of linear and non-linear econometric models.
3. The SVM regression estimates and the least square estimates of econometric model of average total milk cost are similar, whereas the estimates for some independent variables are essentially equivalent.
4. Using of polynomial kernel function gives more acceptable results.
5. The non-linear SVM regression models are sensitive to “overfitting”.
6. The suitable parameter selection allows diminish the “overfitting” problem.
7. Programming Environment R can be used to find the model parameter values.
8. Model combination and Bayesian methods can partially overcome these methods, but require many models to be trained and are hence computationally expensive.
9. SVM regression, as a potential model estimation method, can replace neural networks to solve non-linear problems in econometric modelling.

This analysis has demonstrated that interesting new methods can be implemented for parameter estimation of econometric models. This paper is expected to encourage the use of SVM regression for econometric analysis.

References


