Multiple Regression for High Performance Liquid Chromatography Data Processing

STANISLAVA LABÁTOVÁ
Institute of Informatics
Slovak Academy of Sciences
Dúbravská 9, 845 07 Bratislava
SLOVAKIA
utrrlabt@savba.sk, http://www.sav.sk

Abstract: - Model specification for high performance liquid chromatography using multiple regression analysis is presented. Multiple regression coefficients are established by least square method. An example the dependences of molar masses on retention volumes in THF eluent is shown.

Key-Words: - Multiple regression analysis, model specification, high performance liquid chromatography

1 Introduction
Many of real polymers consist of different chemical macromolecular components. These are mixed intentionally to obtain expected characteristics of system or they arise during the using of polymers (i.e. light degradation). A de:formulation intentionally prepared complex polymers mixtures to render to solve the role of individual elements of system and to try to improve their characteristics in next. Analysis and molecular characterization of system elements that are obtained during construction or degradation poly-reactions improves their better understanding and verifying. The aim of degradation reaction is exact knowledge of process mechanism that allow in the end consequence to suggest the better systems.

Direct analysis of chemical structure of complex polymers and especially of minor (<1%) components is not possible by conventional methods. There are suggested a new unconventional methods of separation of complex multi-component polymer system with emphasis on minor (<1%) components. Experimental results will be mathematically processed in order to obtain improved description and generalized elucidation of retention mechanism. The aim is to describe system with optimal number of parameters, which to do generalized conclusions and predict the behavior of polymer systems.

High performance liquid chromatography (HPLC) belongs to the most important analytical tools of modern chemistry. HPLC separates complex mixtures of both – low and high molecular species to allow conclusions on sample composition in terms of chemical structure and architecture of components and also in terms of their molar mass averages and distributions. HPLC is based on differences in retention of separated substances within appropriate stationary phase. Analyzed phase is introduced into device usually a column filled with stationary phase (eluent). Volume of mobile phase needed to eluate the particular substance from the column is measured. It is retention or elution volume (\(V_R\)) and it reflects extent of retention of analyze substance within the column packing. Concentration of analytes leaving the HPLC columns is monitored by means of detectors which measure light absorption of the column effluent. The dependence the of the detector response on eluent volume is called a chromatogram. Unconventional liquid chromatography methods of separation of complex multi-component polymer systems [3] are mathematically processed in order to obtain improved description and generalized elucidation of the retention mechanisms.

2 Model specification using multiple regression
Multiple regression (MR) analysis is a statistical tool for understanding the relationship between two or more variables in a system. Generally, MR involves a variable to be explained – called the dependent variable – and additional explanatory variables that are thought to produce or be associated with changes in the dependent variable. The important variables that systematically might influence on the dependent variable, and for which
data can be obtained (measured), typically should be included in the model. All remaining influences, which should be small individually, but can be substantial in the aggregate, are included in an additional random error term.

MR is a procedure that separates the systematic effects (associate with the error term) from the random effects and also offers a method of assessing the success of the process.

MR analysis sometimes well suited to the analysis of data about competing theories. There are several possible explanations for the relationship among a number of explanatory variables to assess the statistical data pertinent to these theories.

MR regression also may be useful in:
1. Determining whether or not a particular effect is present
2. Measuring the magnitude of particular effect
3. Forecasting what a particular effect would be, but for an intervening event

Moreover, in interpreting the results of MR analysis, it is important to distinguish between correlation and causality. Two variables are correlated when the events associated with variables occur more frequently together than one would expect by chance. Correlation between two variables does not imply that one event causes the second occur. Therefore, in making causal inferences, it is important to avoid spurious correlation.

MR allows the expert to choose among alternative theories or hypothesis and assists the expert in sorting out correlations between variables that are plainly spurious from those that reflect valid relationships.

Model specification involved several steps, each of which is fundamental to the success of the research effort. Models are often characterized in terms of the parameters – numerical characteristics of the model.

The important variables that systematically might influence the dependent variable, and for which data can be obtained, typically should be included explicitly in a statistical model. All remaining influences, which should be small individually, but can be substantial in the aggregate, are included in additional random error term. MR is a procedure that separates the systematic effects (associated with the explanatory variables) from the random effects (associate with the error term) and also offers a method of assessing the success of the process.

MR uses a sample to obtain estimate of the values of parameters of the model – an estimate associated with a particular explanatory variable is a regression coefficient.

Steps for model specification:
1. Choosing the dependent variable – the variable to be explained should be appropriate variable for analyzing the question in the issue.
2. Choosing the explanatory variable that is relevant to the issues in the case. The explanatory variable that allows the evaluation of alternative hypotheses must be chosen appropriately.
3. Choosing the explanatory variables – an attempt should be made to identify the additional known or hypothesized explanatory variables, some of which are measurable and may support alternative substantive hypotheses that can be accounted for by the regression analyze.
4. Choosing the functional form of the MR model – the expert must also choose the proper form of the regression model. The most frequently selected form of the regression model where the change in the dependent variable associated with the change in any explanatory variables is the same no manner what the level of that explanatory variables will have differential effects on the dependent variables as the values of explanatory variable change. In this case the expert should be consider the use nonlinear system.

Regression results can be interpreted in purely statistical terms, through the use of significance tests, or they can be interpreted in a more practice, non-statistical manner [2]. In our case, we use the last one manner.

3 MR linear model for HPLC
In generally, linear regression model takes the following form:

\[ Y = k_0 + k_1X_1 + k_2X_2 + \ldots + k_nX_n + \epsilon \quad (1) \]
where \( Y \) represents the dependent variable, \( X_1, X_2, \ldots, X_n \) represent the explanatory variables and \( \varepsilon \) – error which represents the collective unobservable influence of any omitted variables. Coefficients \( k_0, k_1, \ldots \) represent multiple regression coefficients which are estimated by fitting the equation to the data some appropriate method.

Using the HPLC method we obtain experimental data set

\[
\{V_R, \log M, \text{THF}\}, \; i = 1, 2, \ldots, n, \tag{2}
\]

where \( V_R \) represents retention volume, \( \log M \) log molar mass, \( \text{THF} \) concentration of eluent

Experimental data for polyst. in neat THF and mixed eluent THF + DMF are shown on the Table 1:

<table>
<thead>
<tr>
<th>M</th>
<th>T</th>
<th>V</th>
<th>NV</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.82</td>
<td>1.000</td>
<td>7.200</td>
<td>7.640</td>
</tr>
<tr>
<td>3.41</td>
<td>1.000</td>
<td>6.200</td>
<td>6.863</td>
</tr>
<tr>
<td>3.95</td>
<td>1.000</td>
<td>5.660</td>
<td>6.152</td>
</tr>
<tr>
<td>4.24</td>
<td>1.000</td>
<td>5.320</td>
<td>5.770</td>
</tr>
<tr>
<td>4.57</td>
<td>1.000</td>
<td>5.250</td>
<td>5.335</td>
</tr>
<tr>
<td>5.00</td>
<td>1.000</td>
<td>5.180</td>
<td>4.769</td>
</tr>
<tr>
<td>5.37</td>
<td>1.000</td>
<td>5.160</td>
<td>4.282</td>
</tr>
<tr>
<td>5.70</td>
<td>1.000</td>
<td>5.170</td>
<td>3.847</td>
</tr>
<tr>
<td>2.82</td>
<td>0.200</td>
<td>10.030</td>
<td>11.430</td>
</tr>
<tr>
<td>3.41</td>
<td>0.200</td>
<td>10.110</td>
<td>10.653</td>
</tr>
<tr>
<td>3.95</td>
<td>0.200</td>
<td>10.030</td>
<td>9.942</td>
</tr>
<tr>
<td>4.24</td>
<td>0.200</td>
<td>9.550</td>
<td>9.560</td>
</tr>
<tr>
<td>4.57</td>
<td>0.200</td>
<td>8.430</td>
<td>9.125</td>
</tr>
<tr>
<td>5.00</td>
<td>0.200</td>
<td>5.750</td>
<td>8.559</td>
</tr>
<tr>
<td>5.37</td>
<td>0.200</td>
<td>5.160</td>
<td>8.071</td>
</tr>
<tr>
<td>5.70</td>
<td>0.200</td>
<td>5.170</td>
<td>7.637</td>
</tr>
</tbody>
</table>

\[
2.82 \quad 0.190 \quad 10.210 \quad 11.477
\]

\[
3.41 \quad 0.190 \quad 10.620 \quad 10.700
\]

\[
3.95 \quad 0.190 \quad 10.790 \quad 9.989
\]

\[
4.24 \quad 0.190 \quad 11.150 \quad 9.607
\]

\[
4.57 \quad 0.190 \quad 10.030 \quad 9.172
\]

\[
5.00 \quad 0.190 \quad 7.170 \quad 8.606
\]

\[
5.37 \quad 0.190 \quad 5.810 \quad 8.119
\]

\[
5.70 \quad 0.190 \quad 5.120 \quad 7.684
\]

\[
2.82 \quad 0.180 \quad 10.150 \quad 11.525
\]

\[
3.41 \quad 0.180 \quad 10.660 \quad 10.748
\]

\[
3.95 \quad 0.180 \quad 11.070 \quad 10.036
\]

\[
4.24 \quad 0.180 \quad 11.350 \quad 9.654
\]

\[
4.57 \quad 0.180 \quad 11.250 \quad 9.220
\]

\[
5.00 \quad 0.180 \quad 9.370 \quad 8.653
\]

\[
5.37 \quad 0.180 \quad 8.550 \quad 8.166
\]

\[
5.70 \quad 0.180 \quad 5.760 \quad 7.731
\]

\[
2.82 \quad 0.170 \quad 11.000 \quad 11.572
\]

\[
3.41 \quad 0.170 \quad 11.480 \quad 10.795
\]

\[
4.24 \quad 0.170 \quad 12.550 \quad 9.702
\]

\[
4.57 \quad 0.170 \quad 12.750 \quad 9.267
\]

\[
5.00 \quad 0.170 \quad 12.060 \quad 8.701
\]

\[
5.37 \quad 0.170 \quad 11.410 \quad 8.213
\]

\[
e = 1.658
\]

\[
k_1 = -1.3170526
\]

\[
k_2 = -4.7372490
\]

\[
k_0 = 16.0913580
\]

Table 2. Results of multiple regression for experimental data from Table 1

where

\[ NV - \quad \text{is a new value of estimate retention value,} \]

\[ k_0, k_1, k_2 - \quad \text{are estimate regression coefficients,} \]

\[ \varepsilon - \quad \text{is error,} \]

\[ M - \quad \text{represents log of molar mass,} \]

\[ T - \quad \text{is concentration of eluent} \]

\[ V - \quad \text{is retention volume} \]

Dependent variable \( V_R \) we wanted estimate in the form:

\[
V_R = k_0 + k_1 \log M + k_2 \text{THF} + \varepsilon \quad (3)
\]

Using least square method for estimate regression coefficient we obtain values as is shown in Table 2:

![Kromasil C-18, 100A, 5um, 2x150x7.8 mm](image)

Table 1

<table>
<thead>
<tr>
<th>log M</th>
<th>Mw</th>
<th>20% THF</th>
<th>20% THF</th>
<th>20% THF</th>
<th>20% THF</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.82</td>
<td>666</td>
<td>7.2</td>
<td>10.03</td>
<td>10.21</td>
<td>10.15</td>
</tr>
<tr>
<td>3.41</td>
<td>2600</td>
<td>6.2</td>
<td>10.11</td>
<td>10.62</td>
<td>10.66</td>
</tr>
<tr>
<td>3.95</td>
<td>9000</td>
<td>5.66</td>
<td>10.03</td>
<td>10.79</td>
<td>11.07</td>
</tr>
<tr>
<td>4.24</td>
<td>17500</td>
<td>5.32</td>
<td>9.55</td>
<td>10.57</td>
<td>11.35</td>
</tr>
<tr>
<td>4.57</td>
<td>37000</td>
<td>5.25</td>
<td>8.43</td>
<td>10.03</td>
<td>11.26</td>
</tr>
<tr>
<td>5</td>
<td>100000</td>
<td>5.18</td>
<td>5.75</td>
<td>7.17</td>
<td>9.37</td>
</tr>
<tr>
<td>5.37</td>
<td>233000</td>
<td>5.16</td>
<td>5.22</td>
<td>5.81</td>
<td>8.55</td>
</tr>
<tr>
<td>5.7</td>
<td>498000</td>
<td>5.17</td>
<td>5.07</td>
<td>5.12</td>
<td>5.67</td>
</tr>
</tbody>
</table>

![Fig.1. Dependences of retention volume on molar masses](image)
Using the experimental data from Table 1 we can express the $V_R$ by next form:

$$V_R = -1.31705 \log M - 4.7372 \text{THF} + 16.0913 + \varepsilon \quad (4)$$

The error $\varepsilon$ is connected with other events that are not expressed in (3). In future we intend to include further additional explained variables that express $V_R$ and that have the influence on interpretation of experiment. In [3] is $V_R$ expressed only as dependences on $\log M$. It is evident that process of HPLC is the function of several other variables.

4 Conclusion

Using multiple regression for establish regression coefficients represents a new form of explanation dependent variable molar mass by dependence on more variables than one. In the past there are studied dependencies molar mass only on one variable – retention volume. It was realized by polynomial regression 2nd and 3rd order. HPLC represents many different events that have influence on this process and that were in previous case not consider. In next we will busy with next events and additive variables that have some importance in this problem.

Acknowledgements: Financial support from the Slovak Grant Agencies APVV (project 0592–07). The author is thankful Dr. D. Berek for his advises and to Mrs. J. Tarbajovska for her excellent technical assistance.

References