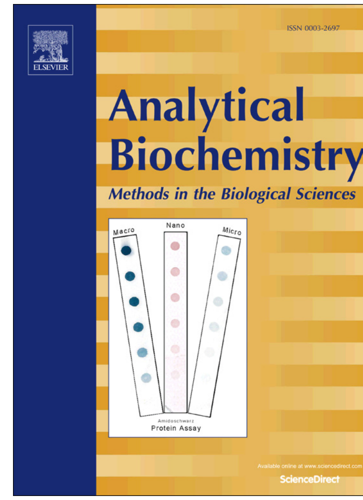
Accepted Manuscript



Notes & tips

Robust regression methods for real-time PCR

Wim Trypsteen, Jan De Neve, Kobus Bosman, Monique Nijhuis, Olivier Thas, Linos Vandekerckhove, Ward De Spiegelaere

PII: S0003-2697(15)00139-6
DOI: http://dx.doi.org/10.1016/j.ab.2015.04.001
Reference: YABIO 12030

To appear in: *Analytical Biochemistry*

Received Date: 6 February 2015
Revised Date: 27 March 2015
Accepted Date: 2 April 2015

Please cite this article as: W. Trypsteen, J. De Neve, K. Bosman, M. Nijhuis, O. Thas, L. Vandekerckhove, W. De Spiegelaere, Robust regression methods for real-time PCR, *Analytical Biochemistry* (2015), doi: http://dx.doi.org/10.1016/j.ab.2015.04.001

Robust regression methods for real-time PCR

Authors: Trypsteen Wim[1], De Neve Jan[2], Bosman Kobus[3], Nijhuis Monique[3], Thas Olivier[2,4], Vandekerckhove Linos[1], De Spiegelaere Ward[1]

Affiliations: [1] Department of Internal Medicine, HIV Translational Research Unit, University Ghent & University Hospital Ghent, Belgium. [2] Department of Mathematical Modelling, Statistics and Bio-informatics, University Ghent, Belgium. [3] Department of Medical Microbiology & Virology, University Medical Center Utrecht, Utrecht, The Netherlands. [4] National Institute for Applied Statistics Research Australia (NIASRA), School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia.

Corresponding author: Vandekerckhove Linos (*); Depintelaan 185, De Pintepark Building, University Hospital Ghent, 9000 Ghent, Belgium; tel: +3293323398, fax: +3293323895; linos.vandekerckhove@ugent.be

Abstract

Current real-time PCR data analysis methods implement linear least squares regression methods for primer efficiency estimation based on standard curve dilution series. This method is sensitive to outliers that distort the outcome and are often ignored or removed by the end-user. Here, robust regression methods are shown to provide a reliable alternative, since they are less affected by outliers and often result in more precise primer efficiency estimators than the linear least squares method.

Keywords: Robust regression, real-time PCR, outliers, qPCR, standard curve, PCR efficiency estimation

Manuscript

The real-time quantitative polymerase chain reaction (qPCR) is a well-established technique for quantification of nucleic acids [1, 2]. Quantification of relative and/or absolute quantities is generally conducted by comparing the quantitative outcome to a calibration curve made by a standard dilution series. In this setup, it is essential to construct a calibration curve that accurately reflects the reaction efficiency of the qPCR to estimate the concentration of the unknown samples [2].

In current qPCR practice, the linear least squares regression is implemented to deduce the PCR reaction efficiency from the calibration curve (for details see Supplementary Material 1) [3]. This method works well with an optimal standard curve preparation and minimal variation between technical replicates. However, in practice, minor variations are regularly observed and these tend to increase in dilutions at the lower end of detection. In addition, the least squares method is sensitive to outliers, especially when these are present at the extremes of the dilution series. In general, variation in standard dilution series is more frequently observed in the lower extreme, when the effect of the random sampling error increases. This issue often results in the manual removal of some replicates by the end-user to intuitively fit a better standard curve, making data analysis subjective to the interpretation of the end-user. In this light, robust regression methods for estimating the slope may be preferred as they are less susceptible to outliers and provide more precise estimators for a variety of error distributions [4]. This concept was recently introduced and explored in the context of real-time PCR by Orenti & Marubini, showing that a robust regression method, the biweight MM estimator, could offer an alternative to the linear least squares method in calibration curve calculations, especially when outliers are present [5].

Here, we extend this concept by comparing the least squares method with three robust regression procedures (the MM estimator [5,6], the robust estimator of Theil and Sen [7,8] and the robust estimator of Siegel [9]), by examining the effect of the error distribution on precision in the absence of outliers and by including the least-squares method after outlier removal according to the Grubbs' outlier test [10]. The Grubbs test is a standardized test for detecting and removing outliers from qPCR standard curves. It tests the null-hypothesis that there are no outliers in the data versus the alternative that there is at least one outlier. The largest Cq value is removed when the p-value of this test is less than 0.05. This process is repeated on the reduced dataset until the p-value exceeds 0.05.

We consider bias caused by outliers and precision as criteria to evaluate and compare the different estimators. More specifically, we examine the bias of the primer efficiency estimator when there is an outlier at the extreme of the dilution series. It is desirable to have a method that is not affected by a single outlier when the other dilution points indicate good primer efficiency.

The precision is inversely proportional to the variance of the estimator. When multiple standard curve dilution series of comparable quality are available for estimating the primer efficiency, then it is desirable to have similar efficiency estimates across all series. This corresponds to an estimator with a small standard error or, equivalently, a high precision.

To illustrate the performance of the different methods for standard curve calculations, a qPCR assay for the quantification of HIV DNA was performed in 8 dilution series. This nested qPCR assay uses two rounds of PCR amplification of the HIV gag-region. DNA from U1 cells, which contain two HIV provirus integrations per cell, were used for the 8 dilution series; for details, see Supplementary Material 2 [11,12]. The regression lines and corresponding efficiencies for each replicate are estimated according to the different regression methods (Supplementary Material 3 & 4). All calculations were performed using the statistical software environment R [13].

A first distinction between linear and robust regression methods can be made by examining a dilution series with and without an outlier (Figure 1). In case no outliers are present (Figure 1A), all four methods provide a comparable and accurate estimate of the efficiency. However, when an outlier is present (Figure 1B), the least squares method results in a decreased primer efficiency estimate, while the estimated efficiencies of the robust methods remain largely unaffected by the outlier.

To examine this in more detail, data simulations were performed to mimic the experimental setups with and without outliers. Seven distinct calibration curves are considered and replicated twice. Quantification cycles are simulated according to a linear model with intercept 26.5 and slope -3.553 corresponding to a primer efficiency of 91.2% (see the Supplementary Material 1 for more details on the linear model and the primer efficiency). To simulate the errors, we consider the normal distribution with mean zero and standard deviation 0.19 which corresponds to the estimated standard deviation of the error distribution of the data in Figure 1A.

Table 1A shows the average estimated efficiency based on 10000 Monte-Carlo simulations. Without the outlier, all estimators are unbiased. However, when a single outlier is present, the least squares estimator is biased and, on average, underestimates the efficiency by 21%. On the other hand, the MM estimator remains unbiased while the Theil-Sen and Siegel estimators underestimate the efficiency by approximately 1%.

A second distinction can be made by examining the variance of the efficiency estimators. For the real data, the robust regressions have lower standard deviations for the estimated efficiency compared to the linear least squares regressions (Supplementary Material 4). These standard deviations may suggest that the least squares estimators are less precise as compared to the robust estimators so that robust regression will likely estimate the true efficiency more accurately. This is of great importance while newly designed qPCR primer pairs are often only tested once or twice on a standard dilution series of reference material.

To examine whether the presence of outliers are solely responsible for this observation, data simulations were performed as described above, but only to mimic the experimental setup without outliers so that all four estimators are unbiased. In addition, the performances of the estimators are tested over different error distributions, as the underlying distribution of the error is unknown. The following error distributions were included: normal, student t with 3 degrees of freedom, lognormal, Gumbel and Laplace. They were all standardized to mean zero and standard deviation 0.19 to make the results comparable (Table 1B). These choices include symmetric, heavy tailed, and skewed distributions.

For the five error distributions considered, the MM and Theil-Sen estimator outperform the least squares estimators except for the normal distribution for which the least squares estimator has a slightly better performance. Hence, even without outliers, robust regression methods can produce more precise estimates.

For the normal distribution, the least squares estimator is most precise. This can be expected since the least squares estimator corresponds to the maximum likelihood estimator for normal distributed errors. In comparison with the least squares estimator, the MM, Theil-Sen and Siegel estimators result in an increased standard error of 5%, 5% and 17% respectively. For the other error distributions, however, the MM and Theil-Sen estimators have an increased precision over the least squares estimator. For the lognormal distribution, for example, the standard error of the least squares estimator is more than twice the standard error of the MM and Theil-Sen estimators. The Siegel estimator has a superior precision over the least squares for the t-distribution, lognormal and Laplace distribution, but underperforms as compared to the MM and the Theil-Sen estimators.

In summary, we have demonstrated that robust regression estimators are less affected by outliers and often prove to be more precise for estimating PCR efficiency compared to standard linear regression with least squares estimation. Especially the MM and Theil-Sen estimators seem to be appropriate for primer efficiency estimation. Therefore, the implementation of robust regression methods in qPCR analysis would provide a reliable alternative. Single outlying dilutions can effectively introduce bias to efficiency estimates, particularly when these occur in the extremes of the dilution series where stochastic sampling effects at the lower levels of detection may increase the variation. However, we do want to stress that outlying data points can be an important indication of poor pipetting and low quality data of the entire PCR run. Therefore, robust regression methods should not be used to falsely expand the dynamic range of the linear interval of the PCR assay. The

Cq values at which outlying data points are observed, should still be considered outside the dynamic range of the assay in the given PCR run.

Acknowledgements

Financial support was provided by the King Baudouin Foundation (Grant 2010-R20640-003) and the HIV-ERA Eranet project EURECA: SBO-IWT (Grant 130442). Prof. Linos Vandekerckhove is supported by the Research Foundation – Flanders (FWO; Grant 1.8.020.09.N.00). Kobus Bosman is supported by the AIDS FONDS (project 2013034), Monique Nijhuis is supported by the amfAR Research Consortium on HIV Eradication (ARCHE) (EPISTEM, project (108930) and the Netherlands Organization for Scientific Research VIDI (grant number 91796349). Jan De Neve and Olivier Thas gratefully acknowledge the IAP research network grant no. P7/06 of the Belgian government (Belgian Science Policy). The research fits within the N2N Multidisciplinary Research Partnership of Ghent University (01MR0310W).

References

[1] C. A. Heid, J. Stevens, K. J. Livak, P. M. Williams, Real time quantitative pcr, Genome Res 6 (1996) 986-994.

[2] S. A. Bustin, V. Benes, J. A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M. W. Pfaffl, G. L. Shipley, J. Vandesompele, C. T. Wittwer, The miqe guidelines: Minimum information for publication of quantitative real-time pcr experiments, Clin Chem 55 (2009) 611-622.

[3] J. M. Ruijter, M. W. Pfaffl, S. Zhao, A. N. Spiess, G. Boggy, J. Blom, R. G. Rutledge, D. Sisti, A. Lievens, K. De Preter, S. Derveaux, J. Hellemans, J. Vandesompele, Evaluation of qpcr curve analysis methods for reliable biomarker discovery: Bias, resolution, precision, and implications, Methods 59 (2013) 32-46.

[4] P.J. Rousseeuw, A.M. Leroy, Robust regression and outlier detection, vol 589, John Wiley and Sons, 205.

[5] A. Orenti, E. Marubini, Performance of robust regression methods in real-time polymerase chain reaction calibration, The International journal of biological markers 29 (2014) e317-327.

[6] V.J. Yohai, High Breakdown-point and High Efficiency Estimates for Regression, The Annals of Statistics, 15 (1987), 642-65.

[7] H. Theil, , A Rank-Invariant method of linear and polynomial regression analysis, I, II, and III, Nederlandsche Akad. van Wetenschappen Proc. 53 (1950) 386-92, 521-5, 1397-412.

[8] P.K. Sen, Estimates of the Regression Coefficient Based on Kendall's Tau, Journal of the American Statistical Association 63 (1968) 1379-1389.

[9] A.F. Siegel, Robust Regression Using Repeated Medians, Biometrika (1982) 242-244.

[10] M.J. Burns, G.J. Nixon, C.A. Foy, N. Harris, Standardisation of data from real-time quantitative PCR methods – evaluation of outliers and comparison of calibration curves, BMC Biotechnology 7 (2005) 5-31.

[11] W. Hu, R. Kaminski, F. Yang, Y. Zhang, L. Cosentino, F. Li, B. Luo, D. Alvarez-Carbonell, Y. Garcia-Mesa, J. Karn, X. Mo, K. Khalili, RNA-directed gene editing specifically eradicates latent and prevents new HIV-1 infection, Proceedings of the National Acadamy of Sciences of the United States of America 111 (2014) 11461-11466.

[12] T.M. Folks, J. Justement, A. Kinter, S. Schnittman, L. Orenstein, G. Poli, A.S. Fauci, Characterization of a promonocyte clone chronically infected with HIV and inducible by 13-phorbol-12-myristate acetate, Journal of Immunology 140 (1988) 1117-1122.

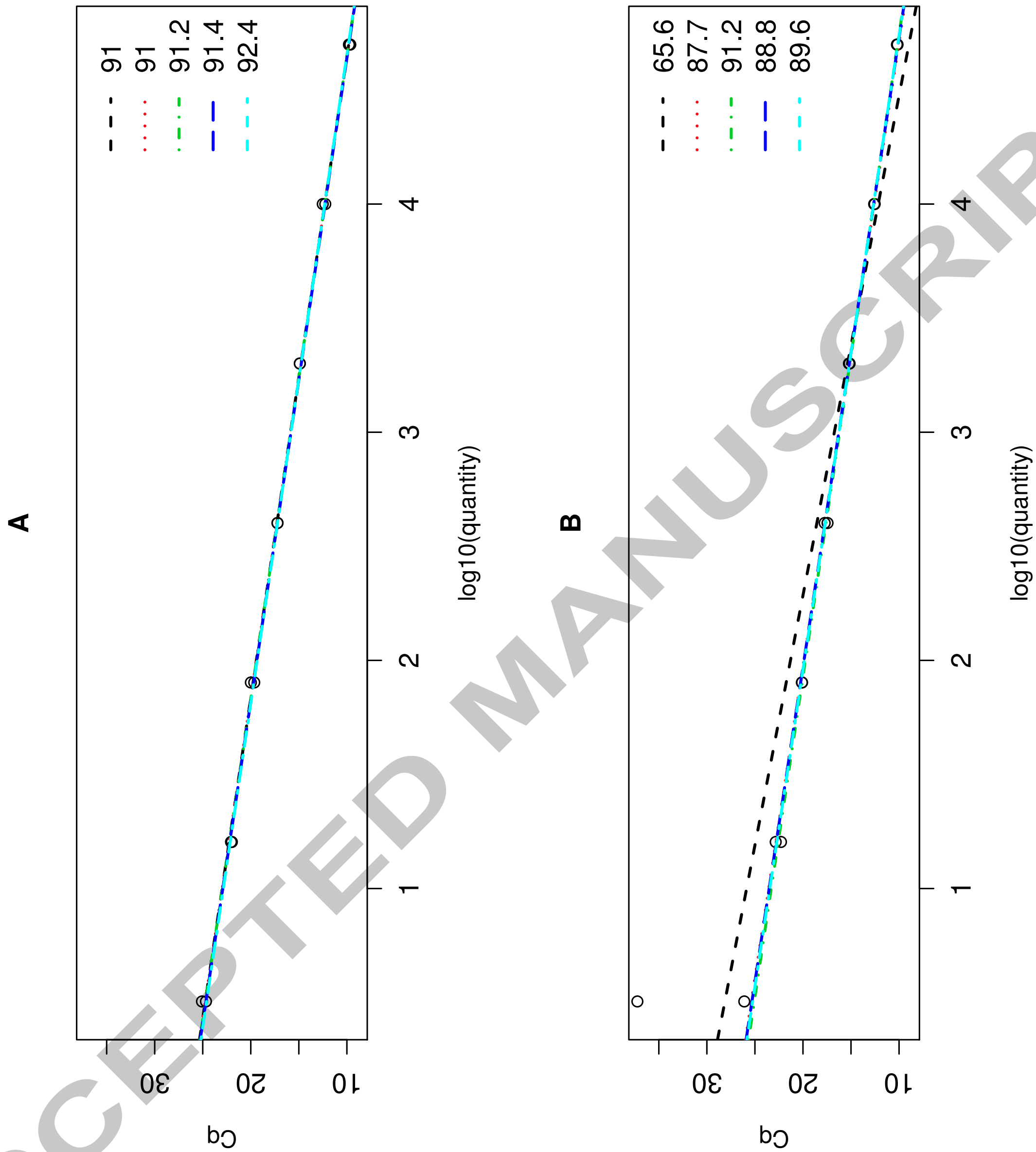
[13] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, (2014) URL http://www.R-project.org/.

Figure 1: Estimated regression lines and estimated efficiencies for two of the eight replicates. Comparison of the five regression estimates without an outlying dilution (A) and in case of an outlying dilution at the lower extreme of the standard curve (B). The least squares (black line, --), the least squares after removal of outliers with the Grubbs test (red line, ···), robust MM (green line, · - ·), the robust Theil-Sen (blue line, —-) and the robust Siegel (turquoise line, — - —) methods show equal efficiency in the upper panel (A). The least squares estimate is heavily affected by the outlier in the lower panel (B), while the estimated efficiencies of the robust methods and the least squares after performing Grubbs' test for removing the outlier remain comparable across the replicated experiments.

Table 1: Results of the simulation study in which the true efficiency equals 91.2%. A: averages of the estimated efficiencies (in %) for the five estimators in the presence and absence of an outlier. B: standard deviations of the estimated efficiencies for the five estimators applied to the simulated data without outlier (for which all estimators are unbiased). Smaller values indicate more precise estimators. All results are obtained based on 10000 Monte-Carlo simulations.

A	LS	LS + Grubbs	MM	Theil-Sen	Siegel
Without outlier	91.2	91.2	91.2	91.2	91.2
With outlier	70.0	91.2	91.2	90.1	90.5

B	LS	LS + Grubbs	MM	Theil-Sen	Siegel
Normal	1.27	1.27	1.33	1.34	1.48
Student t	1.26	1.28	0.99	1.00	1.05
Lognormal	1.27	1.27	0.64	0.62	0.68
Gumbel	1.27	1.27	1.22	1.21	1.34
Laplace	1.27	1.27	1.17	1.16	1.20


A
B
91
91
91.2
91.4
92.4
65.6
87.7
91.2
88.8
89.6
Cq
10
20
30
log10(quantity)
1
2
3
4