Automating regression analysis of heteroscedastic data in non-linear detectors using an integrated CDS platform

Robert Birdsall, Xiangsha Du, Pawel Bigos, Duanduan Han, Nikhil Bhiwankar

Waters Corporation, Milford, MA

Waters™

Overview

A key challenge in developing quantitative methods with non-linear detectors such as charged aerosol detection (CAD) is defining linearity, working range, and appropriate regression weighting in a way that is both objective and aligned with regulatory expectations. Detector response is inherently non-linear and heteroscedastic making traditional trial-and-error evaluation of calibration curves time-consuming and highly dependent on analyst judgment.

The ideal workflow begins with acquisition of calibration data under a single, well-defined reference condition (PFV = 1.00) directly within the chromatography data system (CDS). As shown conceptually in Figure 1, this reference dataset captures the intrinsic detector response and serves as a foundation for objective, data-driven analysis. Rather than empirically acquiring multiple calibration curves at different detector settings, a reference data set is analytically evaluated to predict the concentration range over which the response can be expected to behave linearly.

Using this approach, standard linear regression models and diagnostic statistics are applied to PFV = 1.00 data to estimate a linear operating range (Figure 2). This step enables objective prediction of where subsequent data acquisition should be focused, significantly accelerating method development by limiting experimentation to conditions most likely to meet linearity and range requirements.

Once data are acquired within the predicted linear range, residual analysis becomes the primary tool for evaluating quantitative performance. As demonstrated in Figure 3, assessment of residual variance within defined confidence limits provides an analytical basis for selecting appropriate regression weighting schemes (e.g., unweighted, 1/x, or 1/x²). This process ensures that weighting decisions are driven by observed heteroscedastic behavior across the calibration range for accurate quantitation and unbiased comparison (figure 4).

Critically, this work serves as a proof-of-concept for transitioning this analytical logic into a compliant-ready CDS environment. By formalizing regression evaluation, residual diagnostics, and weighting selection into a defined process, the workflow demonstrates how objective, data-driven decisions can be executed, documented, and reviewed within the CDS itself (Figure 5). This transition supports traceability, consistency, and regulatory readiness while reducing reliance on external tools and subjective analyst intervention.

Conclusion

This work demonstrates that regression analysis for heteroscedastic, non-linear detector responses can be systematically automated within an integrated CDS platform without sacrificing statistical rigor or defensibility. By formalizing detector-specific response models, weighting strategies, and correlation metrics, the workflow removes analyst subjectivity while improving calibration robustness across diverse detector behavior. Integration of these methods directly into Empower™ CDS Software enables consistent method execution, transparent reporting, and scalable deployment across laboratories. Collectively, this framework advances calibration workflows from manual, experience-driven decisions toward reproducible, data-driven selection increasing productivity and confidence in quantitative results in regulated environments.

Experimental

All lipids in this study were used for research and demonstration purposes and were purchased from the following vendors: cholesterol and DSPC from Sigma-Aldrich; DMG-PEG 2000 and SM-102 from Cayman Chemical. Stocks of each lipid were prepared in methanol at 5 mg/mL and diluted to the appropriate concentration at 90/10 water/methanol (v/v). Samples were separated with UPLC™ System using a Waters GTxResolve™ Lipid Phenyl-Hexyl+ RP Column, MaxPeak™ Premier Technology, SPP, 1.6µm, 230Å, 2.1 x 50 mm at 50°C over a 6-min gradient. Data acquisition and analysis was performed in Empower CDS and Microsoft™ Excel™ Software.

LC system:	ACQUITY™ Premier System (BSM)
Detection:	TUV, λ = 200 / 280 nm FC = 5mm Ti, 1Hz, time constant = normal
Column:	Waters GTxResolve™ Lipid Phenyl-Hexyl+ RP Column, MaxPeak Premier Technology, SPP, 1.6µm, 230Å, 2.1 x 50 mm (p/n186011698)
Column temperature:	50 °C
Sample temperature:	Ambient
Injection volume:	3 µL
Flow rate:	0.400 mL/min
Mobile phase:	A: 0.1% formic acid in water B: 50:50 MeOH:MeCN, 0.1% formic acid
Chromatography software:	Empower 3.9.0 CDS
CAD settings:	
Sampling rate:	10 Hz
Time constant:	Normal
Ion trap:	20 V
Evaporation temperature:	35 °C

Gradient Table:	Time (min)	Flow (mL/min)	% A	% B	Curve
	Initial	0.400	50.0	50.0	Initial
	6.00	0.400	10.0	90.0	6
	8.00	0.400	10.0	90.0	6
	8.50	0.400	50.0	50.0	6
	12.00	0.400	50.0	50.0	6

Results & Discussion

Intrinsic CAD Response Behavior

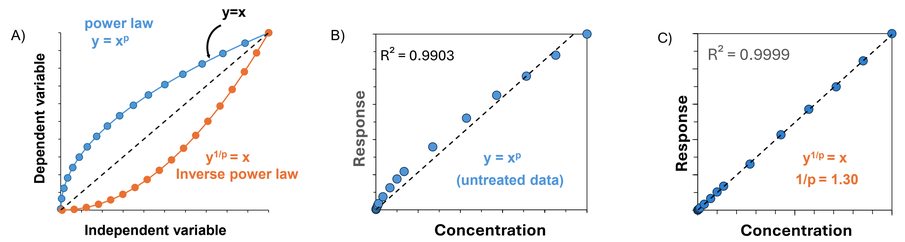


Figure 1. A) Generalized representation of a charged aerosol detector (CAD) non-linear response following a power-law relationship. B) Intrinsic nonlinear CAD response acquired at PFV = 1.00, where detector signal increases as a function of analyte amount raised to an exponent p (untreated). C) An example of linearization of the same calibration curve using an inverse power function transformation 1/p=1.30, illustrating how appropriate selection of PFV mathematically corrects detector nonlinearity and enables definition of an effective linear operating range without altering the underlying analytical data.

Predictive Modeling Using PFV = 1.00 Calibration Data

Predictive Model Definitions

slope

$$m = \frac{\sum w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum w_i (x_i - \bar{x}_w)^2}$$

intercept

$$b = \bar{y}_w - m\bar{x}_w$$

mean weighted x and y

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} \qquad \bar{y}_w = \frac{\sum w_i y_i}{\sum w_i}$$

weighted residual sum of squares

$$WRSS = \sum w_i (y_i - \hat{y}_i)^2$$

weighted total sum of squares

$$WTSS = \sum w_i (y_i - \bar{y}_w)^2$$

coefficient of determination

$$R_w^2 = 1 - \frac{WRSS}{WTSS}$$

standardized residual

$$r_i = \left(\frac{(y_i - \hat{y}_i)\sqrt{w_i}}{RMS\left((y_i - \hat{y}_i)\sqrt{w}\right)} \right) \qquad RMS = \sqrt{\frac{1}{n}\sum w_i (y_i - \hat{y}_i)^2}$$

Monte Carlo simulation slope (m), y-intercept (b), WRSS, and R_w^2 defined above. PFV limits = precentile (e.g. 90%, 95%) of ΔRSS distribution. Absolute noise → 1/x, 1/x².

best fit (min RSS) $\hat{y}_i = mx_i + b, \qquad e_i = y_i - \hat{y}_i$

weighting 1/x or 1/x² $\sigma = \text{std}(e_i), \qquad y_i^* = \hat{y}_i + N(0, \sigma^2)$

confidence interval $\Delta RSS = RSS(p) - RSS_{min}$

Predictive Model Workflow

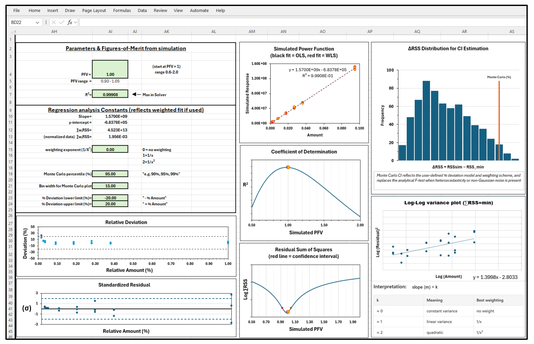
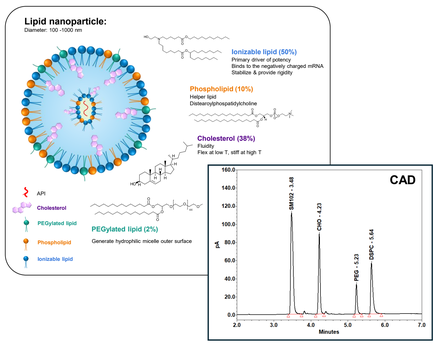


Figure 2. Predictive modeling framework used to characterize intrinsic CAD response and simulate detector behavior at alternative PFV settings using calibration data acquired at PFV = 1.00. A) Statistical definitions shown include weighted least-squares regression parameters, coefficient of determination, residual metrics, and confidence interval criteria. B) The model enables forward simulation of detector response and objective evaluation of PFV-dependent linearity using standard spreadsheet-based tools such as Microsoft Excel Software or CDS-supported informatics.

Please visit our booth (#213) to learn more about the predictive modeling workflow.

Application to a Representative Analytical System



SM102						
Simulator PFV range 1.28-1.32			Deviation (%) weighting			
Experimental PFV 1.30						
conc. (mg/mL)	Area	RSD (%)	none	1/x	1/x²
0.0043	6.31E+06	4.19	-69.84	-12.25	-2.23
0.0065	9.36E+06	3.58	-42.40	-5.62	-0.56
0.0089	1.30E+07	1.68	-24.50	1.18	3.47
0.0230	3.28E+07	0.93	-0.89	6.26	4.17
0.0467	6.45E+07	1.97	3.64	4.92	1.54
0.0704	9.72E+07	0.65	6.47	5.78	1.92
0.0952	1.29E+08	2.22	5.74	4.12	0.12
0.2492	3.08E+08	1.03	-1.39	-4.39	-8.41

Cholesterol					
Simulator PFV range 1.14-1.29			Deviation (%) weighting		
Experimental PFV 1.30					
conc. (mg/mL)	Area	RSD (%)	none	1/x	1/x²
0.0017	2.41E+06	4.33	33.55	11.27	4.97
0.0025	3.47E+06	1.02	14.08	-0.31	-3.70
0.0034	4.84E+06	3.04	8.14	-2.03	-3.74
0.0088	1.27E+07	2.98	-2.67	-5.69	-4.59
0.0178	2.70E+07	2.85	-2.18	-2.81	-0.68
0.0268	4.17E+07	4.03	-0.39	-0.22	2.30
0.0362	5.53E+07	1.76	-2.49	-1.96	0.67
0.0948	1.48E+08	2.69	0.46	1.75	4.77

DMG-PEG 2000					
Simulator PFV range 1.21-1.37			Deviation (%) weighting		
Experimental PFV 1.30					
conc. (mg/mL)	Area	RSD (%)	none	1/x	1/x²
0.0004	1.17E+06	5.43	-43.16	-15.84	-8.40
0.0006	1.62E+06	8.95	-19.74	-2.41	1.18
0.0008	2.29E+06	4.83	4.71	16.54	17.72
0.0021	4.24E+06	1.22	-6.62	-3.04	-4.27
0.0042	8.24E+06	0.95	-1.10	-0.39	-2.69
0.0063	1.27E+07	3.40	5.22	4.91	2.12
0.0086	1.67E+07	0.98	3.20	2.47	-0.41
0.0225	4.08E+07	2.15	-0.77	-2.23	-5.24

DSPC					
Simulator PFV range 1.20-1.29			Deviation (%) weighting		
Experimental PFV 1.30					
conc. (mg/mL)	Area	RSD (%)	none	1/x	1/x²
0.0009	1.81E+06	19.29	14.63	8.32	3.55
0.0013	2.77E+06	2.84	8.32	4.28	1.86
0.0017	3.41E+06	2.63	-4.78	-7.67	-9.08
0.0045	9.70E+06	4.64	-2.55	-3.39	-2.52
0.0091	1.99E+07	4.12	-4.00	-4.19	-2.63
0.0137	3.15E+07	1.81	0.12	0.17	2.06
0.0185	4.36E+07	3.15	2.11	2.28	4.34
0.0485	1.12E+08	0.23	-0.16	0.20	2.41

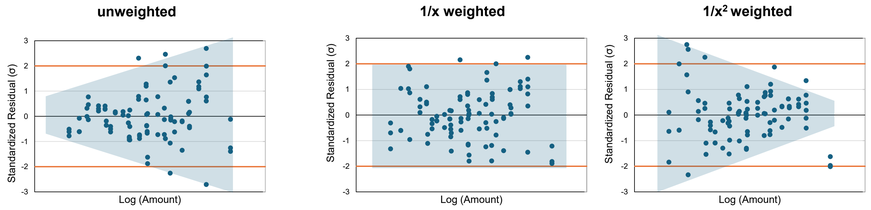


Figure 3. *Determination of appropriate regression weighting for a representative lipid nanoparticle analytical system. Ionizable lipid (SM102), cholesterol, DSPC, and DMG-PEG 2000 were analyzed by CAD across a wide concentration range at the predictive model recommended linear PFV-range. Calibration performance using unweighted, 1/x, and 1/x² regression is compared using standardized residuals versus amount (log-scale). Residual trends reveal heteroscedasticity with unweighted fitting at low concentrations and over-compression with 1/x² weighting, whereas 1/x weighting produces the most uniform residual distribution across the calibration range, supporting its selection as the optimal weighting at PFV = 1.30.*

Comparison of Simulated and Experimental Performance

A)

PFV 1.30 Data

Figure of merit	SM102		Cholesterol		DMG-PEG 2000		DSPC	
Data	Simulated	Experimental	Simulated	Experimental	Simulated	Experimental	Simulated	Experimental
PFV	1.30	1.30	1.30	1.30	1.30	1.30	1.30	1.30
R²	0.9997	0.9968	0.9972	0.9992	0.9980	0.9977	0.9964	0.9988
residual	0.0038	0.0511	0.1440	0.0522	0.3925	0.0477	0.3433	0.1213

B)

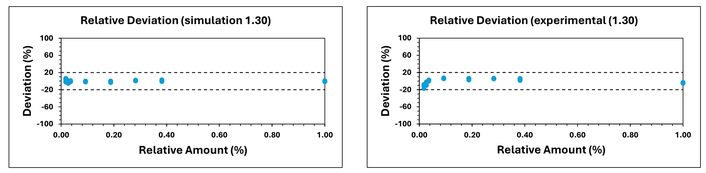


Figure 4. Verifying predictive workflow with experiment. Experimental data acquired at PFV=1.30 was retroactively imported into the model and compared to original simulation data (PFV=1.00) with a simulated PFV setting 1/p=1.30 using 1/x weighting. A) Results exhibited good agreement in correlation data and B) residuals with relative deviations within +/- 20% for all LNP species. Furthermore, the predictive model indicated the experimental results were already near or at optimal linear range (PFV ≤ 0.08), based on "best-fit", demonstrating the validity of the proposed workflow. Results demonstrate that PFV optimization predicted from PFV = 1.00 calibration data accurately reflects experimental detector performance.

Empower CDS Software CAD-based Workflow

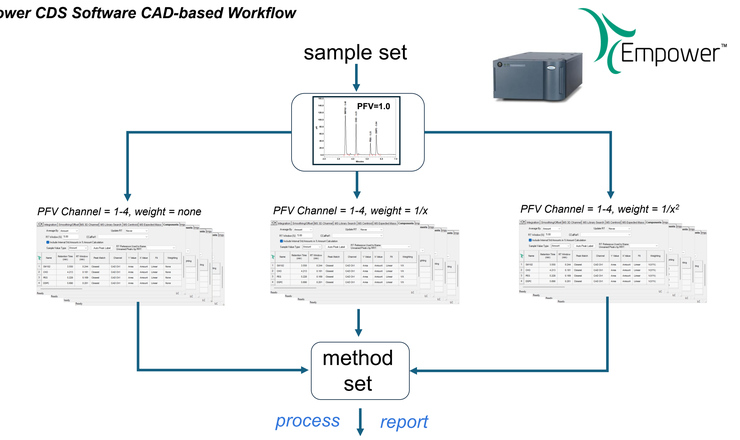


Figure 5. Example workflow illustrating integration of inverse power function modeling within Empower™ CDS Software. Calibration data acquired at PFV = 1.00 are used for predictive evaluation of PFV settings while maintaining data traceability and auditability with regulated laboratory workflows. This information is then used to direct data-driven selection of optimal PFV value and weighting schemes for reporting.

References

1. Fekete S, Doneanu C, Addepalli B, et al. Challenges and emerging trends in liquid chromatography-based analyses of mRNA pharmaceuticals. Journal of Pharmaceutical and Biomedical Analysis. 2023 Feb;224:115174. DOI: 10.1016/j.jpba.2022.115174.
2. DeLaney, K., Han, D., Birdsall R., Yu, Y. Optimized ELSD Workflow for Improved Detection of Lipid Nanoparticle Components. (2022), Waters Application Note 720007740.
3. Han, D., Birdsall, R., Simeone, J., Fogwill, M., Yu, Y. Comparing ELSD and CAD Performance on Polysorbate Quantification in Infliximab Drug Products. (2022), Waters Application Note 720007501.
4. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). *ICH Q2(R2): Validation of Analytical Procedures.* ICH; 2023.