Investigating the Impact of Operating Parameters on Molecular Weight Distributions Using Functional Regression

R. A. Hutchinson,[1] *P. J. McLellan,**[1] *J. O. Ramsay,*[2] *H. Sulieman,*[3] *D. W. Bacon*[1]

[1]Department of Chemical Engineering, Queen's University, Kingston, Ontario, Canada K7L 3N6
E-mail: mclellnj@chee.queensu.ca
[2]Department of Psychology, McGill University, Montréal, Québec, Canada H3A 2T5
[3]Department of Mathematics and Statistics, American University of Sharjah, United Arab Emirates

Summary: Molecular weight distributions (MWDs) are inherently functional observations in which differential weight fraction is expressed as a function of chain length. Conventional approaches for analyzing and predicting MWDs include discretization and treatment as multi-response estimation problems, characterization using moments, and detailed mechanistic modeling to predict fractions for each chain length. However, these approaches can be sensitive to loss of information, complexity and problem conditioning. An alternative is to treat the MWDs as functional observations, and to use techniques from Functional Data Analysis (FDA), notably functional regression. The objective of this paper is to develop and apply empirical modeling techniques based on functional regression for investigating the impact of operating parameters on MWDs.

Keywords: empirical models; functional data analysis; functional regression; molecular weight distribution modeling; polystyrene

Introduction

Molecular weight distributions (MWDs) summarize the occurrence of macromolecules as a function of chain length, describing differential weight fractions of differential chain length intervals. Consequently, they are functional quantities, with weight fractions expressed as a function of chain length. The MWD of a polymer has potential implications for its end-use

 DOI: 10.1002/masy.200450238

properties and processibility. Indeed, these end-use properties may themselves be described in terms of functional behaviour, such as stress-strain relationships.

While the focus of reactor operation may be on producing a polymer of a specific MWD, ultimately the goal is to obtain the desired end-use properties through judicious selection of reactor operating conditions. These conditions can be selected with the aid of a process model that predicts MWD as a function of operating parameters. Both mechanistic and empirical models have been proposed for such tasks, with the former having greater complexity. Describing the relationship between MWD and end-use properties can be harder still, and is most likely to be described using empirical models [3]. While the prediction of end-use properties given operating conditions is the ultimate goal of this research, we have focused in the first step on developing empirical models for predicting the impact of reactor operating conditions on MWDs.

There are several conventional approaches for analyzing and predicting MWDs, including: detailed mechanistic modeling to predict fractions for each chain length, characterization using moments, or discretization and treatment as a multi-response estimation problem. Detailed mechanistic modeling [2] brings substantial complexity both in terms of the model equations as well as the number of parameters to be estimated. Characterization using moments collapses information about the MWD into key moments [4]. Finally, discretization and treatment as a multi-response estimation problem can produce ill-conditioned statistical estimation problems.

An alternative approach is to treat the MWDs as functional observations and to use functional regression to develop an empirical model. Functional regression is a technique in the broader field of Functional Data Analysis (FDA), which is a statistical framework designed to deal with situations in which the data objects are functions as opposed to individual data points. As we will show in this paper, functional regression provides a natural framework for estimating empirical models for predicting MWDs.

The objective of this paper is to develop and apply empirical modeling techniques based on functional regression for investigating the impact of operating parameters on MWDs. The

application of these techniques is demonstrated by developing an empirical model predicting the impact of reactor temperature and initial initiator concentration on MWDs in a bulk polymerization of polystyrene.

Functional Regression and FDA

Functional Data Analysis [8] is a relatively new branch of applied statistics in which the data objects are functions as opposed to individual points. FDA provides a comprehensive statistical framework, including summary statistics, principal components analysis, and functional regression. In this work, functional regression is used to estimate an empirical model predicting the impact of isothermal reactor temperature (T) and initial initiator concentration [I] on the resulting MWD in a bulk polymerization of styrene. Synthetic data have been generated using Predici® for a two-level factorial design (2^2 design) in T and [I].

Conceptually, we will develop a regression model that predicts a functional observation, namely the MWD, given values of the non-functional variables T and [I]. Since the response is functional, the regression coefficients will themselves be functions of log(MW). While conceptually the response is a functional observation, practically the response is measured as a series of observations of weight fractions at (not necessarily uniform) log(MW) observation points. Thus, the first step in the data analysis is to convert these observations into a functional data object. This is accomplished by representing the raw data in a suitable functional basis (e.g., B-splines).

Functional regression analyses typically consist of the following steps: 1) smooth the raw data, expressing them as functional objects in a suitable basis; 2) formulate a functional regression model, 3) estimate parameter functions in order to minimize an integral squared prediction error criterion between the observed functional response and the predicted functional response, and 4) assess diagnostics, adjust model accordingly, and re-estimate parameters as necesssary. The estimation in step 3 is typically performed by representing the coefficient functions in a suitable basis [8]. Additional details are provided below. A B-spline basis was used for all smoothing and parameter function representation.

The Data

Data were generated using Predici® at each of 4 high-low combinations of T and [I] according to a standard 2^2 factorial design [1] summarized in Table 1. The bulk polymerization was simulated for an isothermal batch reactor that included both chemical and thermal initiation mechanisms ([2],[4]), and the simulation stipulated that the polymerization was to go to complete conversion of monomer. A standard coding to +/-1 (i.e., (x-design midpoint)/1/2 range) was used.

Table 1: Run Conditions for the 2^2 Experimental Design

Run	Temperature (°C)	Initial Initiator Concentration (mol/L)
hihi	150	0.02
hilo	150	0.005
lohi	130	0.02
lolo	130	0.005

The four resulting MWDs are summarized in Figure 1. This plot reinforces the notion that the responses are functional, while the operating factors (regressors) are not. There is a marked change in the degree of symmetry of the MWDs, with pronounced shoulders at low values of [I].

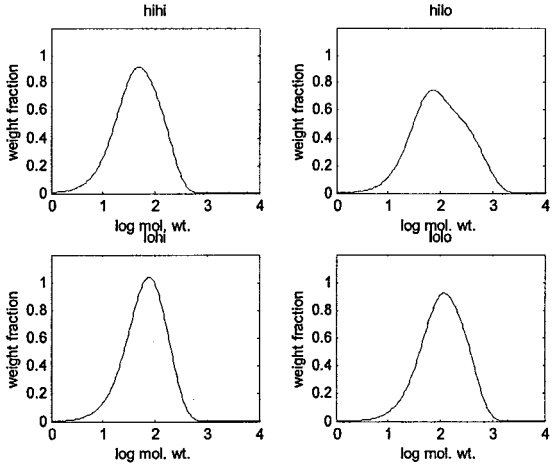

Figure 1: MWDs for the four operating points of the 2^2 experimental design

The Regression Model

The MWD weight fractions are inherently non-negative by definition. Moreover, fundamental modeling has shown that MWDs frequently belong to an exponential family of distributions. The exponential family of distributions can be considered to have density functions g(r) of the form

$$g(r) \propto \exp(W(r)) \tag{1}$$

where W(r) is the kernel of density function (see for example [7]). Ramsay [7] has shown that density functions for the exponential family of distributions can be realized as the solution of the following differential equation:

$$\frac{dg}{dr} = w(r)g(r) \tag{2}$$

Since g(r) is inherently non-negative, we can divide both sides by g(r),

$$\frac{dg/g(r)}{dr} = \frac{d(\ln(g(r))}{dr} = w(r) \tag{3}$$

and finally the density function g(r) can be obtained as

$$g(r) = C\exp\left\{\int_{r_0}^{r} w(u)du\right\} = C\exp(W(r)) \tag{4}$$

The advantage of this approach is that it ensures non-negativity of the predicted weight fractions. Thus, the regression modeling in this paper begins by transforming the response y(r), the weight fractions of the MWD, and working with d(ln(y(r))/dr as the response in the regression analysis. Note that "r" is $\log_{10}$(MW), where MW has units of kg/mol.

The regression model is:

$$\frac{d\ln(y(r))}{dr} = \beta_0(r) + \beta_1(r)T + \beta_2(r)[I] + \beta_{12}(r)T[I] + \varepsilon \tag{5}$$

where the response d(ln(y(r))/dr can be considered as a transformed response z(r), and ε is an independent, identically distributed random error component. Note that because the response is functional, the parameters are themselves functions of log(MW), denoted "r" in kg/mol. The parameter functions describe the impact of the related operating factor (e.g., T in the case of $\beta_1(r)$) over the MWD. Thus, if T has a more significant impact in the middle of the MWD, we would expect $\beta_1(r)$ to have greater magnitude in "r" values in the middle range of the distribution.

Once a satisfactory model has been estimated in z(r), we can obtain a model predicting the MWD by integrating and exponentiating the fitted model:

$$y(r) = C\exp\{\gamma_0(r) + \gamma_1(r)T + \gamma_2(r)[I] + \gamma_{12}(r)T[I]\} \quad (6)$$

where

$$\gamma_i(r) = \int_{r_0}^{r} \hat{\beta}_i(u)du,\ i = 0,1,2,12$$

The scaling, or integrating, constant C can be computed to normalize the area under the predicted MWD.

Finally, the parameter function estimates are computed by minimizing a least squares criterion, namely the integral squared prediction error between the observed MWD functional object and the MWD predicted by the regression model:

$$ISE = \sum_{i=1}^{4} \int_{r_0}^{r_1} \left(y_i(r) - \hat{y}_i(r)\right)^2 dr \quad (7)$$

The total integral squared prediction error (ISE) is obtained by summing the ISE values for each MWD in the dataset. All smoothing and functional regression calculations were computed using the FDA toolbox [6] in Matlab™. The FDA toolbox is available from Jim Ramsay's web site [6].

Model Estimation Results

The data were first log-transformed, and are shown below in Figure 2. The log transformation accentuates the peak areas, and converts the tails into steeper, almost linear, segments. The log-transformed data were converted to functional data objects using spline smoothing. Fourth-order B-splines were used for all functional data and parameter functions in this analysis. The knots were located at points at which weight fractions for the MWDs were provided by Predici®. These points were generally non-uniform, both in terms of spacing for a specific MWD, and between MWDs. Fourteen basis functions were used in the spline smooths of the data and in the parameter function representations.

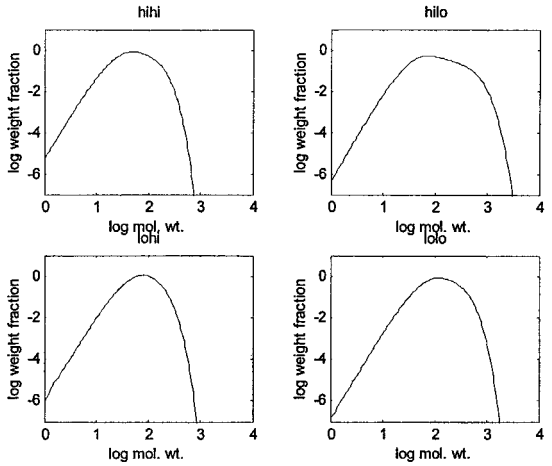

Figure 2: ln(MWD) profiles for the four operating points of the 2^2 experimental design

Each log-transformed MWD in the dataset was smoothed and converted to a functional data object. Noise in the tails can be problematic. For example, if one works in the original MWD domain, the smoothing problem is ill-conditioned at the tails relative to the central area of the distribution because of the difference in magnitudes of the weight fractions. In the log-transformed case, spurious jumps from 10^{-4} to 10^{-8} in the tail areas produce substantial rippling behaviour in ln(y(r)) at high and low r values. In our analysis, we chose to truncate the interval for log(MW) over which the MWDs were considered to the interval [1,3] (MW in units of kg/mol).

After smoothing the log-transformed MWDs, the resulting functional objects were differentiated in order to obtain the responses z(r) = d(ln(y(r))/dr shown below:

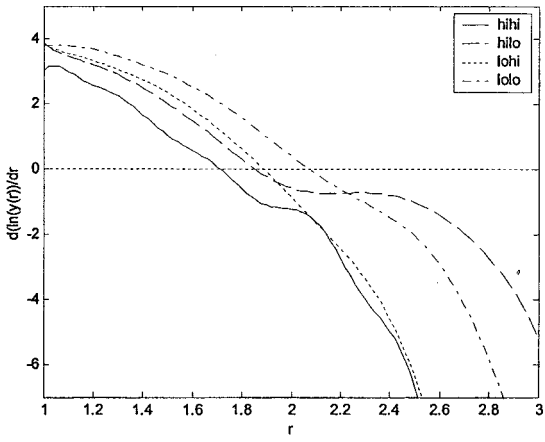

Figure 3: d(lny(r))/dr profiles for the four operating points of the 2^2 experimental design

Differentiation accentuates the shoulders in the distributions. For example, the MWD for the high T - low [I] case ("hilo") has a particularly pronounced segment in which the first derivative is constant over a log(MW) range from 2 to 2.4, and then decreases. This is associated with the pronounced shoulder in the original MWD for this case. The corresponding second derivative would be 0 over this interval, and would then become negative.

The parameter function estimates were computed by representing the parameter functions in the spline basis described above, and estimating coefficients for this basis representation [8]. The resulting estimates are shown below in Figure 4. The responses (d(ln(y))/dr) are also presented to provide a basis against which to compare magnitudes of the parameter function estimates.

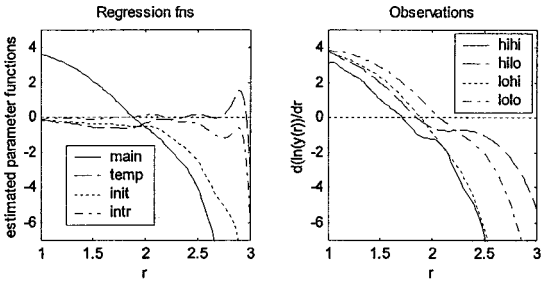

Figure 4: Parameter function estimates and transformed responses

The "main" curve is the mean effect, i.e., the intercept parameter function. This function provides the base MWD in the original (non-transformed) domain. Examining the parameter function estimates for temperature, initiator, and their two-factor interaction ("intr") indicates that the these factors have their main influence at higher log(MW) ranges, as expected.

The predicted MWDs for each of the operating points can be obtained by reversing the transformation as described earlier. The resulting predictions are shown below in Figure 5. The predicted MWDs match the observed MWDs almost exactly. This result is consistent with what one would expect when fitting a main effects plus two-factor interaction model to a conventional problem in which non-functional data (e.g., yield) were collected using a 2^2 design.

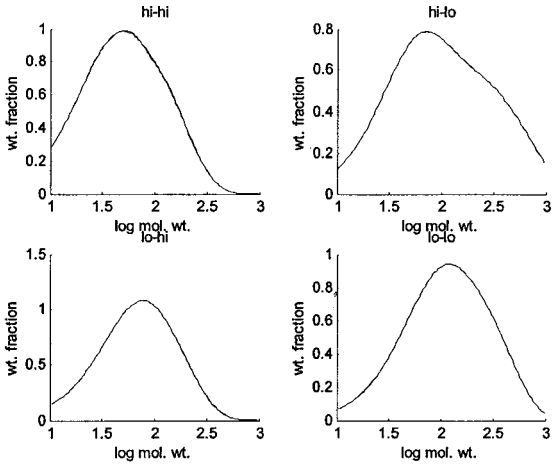

Figure 5: Predicted and observed MWDs for the four operating points of the 2^2 experimental design

The need to retain the two-factor interaction term can be confirmed by examining the predicted MWDs for the model in which this term has been deleted. As Figure 6 indicates, some predictive ability is lost for the hilo and lolo cases in which the initiator concentration is low.

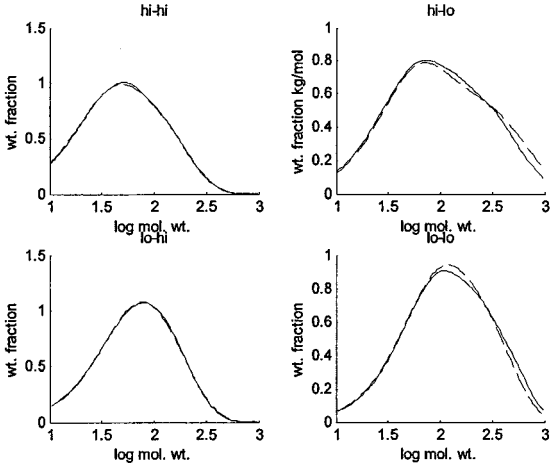

Figure 6: Predicted (solid line) and observed (dashed line) MWDs for the four operating points of the 2^2 experimental design

In conventional analyses of factorial screening designs, one is often interested in the magnitudes of the main effects [1]. Each main effect represents the sensitivity of the response to perturbations in a given operating variable. These can be computed in terms of the parameter estimates for the main effect terms, or summarized graphically using effects plots. We could perform a similar analysis in terms of the parameter function estimates for the transformed response d(ln(y))/dr. However, we are more interested in the impact of the operating factors on the original MWD. The parameter functions appear nonlinearly in the MWD (in the exponential and normalization constant). The best way to summarize the sensitivity information is to take the derivative of the predicted response with respect to the operating variable of interest ($\partial y(r,x)/\partial x$, where x is T or [I]), and to plot the sensitivity as a function of r. The sensitivities can be examined at different operating points. Figure 7 depicts the sensitivity of weight fraction with respect to temperature and initiator concentration, evaluated at the centre point of the design (i.e., (0,0) in coded units). Temperature has a predominantly symmetric effect on the MWD. Increasing T at the centre point reduces the weight fractions, thus broadening the distribution. In contrast, initiator has an asymmetric impact on MWD, with a more pronounced impact at higher values of r.

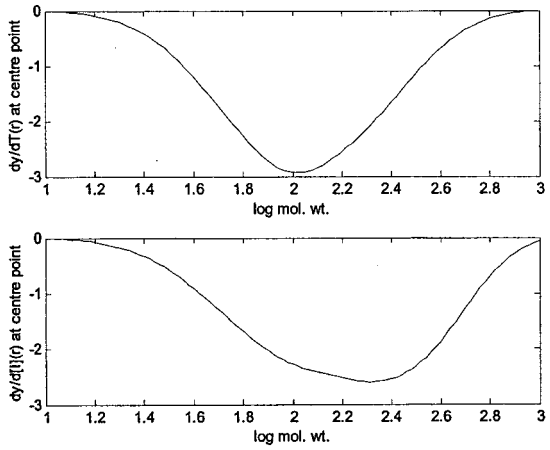

Figure 7: Sensitivity profiles with respect to temperature and initiator concentration evaluated at the centre point of the experimental design

Using the Model to Guide Further Experimentation

One of the major benefits of process modeling is to guide further experimentation. The efficacy of such an approach for the empirical MWD modeling presented in this paper was assessed in a blind study. MWDs were generated independently using Predici® for two test cases, and operating points were computed without prior knowledge of the operating points that had been used to generate the test distributions. Temperature and initiator concentration values to match a given target distribution were obtained via an optimization which minimized the integral squared error between the predicted and target distributions:

$$\min_{T,[I]} \int_{r_0}^{\eta} \left(y_{target}(r) - \hat{y}(r;T,[I])\right)^2 dr \qquad (8)$$

where the dependence of the predicted MWD on T and [I] through the regression model is noted explicitly. The optimization problem was readily solved using the Nelder-Mead Simplex Algorithm presented [5] available in Matlab ™ using the "fminsearch" command. Two blind tests were run, with promising results. The target and predicted distributions are shown below, and the match is quite good.

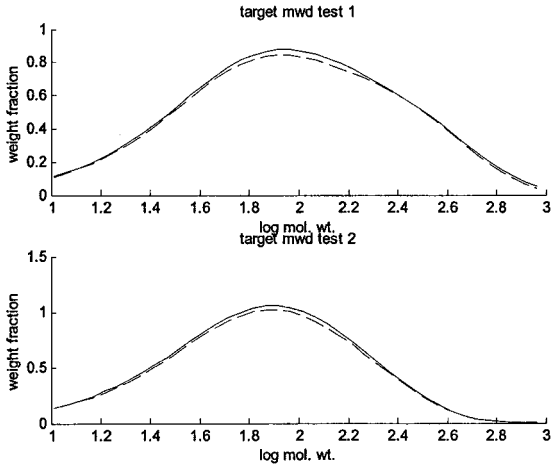

Figure 8: Target and Predicted MWDs for two test points

The operating point estimates in coded and uncoded form are summarized in Table 2, together with the actual values used to generate the test distributions.

Table 2: Estimated and Target Operating Points for Matching Target MWDs

Operating Point	T (coded)	[I] (coded)	T (uncoded, °C)	[I] (uncoded, mol/L)	T (actual, °C)	[I] (actual, mol/L)
Test 1	0.1483	-0.7558	141.5	6.8E-3	143	6.5E-3 mol/L
Test 2	-0.8462	0.7650	131.5	0.018	133	0.017 mol/L

The percentage errors, expressed as percentages of the target values, for the two test points are (1.1%, 5.1%) and (1.1%, 7.3%). The first element of each pair corresponds to temperature, while the second element refers to initiator concentration. From these results, we can conclude that the model is particularly effective at predicting temperature effects, while there is more error associated with predictions of initiator effects. This could be related to the thermal initiation mechanism also included in the model. These results suggest that a higher-order design, enabling the estimation of a nonlinear model in T and [I] in the transformed domain, should be considered.

Conclusion

Functional regression is an effective technique for developing empirical models to predict the impact of operating factors on MWDs. The parameter functions provide insight into the impact of T and [I] on different portions of the MWD, and the empirical models can be used to guide further experimentation and process development. Treating the MWDs as members of an exponential family of distributions and working with transformed responses ensures non-negative predictions. Furthermore, the transformation approach allows the extension of standard techniques for analyzing two-level factorial experimental design results to the MWD case. Main effects plus two-factor interaction models can be appropriately specified in the transformed domain when it would be inappropriate in the original MWD domain. Operating factor effects can be assessed by examining the sensitivities of the predicted MWD to the operating factors at specific operating points, and can be summarized graphically. Further work will address spurious

behaviour at the tails resulting from analytical noise using penalization or truncation, and extensions to higher-order designs (e.g., central composite designs). While styrene polymerization kinetics are reasonably well-understood and mechanistically modeled, the technique can also be applied for the estimation of empirical models relating operating factors to end-use properties, and MWDs to end-use properties.

Acknowledgements

The financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC) in the form of a Discovery Grant is gratefully acknowledged.

[1] Box, G.E.P., W.G. Hunter and J.S. Hunter, Statistics for Experimenters: An Introduction to Design, Data Analysis and Model-Building, Wiley, New York (1978).

[2] Campbell, J.D., F. Teymour, M.Morbidelli, "High Temperature Free Radical Polymerization. 2. Modeling Continuous Styrene Polymerization", Macromolecules **36**, 5502-5515 (2003).

[3] Elizalde, O., M. Vicente, J.R. Leiza and J.M. Asua, "Control of the Adhesive Properties of Acrylate/Styrene Latexes", Polymer Reaction Engineering, **10**(4), 265-283 (2002).

[4] Kotouslas, C., A. Krallis, P. Pladis and C. Kirpassides, "A Comprehensive Kinetic Model for the Combined Chemical and Thermal Polymerization of Styrene up to High Conversions", Macromol. Chem. Phys. **204**, 1305-1314 (2003).

[5] Lagarias, J.C., J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions," SIAM Journal of Optimization, **9**(1), 112-147 (1998).

[6] Ramsay, J.O., Research Web Site, McGill University, www.psych.mcgill.ca/faculty/ramsay.htm (2003).

[7] Ramsay, J.O., Differential equation models for statistical functions. Canadian Journal of Statistics, 28, 225-240 (2000).

[8] Ramsay, J.O. and B.W. Silverman, Functional Data Analysis, Springer-Verlag, New York (1997).