The choice of prediction curve method and its effect on the estimated amount of DNA

by

Morgan Sophia Xiao Magee
B.Sc., University of Victoria, 2022

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

The choice of prediction curve method and its effect on the estimated amount of DNA

by

Morgan Sophia Xiao Magee

B.Sc., University of Victoria, 2022

Supervisory Committee

Dr. Mary Lesperance, Supervisor

Department of Mathematics and Statistics

Dr. Caren Helbing, Outside Member

Department of Biochemistry and Microbiology

Abstract

Water samples from the field data that contained environmental DNA (eDNA) were taken from multiple rivers where oolichan fish (*Thaleichthys pacificus*) are known to spawn. The samples were split into 8 technical replicates and analyzed using quantitative real-time polymerase chain reaction (qPCR). A qPCR experiment is the real time quantification of DNA amount at the end of a full cycle of heating and cooling. C_T values were determined from a qPCR experiment or a replicate was given N/A (not available) if no DNA was detected. Four data sets from two different labs were used, Bureau Vertitas Lab (BVL) and University of Victoria (UVic). Both labs have a gblock and field data set with chemical assays named eTHPA2 and eTHPA6. Gblock data is comprised of gblock samples which were synthetically constructed genes of known concentration (copy number) and measured using qPCR. Field data is comprised of samples taken from river sites in British Columbia where eDNA naturally occurs, and the copy numbers were unknown for the field samples. The field samples were analyzed using qPCR technique to determine the C_T value for each technical replicate. Each data set was split into two subsets named full and partial detect, resulting in eight working data sets. The full data sets were comprised of samples whose technical replicates had (8/8) detects. The partial data set was comprised of samples whose technical replicates had less than (8/8) detects. For the partial detect data, a Binomial model for the proportion of detects in a sample was defined, where a replicate with a C_T value was an “event” and N/A was not an “event”. Assuming the number of molecules in a sample followed a Poisson distribution with mean λ, we estimated the λ as $\hat{\lambda} = -ln(1-\hat{p})$, where $\hat{p}$ is the estimated sample proportion of detect from the Binomial model.

Standard/calibration and prediction curves were built from the gblock data. Standard curves were built using gblock data with known copy number values, and relate C_T and $\hat{\lambda}$ values to copy number values. Standard curves were used to estimate copy numbers given C_T or $\hat{\lambda}$ values for samples with unknown copy numbers. Prediction curves were built by fitting least squares and orthogonal regression using an unweighted and weighted method for each, to the gblock data. Prediction curves were used to estimate eTHPA6 C_T or $\hat{\lambda}$ values given eTHPA2 C_T or $\hat{\lambda}$ values. Plots and model summaries for the four prediction curves for each data set were analyzed. Based off the analysis and recommendation of the literature, weighted orthogonal regression was chosen as the best prediction model for each gblock data set. The prediction curves were applied to the corresponding field data to investigate how well the models predict the values of eTHPA6 given eTHPA2. All of the data sets saw majority well predicted final eTHPA6 copy number values, which indicated that the weighted Deming model was a good prediction method.

The purpose of this study was to determine the best statistical methods for eDNA assay prediction for biologists and other researchers to use. From the methods validated in this study researchers can go on to connect the population estimates of the oolichan species made from the older and newer assays, make conclusions on the health of the species population, produce plan(s) to safeguard the population against over harvesting, and more conservation work.

Table of Contents

List of Figures

List of Tables

Abbreviations

Starting Quantities (SQ) = known concentration (copy number);

cn = copy numbers;

Ct or C_T = cycle threshold values;

CI or CIs = confidence interval(s);

se = standard error;

f = factors;

eTHPA2 ct = eTHPA2 cycle threshold values;

eTHPA6 ct = eTHPA6 cycle threshold values;

n - set number of technical replicates;

detect - number of technical replicates per sample with a detect;

phat - proportion of detects per sample, $\hat{p}$;

vphat - variance of phat;

lamhat - estimated poisson mean, $\hat{\lambda}$;

vlamhat - variance of $\hat{\lambda}$;

sd - standard deviation of variable;

ln(cn) - natural log of the copy numbers;

eDNA - Environmental DNA;

qPCR - quantitative real-time polymerase chain reaction;

LS - Least squares regression;

OR - Orthogonal/Deming regression;

WOR - weighted orthogonal/Deming regression;

BVL - Bureau Veritas Lab;

UVic - University of Victoria Lab;

Dedication

I would like to thank Dr. Mary Lesperance for her unwavering support, guidance, and encouragement throughout my time working with her. Dr. Lesperance has impacted my life in a profound way; she has shown myself and many others the joy of statistics through her enthusiastic, humorous, and insightful lectures. Words cannot describe how fortunate I feel to have been granted the opportunity to study under her remarkable mentorship. Dr. Lesperance is a role model whose impact on my life I will never forget.

I would also like to thank the dedicated team of researchers, whom I had the honor of working with during my time in the UVic masters program.

On the personal side, I thank my parents and partner for their endless support.

1 Introduction

The utilization of environmental DNA (eDNA) by biologists, environmental scientists, and other academic researching groups has grown greatly over the past few years. Environmental DNA is defined as genetic material that is isolated from water, soil, air, or sediment samples (Lesperance et al. 2021). This growth in popularity can be attributed to its low cost due to the reduction in need for trained biologists for collection, and its non-invasive procedure (Lesperance et al. 2021). With the growth in eDNA usage, new techniques are being developed to detect organism DNA and to estimate the amount of DNA present in a sample. As the technology has advanced, the species-specific detection assays have changed to ones that show complete species specificity (Allison et al. 2023). With the change in preferred detection assay, past samples need to be transformed to match the scale of the new samples for comparison purposes. In order to transform these older samples, prediction curves need to be built and studied. In our case we want to predict eTHPA6 from eTHPA2, as eTHPA6 is now the current detection assay for oolichan fish (Helbing 2023).

There are many options to build a prediction curve, however one must take into consideration how the data was collected and what errors may accompany them. When predicting one measurement from another from within the same sample, in most cases, a curve is built that relates the two measurements. This curve can be linear or nonlinear, and usually the prediction curve relies on one measurement acting as the predictor and the other as the response. To estimate the curve, least squares (LS) is often employed. However, LS assumes that only one of the variables has an associated error, namely the response (Souza et al. 2018). In our case, both variables are measured with an associated error which suggests the use of orthogonal regression (OR) (Souza et al. 2018). Orthogonal regression is often used by researchers to build a prediction curve between two measurements (Taberner and Dufty 1995). Noblitt et al. (2016) indicated that weighted regressions often outperform non-weighted regressions when estimating parameters for prediction curves; we will also use a weighted orthogonal regression (WOR) model. This paper will examine the differences in the estimates of LS and OR, and examine the differing mathematical ways in which OR estimates can be calculated with or without weighting the variables.

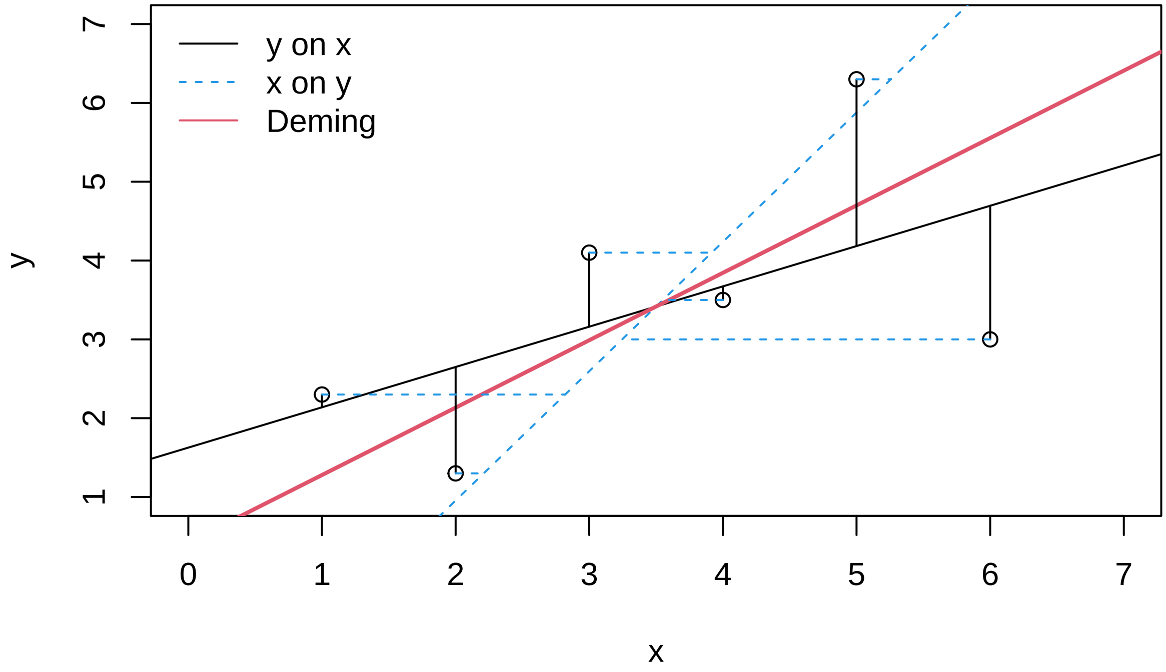


Figure 1: An example by Therneau (2018) of the difference between the symmetric OR line and the asymmetric LS lines.

The orthogonal regression algorithm builds a symmetric line between variables, which means that either order of the variables theoretically should yield the same line (Therneau 2018). This is because OR minimizes the distance perpendicular to the line whereas LS minimizes the vertical distance (Therneau 2018). Figure 1 demonstrates how the OR line differs from the LS lines. The y on x line (black) and the x on y line (blue) are not symmetric when interchanging the variables. This is because the the y on x line (black) minimizes the vertical distances whereas, the x on y line (blue) minimizes the horizontal distances. The OR line (Deming, red) minimizes the perpendicular distances, which creates a more balanced line compared to the LS estimates.

Orthogonal regression is often used to determine if two instruments of measurement provide comparable measurements. This is done by minimizing the perpendicular distance from the line to the observed values (Kane and Mroch 2020; Markovsky and Van Huffel 2007). Performing OR can be computed directly, through Principal Component Analysis (PCA) or R packages. In our study, we use the Deming package's algorithm for OR. We use OR to build a prediction line that allows us to estimate eTHPA6 quantities from eTHPA2 measurements. Prediction models are also built on the copy number scale, although it is not recommended due to introduction of new sources of error; the results are in section 8.4 of the Appendix. The eTHPA2 measurements were used in older studies to determine the abundance of oolichan. With newer technology, biology labs have switched to eTHPA6 detection assay to measure the amount of eDNA in a given sample. The estimation of eTHPA6 from eTHPA2 allows researchers to compare the data from older studies with new data which produces a wider timeline for studying population trends in fish, breeding cycles, and more (Hocking et al. 2022).

2 Methods

This section details the methods used in the lab to process the samples into usable data, how the standard curves are built, the mathematics used to perform orthogonal regression (unweighted and weighted), how we predict copy number values, and how we derive confidence intervals and standard error estimates.

2.1 Data lab-processing

2.1.1 Gblock data

Standard or calibration curves are built to relate the cycle threshold (C_T or Ct) values from quantitative real-time polymerase chain reaction (qPCR) experiments to concentration (copy number) for known concentration samples. The cycle threshold value is defined as the number of cycles for the qPCR test to indicate detection of DNA in a sample. In order to relate cycle threshold value to the more informative estimated amount of DNA per sample (concentration or copy numbers), a standard curve is applied to the C_T values which produces estimated copy number values. Hocking et al. (2022) and Lesperance et al. (2021) provided methodology to build standard curves from serial dilution experiment(s) of known copy numbers, to estimate copy number values from C_T values.

A combined overview of the serial dilution methods from Hocking et al. (2022) and Lesperance et al. (2021) is as follows. Let the dilution series contain J samples (S_j) each of known concentration, c_j for $j = 1$ to J many samples taken. Each sample is split into a pre-determined number of technical replicates or reactions ($r_j, ;\ j = 1, ..., J$), which are run on qPCR plates to detect the presence of DNA for the organism of interest. Each reaction is subject to amplification cycles, for a maximum of 50 cycles. If DNA is detected during the reaction, then the cycle threshold value (C_T) at the time of detection is recorded. This is repeated for each individual reaction. If the cycles surpasses the maximum without reaching the threshold, the given reaction is valued at N/A (not available) for their C_T number. Thus, the outcome for the experiment is a collection of $C_{Tij},\ i = 1, ...r_j$ corresponding to the number of amplification cycles required to detect gene fragments or N/A if there are no detects. The result for each reaction is also recorded as “detect” or “non detect” corresponding to whether the given replicate has a C_T value or N/A, respectively. This determines the method that is employed to estimate the copy number (amount of DNA) and standard error (se) (Lesperance et al. 2021).

In our experiment, the number of technical replicates per sample (r_j) is typically 8 or 24. Our data is partitioned into two sets, a full detect set and partial detect set. A sample from the serial dilution experiment is classified as full detect if it has r_j out of r_j detects during the serial dilution experiment in both the eTHPA2 and eTHPA6 assays. Samples with less than r_j/r_j detects in both assays are classified as partial detect. To estimate the copy numbers per reaction for the full detect data’s cycle threshold values, a standard curve is used to relate the C_T value to the estimated copy number. The standard curves are built from a serial dilution experiment of known synthetic or plasmid DNA and estimated using a linear model, with BVL and UVIC labs producing and using unique curves (Hocking et al. 2022; Lesperance et al. 2021).

Standard curves for the partial detect data are built using the eLowQuant algorithm by Lesperance et al. (2021), which relates the sample proportion of detects using a Binomial-Poisson model. It is assumed that the number of DNA molecules in a qPCR reaction follows a Poisson distribution with mean λ molecules per reaction. The probability of zero DNA molecules in a reaction, under the Poisson model is $\exp(-\lambda)$. Thus, let $p = 1 - e^{-\lambda}$ be the probability that a reaction contains at least one DNA molecule. We estimate p using $\hat{p}$, the sample proportion of reactions that detect at least one molecule, that is the sample proportion of detects. Then for each sample the Poisson mean is λ, which is estimated by $\hat{\lambda} = -\ln(1 - \hat{p})$. Returning to the dilution series, for each sample S_j, we estimate λ_j as $\hat{\lambda}_j = -ln(1 - \hat{p})$. Lesperance et al. (2021) demonstrated that the linear model $\hat{\lambda}_j = \beta * c_j$ provides a good fit for many lower copy number data sets. The standard curves from eLowQuant are built using no intercept linear models that relate copy number values with $\hat{p}$ values using Maximum Likelihood, assuming a Binomial model for the number of replicates with detections. A no intercept model is used for the partial detect data as it has starting quantity $= 0$ with $C_T = 0$.

2.1.2 Field data

Our data was collected from multiple sites along the west coast of British Columbia where oolichan are known to spawn. The field data is processed at the Bureau Vertitas Lab (BVL) and University of Victoria (UVIC) lab using similar methodology and data classification as the gblock data, but for samples with unknown copy numbers. Field data samples are processed using both eTHPA2 and eTHPA6 detection assays. Therefore, each replicate has two associated C_T values, one from each assay. Field data samples where each assay had full detects r_j/r_j in one but not the other, were classified on a case-by-case basis. Some data pairs with one assay having full detects and the other having (5/8) or more had full detect data analysis applied, after the non-detect technical replicates were removed. Other samples that have one assay with full detects (8/8) and one with less than (4/8) detects are removed from analysis completely as the uneven detection ratio produced unusual final estimates. Technical replicates of samples are removed if the cycle threshold values exceed the range of C_T values for which the standard curve is built.

2.2 Standard curves

Standard curves are employed by many Biochemist and are a method in which a curve is built to relate the outcome of a qPCR experiment to the copy number of a sample or technical replicate (Lavagnini and Magno 2007). The summaries of the fitted least squares models for the standard curves from the full detect gblock data sets and the maximum likelihood models for the partial detect gblock data sets are presented in this section. For the full detect data the standard curve estimates the natural log copy number ($\ln(cn)$) values from C_T values for a given assay. The summaries for the partial detect data sets are from the eLowQuant algorithm by Lesperance et al. (2021).

2.2.1 UVic full detect data: standard curves

The standard curve summaries for eTHPA2 and eTHPA6 are derived from the UVic full detect gblock data.

```
## [1] "eTHPA2 standard curve summary"

##
## Call:
## lm(formula = tab2.ct$Mean ~ log(SQ))
##
## Residuals:
##       1       2       3       4       5       6
##  0.2133 -0.0180 -0.1876 -0.1692 -0.0934  0.2549
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.9618     0.2397     179  5.8e-09 ***
## log(SQ)      -1.6540     0.0318     -52  8.2e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.214 on 4 degrees of freedom
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.998
## F-statistic: 2.71e+03 on 1 and 4 DF,  p-value: 8.18e-07
```

```
## [1] "eTHPA6 standard curve summary"

##
## Call:
## lm(formula = tab6.ct$Mean ~ log(SQ))
##
## Residuals:
##       1       2       3       4       5       6
##  0.1326 -0.0618 -0.0754 -0.1428  0.0961  0.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.0969     0.1364   279.3  9.9e-10 ***
## log(SQ)      -1.5589     0.0181   -86.2  1.1e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.122 on 4 degrees of freedom
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.999
## F-statistic: 7.42e+03 on 1 and 4 DF,  p-value: 1.09e-07
```

The standard curves for eTHPA2 and eTHPA6 from the UVic full detect gblock data are:
eTHPA2: $C_{T_2} = -1.654 * ln(c\hat{n}_2) + 42.962$ with inverse $ln(c\hat{n}_2) = \frac{C_{T_2}-42.962}{-1.654}$, where the standard errors for the slope and intercept are 0.0318 and 0.2397, respectively.

eTHPA6: $C_{T_6} = -1.559 * ln(c\hat{n}_6) + 38.097$ with inverse $ln(c\hat{n}_6) = \frac{C_{T_6}-38.097}{-1.559}$ where the standard errors for the slope and intercept are 0.0181 and 0.1364, respectively.

2.2.2 UVic partial detect data: standard curves

The standard curve summaries for eTHPA2 and eTHPA6 are derived from the UVic partial detect gblock data.

```
## [1] "eTHPA2 standard curve summary"

##
##  eTHPA2

## Convergence = 0

##      Estimate Std.Err Z value  Pr(>z)
## beta    0.566   0.115    4.94 7.7e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## LLR test stat = 1.743 , df =  4 , p-value = 0.7828

## [1] "eTHPA6 standard curve summary"

##
##  eTHPA6

## Convergence = 0

##      Estimate Std.Err Z value  Pr(>z)
## beta    0.652   0.129    5.06 4.2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## LLR test stat = 1.812 , df =  4 , p-value = 0.7704
```

The standard curves for eTHPA2 and eTHPA6 from the partial detect UVic gblock data are:
eTHPA2: $-ln(1-\hat{p}_2) = \hat{\lambda}_2 = .566 * c\hat{n}_2$ with inverse $c\hat{n}_2 = \dfrac{\hat{\lambda}_2}{0.566}$, where $\hat{p_2}$ is the proportion of detects and the standard error for $\hat{\beta} = 0.566$ is 0.115 ,
eTHPA6: $-ln(1-\hat{p}_6) = \hat{\lambda}_6 = .652 * c\hat{n}_2$ with inverse $c\hat{n}_6 = \dfrac{\hat{\lambda}_6}{0.652}$, where $\hat{p_6}$ is the proportion of detects and the standard error for $\hat{\beta} = 0.652$ is 0.129.

2.2.3 BVL full detect data: standard curves

The standard curve summaries for eTHPA2 and eTHPA6 are derived from the BVL full detect gblock data.

```
## [1] "eTHPA2 standard curve summary"

##
## Call:
## lm(formula = tab2.ctbvl$Mean ~ log(SQb))
##
## Residuals:
##       1       2       3       4       5       6       7
##  0.1477 -0.1865  0.0315  0.0305 -0.0766 -0.0181  0.0715
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.9815     0.0971     432  1.3e-12 ***
## log(SQb)     -1.6105     0.0139    -116  9.0e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.118 on 5 degrees of freedom
## Multiple R-squared:     1,   Adjusted R-squared:     1
## F-statistic: 1.35e+04 on 1 and 5 DF,  p-value: 9e-10

## [1] "eTHPA6 standard curve summary"

##
## Call:
## lm(formula = tab6.ctbvl$Mean ~ log(SQb))
##
## Residuals:
##       1       2       3       4       5       6       7
##  0.0729 -0.0997  0.0294 -0.0338 -0.0152  0.0754 -0.0290
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.92955    0.05695     719  9.9e-14 ***
## log(SQb)    -1.63035    0.00814    -200  5.9e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0693 on 5 degrees of freedom
## Multiple R-squared:     1,   Adjusted R-squared:     1
## F-statistic: 4.01e+04 on 1 and 5 DF,  p-value: 5.88e-11
```

The standard curves for eTHPA2 and eTHPA6 from the full detect BVL gblock data are:
eTHPA2: $C_{T_2} = -1.6105 * ln(c\hat{n}_2) + 41.9815$ with inverse $ln(c\hat{n}_2) = \frac{C_{T_2} - 41.9815}{-1.6105}$, where the standard errors for the slope and intercept are 0.0139 and 0.0971, respectively.
eTHPA6: $C_{T_6} = -1.6304 * ln(c\hat{n}_6) + 40.9296$ with inverse $ln(c\hat{n}_6) = \frac{C_{T_6} - 40.9296}{-1.6304}$, where the standard errors for the slope and intercept are 0.0081 and 0.0570, respectively.

2.2.4 BVL partial detect data: standard curves

The standard curve summaries for eTHPA2 and eTHPA6 are derived from the BVL Partial detect gblock data.

```
## [1] "eTHPA2 standard curve summary"

##
##  eTHPA2

## Convergence = 0

##      Estimate Std.Err Z value  Pr(>z)
## beta    1.358   0.308    4.41 1.1e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## LLR test stat = 2.446 , df =  3 , p-value = 0.485

## [1] "eTHPA6 standard curve summary"

##
##  eTHPA6

## Convergence = 0

##      Estimate Std.Err Z value  Pr(>z)
## beta    1.691   0.366    4.62 3.9e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## LLR test stat = 1.104 , df =  3 , p-value = 0.776
```

The standard curves for eTHPA2 and eTHPA6 from the partial detect BVL gblock data are: **eTHPA2:** $-ln(1-\hat{p_2}) = \hat{\lambda_2} = 1.358 * \hat{cn_2}$ with inverse $\hat{cn_2} = \frac{\hat{\lambda}}{1.358}$, where $\hat{p_2}$ is the proportion of detects and the standard error for $\hat{\beta} = 1.358$ is 0.308, **eTHPA6:** $-ln(1-\hat{p_6}) = \hat{\lambda_6} = 1.691 * \hat{cn_2}$ with inverse $\hat{cn_6} = \frac{\hat{\lambda}}{1.691}$, where $\hat{p_6}$ is the proportion of detects and the standard error for $\hat{\beta} = 1.691$ is 0.366.

Cycle threshold values estimation under ideal contidions

The general form of a full detect data standard curve for predicting the natural log copy numbers from C_T values is: $ln(\hat{cn}) = \frac{C_T - \hat{b}}{\hat{m}}$, where $\hat{b}$ and $\hat{m}$ are the estimated intercept and slope, respectively. Under ideal conditions, the standard curves for eTHPA2 and eTHPA6 should yield the same natural log copy number values given C_T values derived from the same sample. Equating the standard curve general forms for eTHPA2 to eTHPA6 gives

$$\frac{C_{T_6} - \hat{b_6}}{\hat{m_6}} = \frac{C_{T_2} - \hat{b_2}}{\hat{m_2}}. \tag{1}$$

Taking the natural logarithm of both sides and isolating for C_{T_6} produces the following relationship between the assays

$$C_{T_6} = \frac{\hat{m_6}}{\hat{m_2}} C_{T_2} - \frac{\hat{m_6}}{\hat{m_2}} \hat{b_2} + \hat{b_6}, \tag{2}$$

where $-\frac{\hat{m_6}}{\hat{m_2}}\hat{b_2} + \hat{b_6}$ is the intercept and $\frac{\hat{m_6}}{\hat{m_2}}$ is the slope. Substituting the $\hat{m}$ and $\hat{b}$ values from the eTHPA2 and eTHPA6 standard curves produces Table 1, containing the estimates for the slope and intercept under the assumption of the ideal relationship between eTHPA6 and eTHPA2. Comparing the estimated slopes of the two labs, we observe that the UVic slope is less than one and the BVL slope is greater than one (Table 1). The estimated intercepts for the two labs differ greatly from each other by almost one unit. In section 4, we compare the intercept and slope to that of the chosen prediction model.

Table 1: The slopes and intercepts for the ideal relationship between eTHPA6 and eTHPA2. The slopes and intercepts are calculated using equation (2) for both UVic and BVL labs.

	Slope	Intercept
UVic	0.9426	-2.397
BVL	1.0124	-1.571

2.3 Orthogonal regression

Orthogonal regression is a method recommended by the literature when fitting a linear model to data, where both variables have error. Since both eTHPA2 and eTHPA6 have associated error, we use unweighted and weighted orthogonal regression to fit a prediction curve to the data. We compare the fits of the orthogonal regression curves (unweighted and weighted) to unweighted and weighted least squares regression fitted to the same data set. For the computation of the estimated orthogonal regression line one often prefers Principle Component Analysis (PCA) over the difficult to compute direct method. The theory behind PCA and OR are the same as both seek to minimize the perpendicular distances from the line to the data (Wu 2021). The PCA algorithm uses matrix calculus, which is easier and faster to code and apply to the data sets. PCA aims to find a vector **w** such that, when our data is projected onto the vector, the perpendicular distance from the data to **w** is minimized. This is equivalent to maximizing the variation in the data after the projection. Wu (2021) presents the proof of the minimization of the perpendicular distances using matrix calculus.

Let $\boldsymbol{z_i} = (x_i, y_i)$, where $x_i = C_{T_{2i}}$ is the predictor variable and $y_i = C_{T_{6i}}$ is the response variable. Then define $\mathbf{w}^*$ of size (2x1) as

$$\begin{gathered}\mathbf{w}^* = \arg\max_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \|(\boldsymbol{z_i} - \bar{\boldsymbol{z}})^T \mathbf{w}\|^2 \\ \text{such that } \mathbf{w}^T \mathbf{w} = 1,\end{gathered} \tag{3}$$

where $\boldsymbol{z_i}$ is the i^{th} row of size (2x1) from the data set and $\bar{\boldsymbol{z}}$ is a (2x1) vector containing the means of the predictor and response data overall samples. Through the usage of vector transformations, we simplify equation (3) as follows

$$\begin{aligned} \|(\mathbf{z}_i - \bar{\mathbf{z}})^T\mathbf{w}\|^2 &= \left[(\mathbf{z}_i - \bar{\mathbf{z}})^T\mathbf{w}\right]^T (\mathbf{z}_i - \bar{\mathbf{z}})^T\mathbf{w} \\ &= \mathbf{w}^T(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T\mathbf{w}. \end{aligned} \tag{4}$$

Substituting equation (4) into equation (3) produces the following

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \mathbf{w}^T \left[\sum_{i=1}^{n} \frac{1}{n}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T\right] \mathbf{w}. \tag{5}$$

We note that the summation portion in equation (5) is the formula for the covariance of $\boldsymbol{z_i}$. Using Lagrangian multipliers on equation (5) subject to $\mathbf{w}^T\mathbf{w} = 1$ we get

$$f(\mathbf{w}, \lambda) = \mathbf{w}^T\text{cov}(\mathbf{z})\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1). \tag{6}$$

Taking the partial derivative of equation (6) and setting it equal to 0 produces

$$\frac{\partial f}{\partial \mathbf{w}} = 2\text{cov}(\mathbf{z})\mathbf{w} - 2\lambda\mathbf{w} = 0, \tag{7}$$

rearranging equation (7) produces the following $\text{cov}(\mathbf{z})\mathbf{w} = \lambda\mathbf{w}$. Therefore, solving the eigenvector/value problem we note that the largest eigenvector of the covariance matrix of the input data, $\mathbf{w} = \boldsymbol{\xi}_1$ of size (2x1), maximizes equation (3). From the largest eigenvector, $\boldsymbol{\xi}_1$, we can build a line that has minimized orthogonal distances from the data

$$g = m * f + b, \tag{8}$$

where m is the estimated slope, b is the estimated intercept, g is the response variable (eTHPA6), and f is the predictor variable (eTHPA2). Calculating the estimated intercept and slope for a given matrix of data, set the first column as the predictor x and the second column as the response y. For our purposes eTHPA2 is the predictor and eTHPA6 is the response. Calculate the eigendecomposition of the covariance matrix for the data, then manipulate the largest eigenvector $\boldsymbol{\xi_1}$ from the decomposition as follows:

1. The slope $m = \frac{\xi_2}{\xi_1}$ of the estimated line is calculated by dividing the second element by the first element from the largest eigenvector $\boldsymbol{\xi_1}$.
2. The intercept is calculated using $\bar{y} - m\bar{x} = b$.

Thus, PCA produces the estimated line $g = m * f + b$ such that the sum of its perpendicular distances to the set of data points is minimized.

To incorporate weights into orthogonal regression, Therneau (2018) uses theory from Ripley and Thompson (1987) to build an R package which includes weights. Ripley and Thompson (1987) developed a special case of maximum-likelihood fitting of a functional relationship (MLFR) that minimizes the sum of squared perpendicular distances to the fitted line. An overview of the mathematics for the minimization is as follows.

Let ϵ be the error term and let any variable with a hat ($\hat{\alpha}$) denote that variable's estimate. Define x and y as the predictor (eTHPA2) and response (eTHPA6) vectors with associated error ϵ_x and ϵ_y, respectively, for a given data set. The dimensions of x and y are n rows for the technical replicates and 1 column. The data has the form of

$$\begin{aligned} x_i &= u_i + \epsilon_x, \\ y_i &= v_i + \epsilon_y, \end{aligned} \tag{9}$$

with each x_i and y_i measurement containing a small amount of error. The relationship between v_i and u_i is defined by $v_i = f(u_i) = \alpha + \beta u_i$. To estimate α and β, Ripley and Thompson (1987) assume the errors to be independent and normally distributed which leads to the log-likelihood function of

$$L = \text{constant} - \frac{1}{2}\sum \frac{(x_i - u_i)^2}{\kappa_i} - \frac{1}{2}\sum \frac{(y_i - v_i)^2}{\lambda_i}, \tag{10}$$

where κ_i= variance of x_i and λ_i= variance of y_i. Maximizing equation (10) over u_i incorporating the relationship between u_i and v_i, is equivalent to minimizing

$$Q = \sum\left(\frac{(x_i - u_i)^2}{\kappa_i} + \frac{(y_i - \alpha - \beta u_i)^2}{\lambda_i}\right). \tag{11}$$

Expanding equation (11) results in

$$\begin{aligned} Q_i &= \frac{1}{\kappa_i \lambda_i w_i}\{(u_i^2 - 2w_i[\lambda_i x_i + \kappa_i \beta(y_i - \alpha)]u_i) + w_i[\lambda_i x_i^2 + \kappa_i(y_i - \alpha)^2]\} \\ &\text{where } w_i = \frac{1}{\lambda_i + \beta^2 \kappa_i}. \end{aligned} \tag{12}$$

Completing the square for equation (12) gives

$$Q_i = \frac{1}{\kappa_i \lambda_i w_i}[\{u_i - w_i[\lambda_i x_i + \kappa_i \beta(y_i - \alpha)]\}^2 + w_i^2 \kappa_i \lambda_i (y_i - \alpha - \beta x_i)^2]. \tag{13}$$

After completing the square we note that equation (13) is minimized over u_i when $\hat{u}_i = w_i[\lambda_i x_i + \kappa_i \beta(y_i - \alpha)]$.

Substituting $\hat{u}_i$ into (13) yields equation (14)

$$Q_m(\alpha, \beta) = min_{u_i}\, Q = \sum w_i (y_i - \alpha - \beta x_i)^2. \tag{14}$$

Estimates for α and β are derived from equation (14) and are as follows:

the estimate for α is

$$\hat{\alpha} = \frac{\sum w_i (y_i - \beta x_i)}{\sum w_i}, \tag{15}$$

and the estimate for β, $\hat{\beta}$ is found numerically by minimizing (14)

$$Q_m(\hat{\alpha}, \beta) = Q = \sum w_i (y_i - \hat{\alpha}(\beta) - \beta x_i)^2. \tag{16}$$

Ripley and Thompson (1987) note the estimate of $\hat{\alpha}$ depends on β. Ripley and Thompson (1987) recommend an algorithm developed by Nash (1990), that numerically solves for the estimate of β. In Nash's (1990) book he suggests multiple ways to numerically solve for the β in equation (14), however Therneau (2018) derived a shorter way to calculate the estimate for β. Therneau (2018) follows the principles outlined by Nash's (1990) Gauss-Newton algorithm.

The core of the Gauss-Newton algorithm is to fit a linear equation at each iteration of the algorithm until it converges on a minimum (Nash 1990). Therneau (2018) does this by fitting two weighted linear regressions, where the squared standard deviation's (sd) for x and y are the weights in the model (Therneau 2018). The inputs are x and y variables, where both x and y are interchanged, taking turns acting as the predictor and response variables. When a given variable is acting as the response the corresponding squared sd's are the weights for the model. The formula for the two weighted regressions are lm.wfit(predictor, response, weight vector of the response). Applying the formula by Therneau (2018) with interchanging the variables gives

$$\text{lm.wfit}\left(x, y, \frac{1}{sd_y^2}\right) \text{ and lm.wfit}\left(y, x, \frac{1}{sd_x^2}\right). \tag{17}$$

From the two weighted regressions (equation 17) the slopes are used to build an interval where equation (16) will be minimized over to find the optimal β. To perform this minimization, estimates for w_i (equation 12) and $\hat{\alpha}(\beta)$ (equation 15) are found first by optimizing these equations over the built interval; then they are used to find the optimal β in equation (16). Once the optimal β is found the true estimate for $\hat{\alpha}$ is produced using equation (15) based off the optimum $\hat{\beta}$.

2.4 Weighting factor calculation

The weights for the weighted Deming regression prediction curves are calculated using the sample standard deviation. For the full detect data sets, the sample standard deviation's s_j^2, are calculated from the C_T values for a given sample S_j and are used as the weight for each technical replicate r_j in S_j. The weights for the partial detect data sets are sample standard deviations calculated based off the $\hat{\lambda}$ values. Lesperance et al. (2021) used the Delta method to determine the variance of $\hat{\lambda}$, and the square root of the variance was taken to produce the sample standard deviation for the partial detect data. For the weighted linear regression, the methods above are used to calculate the sample variance σ_j^2 for the full and partial detect data sets and $\frac{1}{\sigma_j^2}$ are used as the weights.

2.5 Prediction curve application

The details of the prediction curve validation are as follows for the full and then partial detect field data, respectively.

Full detect data procedure: Figure 2 shows the procedure to predict natural log copy number values from C_T values.

1. Start with the field data processed in the labs to get eTHPA2 and eTHPA6 C_T values.
2. Apply the corresponding eTHPA6 standard curve (BVL or UVic) to the eTHPA6 C_T values from the field data.
3. The results after applying the corresponding eTHPA6 standard curve (BVL or UVic) are called "actual" eTHPA6 natural log copy number values.
4. Use the corresponding prediction curve (BVL or UVic) on eTHPA2 C_T values to produce predicted eTHPA6 $\hat{C_T}$ values.
5. From the predicted $\hat{C_{T_6}}$ values for eTHPA6, apply the corresponding standard curve for eTHPA6.
6. The results after applying the corresponding eTHPA6 standard curve (BVL or UVic) are estimated natural log copy number values for eTHPA6 given eTHPA2.

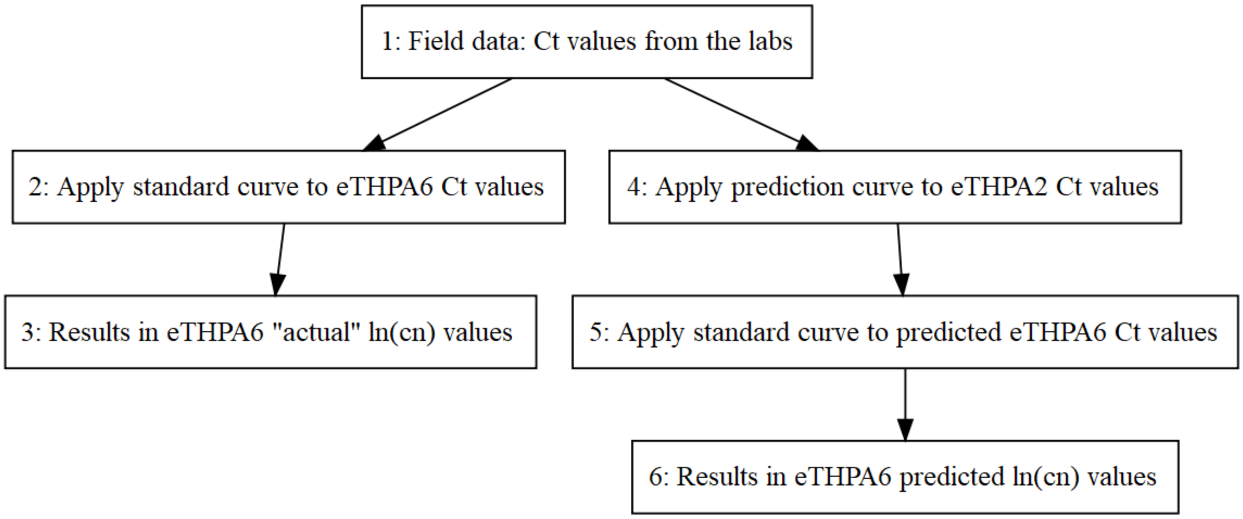


Figure 2: Overview of the procedure used to be able to compare predicted and "actual" cycle threshold and natural log copy number values of eTHPA6. This shows what equations are used to predict eTHPA6 given eTHPA2 Ct values and when the standard curves are applied to produce the final natural log copy number estimates for eTHPA6.

Partial detect data procedure: Figure 3 shows the procedure to predict copy number values from $\hat{p}$ values.

1. Use the field data to calculate $\hat{p_j} = \frac{\sum_{i=1}^{n} \text{detects per sample}}{n}$ for both eTHPA2 and eTHPA6, $j = 1, ..., J$.
2. Convert the $\hat{p_j}$ values to the mean Poisson scale by $\hat{\lambda_j} = -ln(1 - \hat{p_j})$.
3. Use $\hat{\lambda_j}$ values for eTHPA6 and apply the corresponding standard curve for eTHPA6 (BVL or UVic).
4. The results from applying steps 1-3 are "actual" copy number values for eTHPA6.
5. Use the $\hat{\lambda_j}$ values for eTHPA2 and apply the corresponding prediction curve (BVL or UVic).
6. From the predicted $\hat{\lambda_j}$ values for eTHPA6 from eTHPA2, apply the corresponding standard curve for eTHPA6 BVL or UVic).

7. The results after applying the corresponding eTHPA6 standard curve (BVL or UVic) are estimated copy number values for eTHPA6 given eTHPA2.

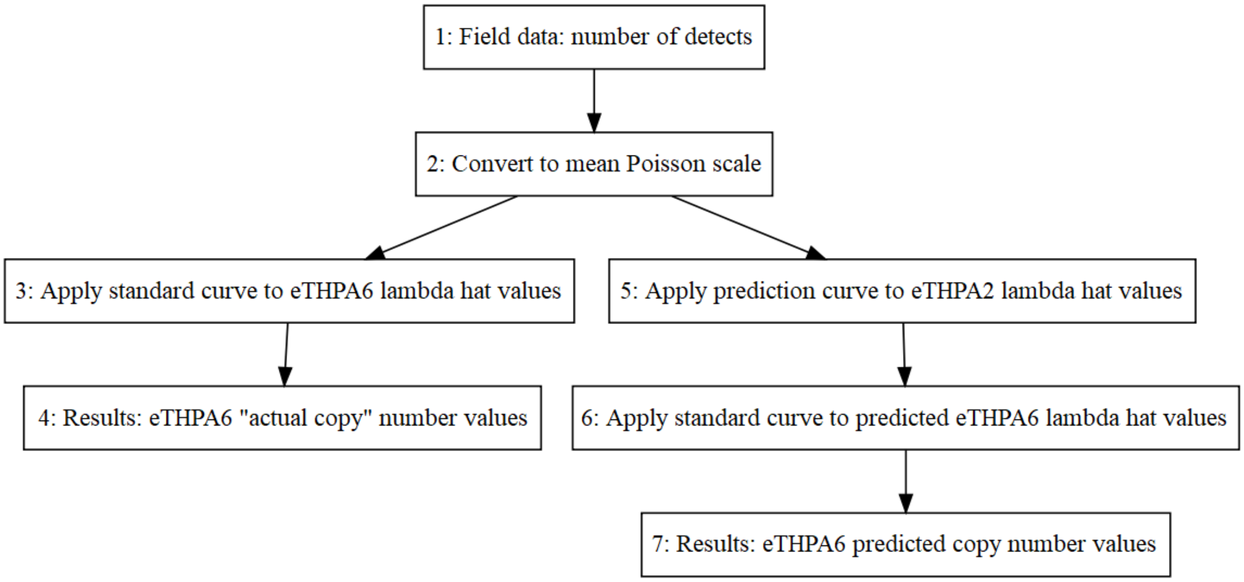


Figure 3: Overview of the procedure used to be able to compare predicted and “actual” lambda hat and copy number values of eTHPA6. This shows what equations are used to predict eTHPA6 given eTHPA2 lambda hat values and when the standard curves are applied to produce the final copy number estimates for eTHPA6.

2.6 Standard error estimation and Confidence interval calculation

In this section we present methods for calculating the standard error’s (se) of the predicted copy number values for the full and partial detect data sets and the standard error’s for the predicted $\hat{\lambda}$ values for the partial detect data. All confidence intervals are approximately normal 95%.

2.6.1 Standard error calculation methods for the predicted values

Standard error for the weighted Deming regression

Alongside the prediction curve’s predicted values, we estimate the variance for the predicted eTHPA6 C_T values using equation (21) from Rio, Riu, and Rius (2001). The derivation for this equation is in a paper written by Lisy, Cholvadova, and Kutej (1990). Let $\mathbf{Y} = b_o + b_1 * \mathbf{X}$ be the general prediction curve equation, where b_o is the intercept estimate and b_1 is the slope estimate. The general prediction curve produces "new" y_o values, in our case predicted eTHPA6 C_T values from "new" x_o, which are eTHPA2 C_T values. The equation for the predicted variance from Rio, Riu, and Rius (2001), in our case produces variance estimates for estimated eTHPA6 C_T values (y_o) using the estimated slope and intercept from the weighted Deming prediction curve. The equation for calculating the predicted variance for the predicted eTHPA6 C_T (y_o) values is as follows

$$s^2_{y_o} = s^2_{b_o} + x^2_o s^2_{b_1} + b^2_1 s^2_{x_o} + 2x_o \, cov(b_o, b_1), \tag{18}$$

where $s^2_{y_o}$ and $s^2_{x_o}$ denotes the sample variance of Y_o and X_o, and $cov(b_o, b_1)$ is the covariance between the estimated intercept and slope. Therefore, each predicted natural log copy number value has an associated sample variance value. The estimated standard errors come from the square root of the variances for the predicted natural log copy number values for all data sets and can be found in section 8.2 of the Appendix.

Standard error for the full detect data standard curves

Equation (10) from Lavagnini and Magno (2007) formulates the variance of the predicted natural log copy number values. The inverse of the general standard curve is $ln(\hat{cn}) = \dfrac{\hat{C_T} - b}{a}$, where a is the slope and b is the intercept. The inverse of the general standard curve intakes predicted or "actual" eTHPA6 C_T values (y_o) and estimates the natural log of their copy number (x_0). The equation Lavagnini and Magno (2007) formulates is for calculating the variances of predicted values from inverse regression curves and is as follows

$$s^2_{\hat{x}_o} = \frac{s^2_{y/x}}{b^2_1}\left(\frac{1}{m} + \frac{1}{n} + \frac{(\hat{x}_o - \bar{x})^2}{\sum_{i=1}^{n} .(x_i - \bar{x})^2}\right), \tag{19}$$

where $s^2_{y/x}$ is the residual sum of squares divided by the degrees of freedom from the error from the standard curve, $\sum_{i=1}^{n}(x_i - \bar{x})^2 = var(x) * (n-1)$ with the x values being those used to build the standard curve, b_1^2 is the slope of the standard curve, n is the total number of observations used to build the standard curve, m is the number of technical replicated in a field sample, $\bar{x}$ is the mean of the x values that are used to build the standard curve (log of the starting quantities), and $\hat{x}_o$ is the sample mean of the natural log predicted values, and in our case this is the mean of the predicted eTHPA6 natural log copy number values by sample. The calculations for all data sets using this method are found in section 8.2 of the appendix.

Standard errors for the partial detect data standard curves

To derive the standard error for the copy number values we employ the Delta method. The general standard curve equation from the eLowQuant algorithm is

$$\hat{cn} = \frac{\hat{\lambda}}{\hat{\beta}}, \tag{20}$$

where $\hat{\beta}$ the slope estimate calculated from the eLowQuant algorithm and $\hat{\lambda}$ is the estimate of λ. Applying the variance function (Var) to both sides of equation (20) given $\hat{\lambda}$ we get

$$Var(\hat{cn}) = \hat{\lambda}^2 * Var\left(\frac{1}{\hat{\beta}}\right). \tag{21}$$

The $Var\left(\frac{1}{\hat{\beta}}\right)$ is estimated using the Delta method, where the $g(x) = \frac{1}{x}$ is the continuously differentiable function transformation. From the Delta method it follows that if $X \approx N(\mu, V)$ then $g(X) \approx N(g(\mu), g'(\mu)^2 * V)$. The first derivative of $g(x)$ is

$$g'(x) = -\frac{1}{x^2}. \tag{22}$$

Here, we assume that $\hat{\beta} \approx N(\beta, \hat{V})$, where β is the true slope and $\hat{V}$ is the estimated variance of $\hat{\beta}$. For our purposes we estimate the true slope β with $\hat{\beta}$, the estimated slope of the regression. Therefore, the $Var\left(\frac{1}{\hat{\beta}}\right) \approx \left(-\frac{1}{\hat{\beta}^2}\right)^2 * \hat{V}$. To calculate the standard error of the copy number values we take the square root of equation (21) which produces $\sqrt{Var(\hat{cn})} \approx \hat{\lambda} * \sqrt{Var\left(\frac{1}{\hat{\beta}}\right)}$, where the square root of the variance is the standard error of the given quantity.

2.7 Error propogation from the standard curves

Due to the multistep process of value conversion through prediction and standard curve application, the error is compounded. We can quantify this error growth using the following equation (1) from Tellinghuisen (2001)

$$\sigma_f^2 = \boldsymbol{g}^{\boldsymbol{T}} \boldsymbol{V} \boldsymbol{g}, \tag{23}$$

where σ_f^2 is the variance, $\boldsymbol{V}$ is the variance-covariance matrix of the β parameters of the least squares model for the standard curve, and $\boldsymbol{g}$ is a vector who's i^th element is the partial derivative of f with respect to β_i. We use equation (23) in section 8.3 of the appendix to calculate the final error growth of the copy number values from standard curve for all field data sets.

Matrices for the full detect data

For the full detect data, we have the standard curve for eTHPA6 as $f = ln(cn_6) = \dfrac{C_{T6} - a_6}{b_6}$, where b and a are the slope and intercept, respectively. Then $\boldsymbol{g}$, the vector of partial derivatives of the parameters of the standard curves has form

$$\boldsymbol{g} = \begin{bmatrix} \dfrac{\partial f}{\partial C_{T6}} = \dfrac{1}{b_6} \\ \dfrac{\partial f}{\partial a_6} = -\dfrac{1}{b_6} \\ \dfrac{\partial f}{\partial b_6} = -\dfrac{C_{T6} - a_6}{b_6^2} \end{bmatrix}. \tag{24}$$

$\boldsymbol{V}$ the variance-covariance matrix of the parameters from the standard curves has form

$$\boldsymbol{V} = \begin{bmatrix} Var(C_{T6}) & 0 & 0 \\ 0 & Var(a_6) & Cov(a_6, b_6) \\ 0 & Cov(b_6, a_6) & Var(b_6) \end{bmatrix}, \tag{25}$$

where the $Var(C_{t6})$ is the variance estimate from equation (18) for C_T values.

Matrices for the partial detect data

For the partial detect data, we have the standard curve for eTHPA6 as $f = cn_6 = \dfrac{\hat{\lambda}_6}{\beta_6}$, where beta is the slope. Then $\boldsymbol{g}$, the vector of partial derivatives of the parameters of the standard curve has form

$$\boldsymbol{g} = \begin{bmatrix} \dfrac{\partial f}{\partial \hat{\lambda}_6} = \dfrac{1}{\beta_6} \\ \dfrac{\partial f}{\partial \beta_6} = -\dfrac{\hat{\lambda}_6}{\beta_6^2} \end{bmatrix}. \tag{26}$$

$\boldsymbol{V}$ the variance-covariance matrix of the parameters from the standard curve has form

$$\boldsymbol{V} = \begin{bmatrix} Var(\hat{\lambda}_6) & 0 \\ 0 & Var(\beta_6) \end{bmatrix}, \tag{27}$$

where $Var(\hat{\lambda}_6)$ is the estimated variance from equation (18) for $\hat{\lambda}$ values.

3 Data

This section provides samples of the data sets, plots and their statistical summaries.

3.1 gblock data

This section displays the summary statistics for the full and partial UVic and BVL gblock data, alongside plots for their C_T and copy number values.

3.1.1 UVic: full detect data

This portion of the UVic full detect gblock data displays the starting quantity (copy numbers) and the cycle threshold values for eTHPA2 and eTHPA6 (Table 2). Figure 4 displays the cycle threshold values from the full detect gblock data set. eTHPA6 and eTHPA2 display a possible linear relationship with a decrease in variance corresponding to increasing starting quantity (Figure 4). Cycle threshold values increase with decreasing starting quantity (Figure 4). The values naturally group together according to their starting quantity.

Table 2: Sample of UVic full detect gblock data: first 10 rows. Starting quantity is quantified by copy number and Ct values for eTHPA2 and eTHPA6.

Starting Quantity	eTHPA2 Ct	eTHPA6 Ct
62500	25.15	20.61
62500	25.06	20.89
62500	24.61	20.93
62500	24.90	21.20
62500	25.15	20.85
62500	24.62	20.95
62500	24.88	20.97
62500	25.24	21.07
12500	27.30	23.29
12500	27.58	23.51

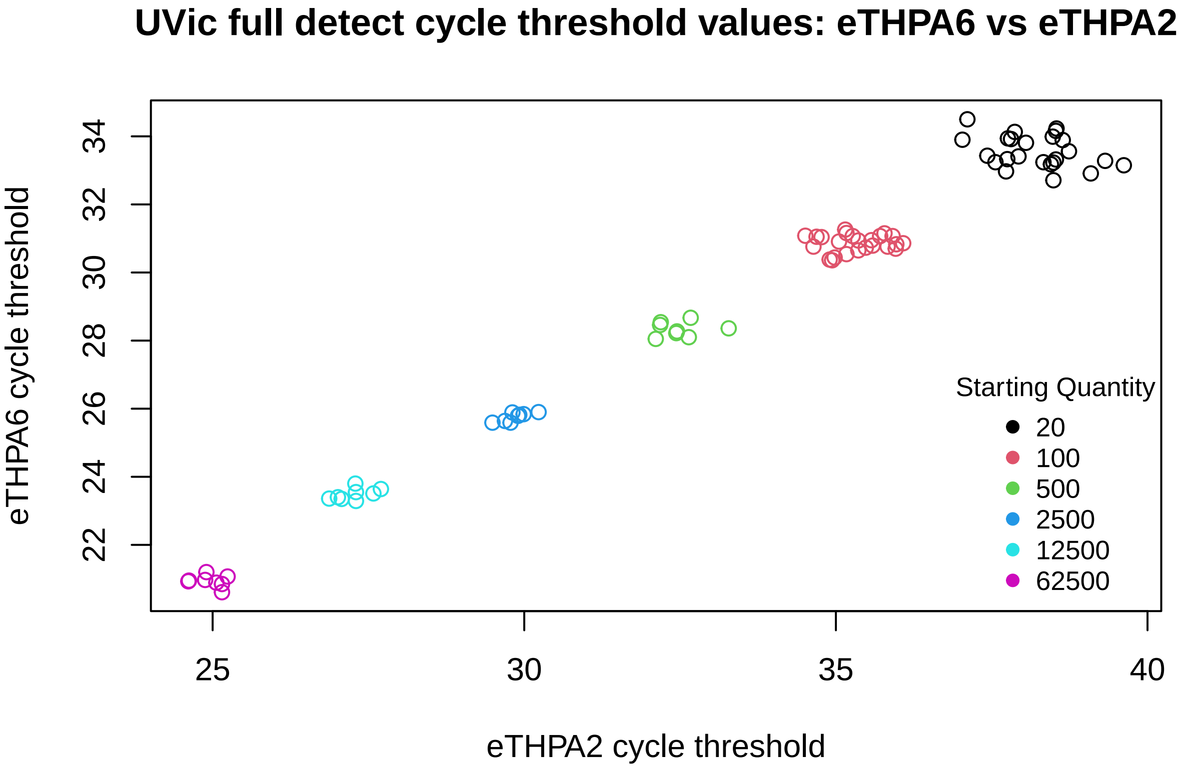


Figure 4: The UVic full detect gblock data. the cycle threshold values are grouped by starting quantity (copy numbers).

Table 3: Summary statistics for the UVic full detect gblock data: cycle threshold values. Starting quantity is quantified by copy number, N is the total number of technical replicates in a sample. Estimates for the minimum, maximum, mean, median, sample variance and standard deviation by starting quantity are shown.

Starting Quantity	N	Min	Max	Mean	Median	Sample variance	Sample standard deviation
eTHPA2							
20	24	37.03	39.62	38.22	38.39	0.423	0.650
100	24	34.51	36.08	35.33	35.31	0.215	0.464
500	8	32.11	33.28	32.49	32.45	0.145	0.380
2500	8	29.49	30.23	29.85	29.86	0.048	0.218
12500	8	26.87	27.70	27.27	27.30	0.079	0.280
62500	8	24.61	25.24	24.95	24.98	0.058	0.241
eTHPA6							
20	24	32.71	34.50	33.56	33.42	0.222	0.471
100	24	30.36	31.26	30.86	30.89	0.064	0.253
500	8	28.05	28.67	28.33	28.32	0.046	0.215
2500	8	25.59	25.90	25.76	25.80	0.017	0.131
12500	8	23.29	23.80	23.49	23.45	0.030	0.172
62500	8	20.61	21.20	20.93	20.94	0.029	0.171

From Table 3 range for eTHPA2 and eTHPA6 C_T values are (24-40) and (20-35), respectively. We use these ranges to define the acceptable range for field data C_T values and use this acceptable range to remove any field data that is out of their corresponding range. The variance of the cycle threshold values decreases with increasing starting quantity. The construction of the standard curve from the full detect gblock data uses regular least squares, however the variability seen in the standard deviations suggest that a weighted regression should be used to build the standard curves. This issue is left for future investigation.

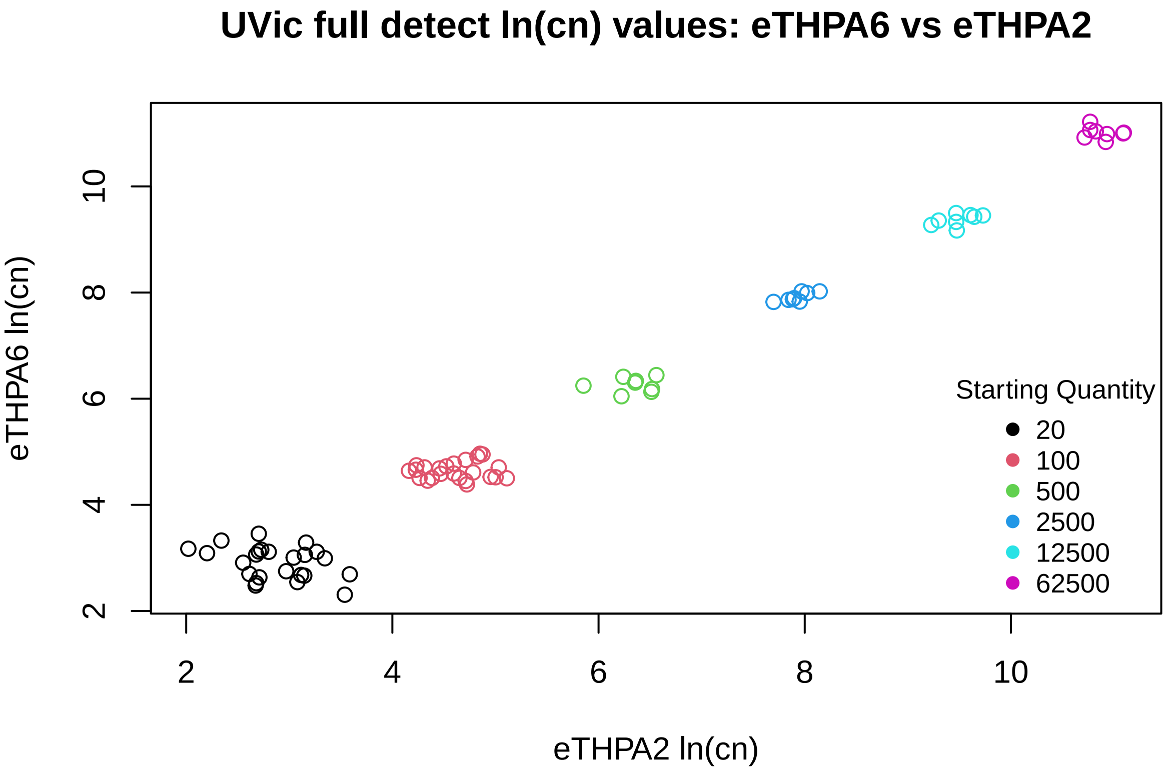


Figure 5: The UVic full detect gblock data. The Ct values are converted to natural log copy number values using the UVic standard curve for full detect data, grouped by starting quantity (copy numbers).

Figure 5 displays the natural log copy number values converted from the UVic full detect C_T values, using the corresponding standard curves in section 2.2 of the methods. eTHPA6 and eTHPA2 natural log copy number values display a possible linear relationship with a decrease in variance corresponding to increasing starting quantity (Figure 5). The natural log copy number values increase with increasing starting quantity and group together according the their starting quantity (Figure 5).

3.1.2 UVic: partial detect data

Table 4 displays a portion of the UVic partial detect gblock data. The partial detect UVic data set has 120 observations; 5 included starting quantities with 24 technical replicates. The N/A cycle threshold values within the data set indicate the given replicate did not detect any DNA. The cycle threshold N/A's are represented by 0 in the detect vector, and C_T values are represented by a 1 in the detect vector for a given replicate. The UVic partial detect gblock data includes starting quantity 4.

Table 4: Sample of UVic partial detect gblock data: first 10 rows. Starting quantity is quantified by copy number and Ct and detect values for eTHPA2 and eTHPA6.

Starting quantity	eTHPA2 Ct	eTHPA2 detect	eTHPA6 Ct	eTHPA6 detect
4	NA	0	36.73	1
4	38.28	1	36.78	1
4	39.70	1	36.40	1
4	39.46	1	37.41	1
4	38.95	1	37.50	1
4	41.75	1	35.91	1
4	39.11	1	36.67	1
4	41.46	1	35.76	1
4	40.77	1	37.47	1
4	38.94	1	37.71	1

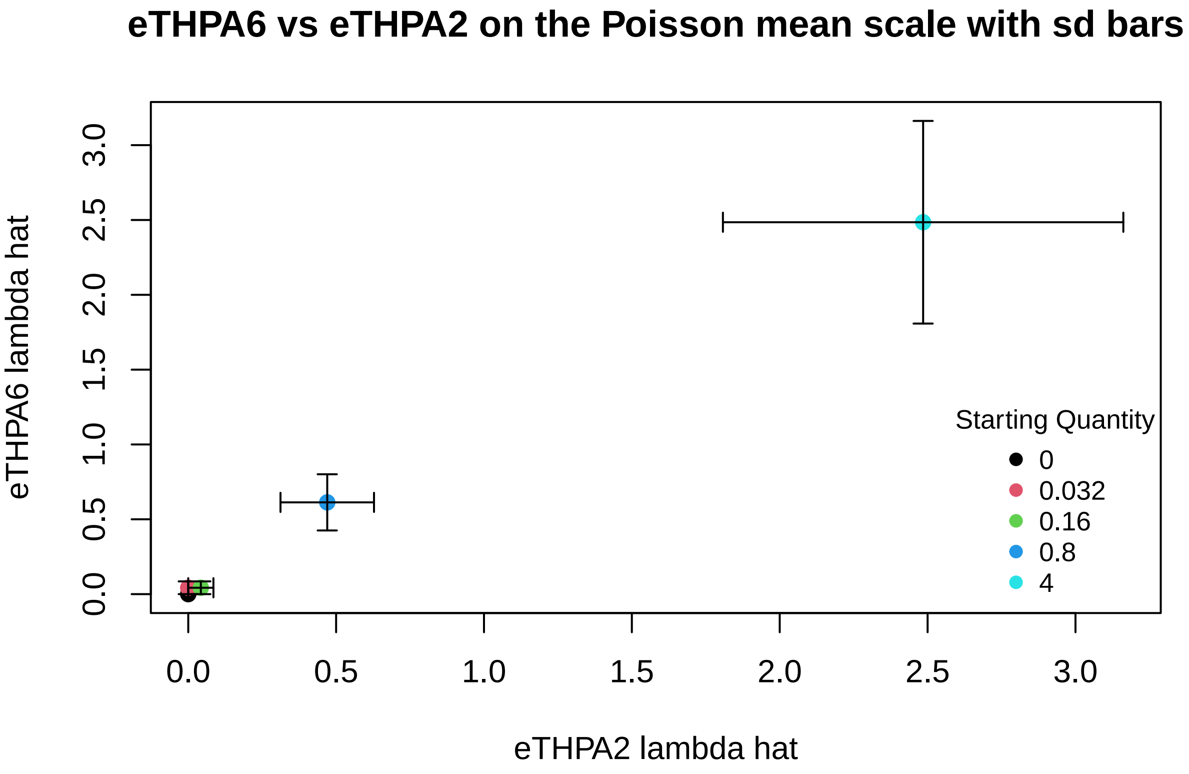


Figure 6: The relationship between eTHPA6 and eTHPA2 on the Poisson mean (lambda hat) scale with standard deviation (sd) bars. Lambda hat estimates are calculated from the proportion of detects per sample, phat, as lambda hat = -ln(1 - phat) from the UVic partial detect gblock data.

The $\hat{\lambda}$ values calculated from the partial detect gblock data using methodology from (Lesperance et al. 2021). Figure 6 shows a potential linear relationship between eTHPA6 and eTHPA2, with an increase in variance with increase in starting quantity shown by the increasing standard deviation bars around the points. The $\hat{\lambda}$ values for eTHPA6 and eTHPA2 increase with increasing starting quantity (Figure 6).

Table 5: Summary statistics by sample for the UVic partial detect gblock data: eTHPA2 and eTHPA6. Starting quantity is quantified by copy number, n is the number of technical replicates, v indicates variance of variable, sd implies standard deviation of variable, phat is the proportion of detects, and lamhat (lambda hat) = -ln(1 - phat).

Starting quantity	n	Detect	phat	Sample vphat	lamhat	Sample vlamhat	Sample sd lamhat
eTHPA2							
0	24	0	0.0000	0.0000	0.0000	0.0000	0.0000
0.032	24	0	0.0000	0.0000	0.0000	0.0000	0.0000
0.16	24	1	0.0417	0.0017	0.0426	0.0018	0.0426
0.8	24	9	0.3750	0.0098	0.4700	0.0250	0.1581
4	24	22	0.9167	0.0032	2.4849	0.4583	0.6770
eTHPA6							
0	24	0	0.0000	0.0000	0.0000	0.0000	0.0000
0.032	24	1	0.0417	0.0017	0.0426	0.0018	0.0426
0.16	24	1	0.0417	0.0017	0.0426	0.0018	0.0426
0.8	24	11	0.4583	0.0103	0.6131	0.0353	0.1878
4	24	22	0.9167	0.0032	2.4849	0.4583	0.6770

Table 5 above shows the summary statistics for eTHPA2 and eTHPA6 in terms of phat and $\hat{\lambda}$ variables with their variances and standard deviations. Looking at the number of detects, there is a sufficient amount of data to perform analysis (Table 5). Here the acceptable range for $\hat{\lambda}$ for eTHPA2 and eTHPA6 is (0-2.4849) (Table 5).

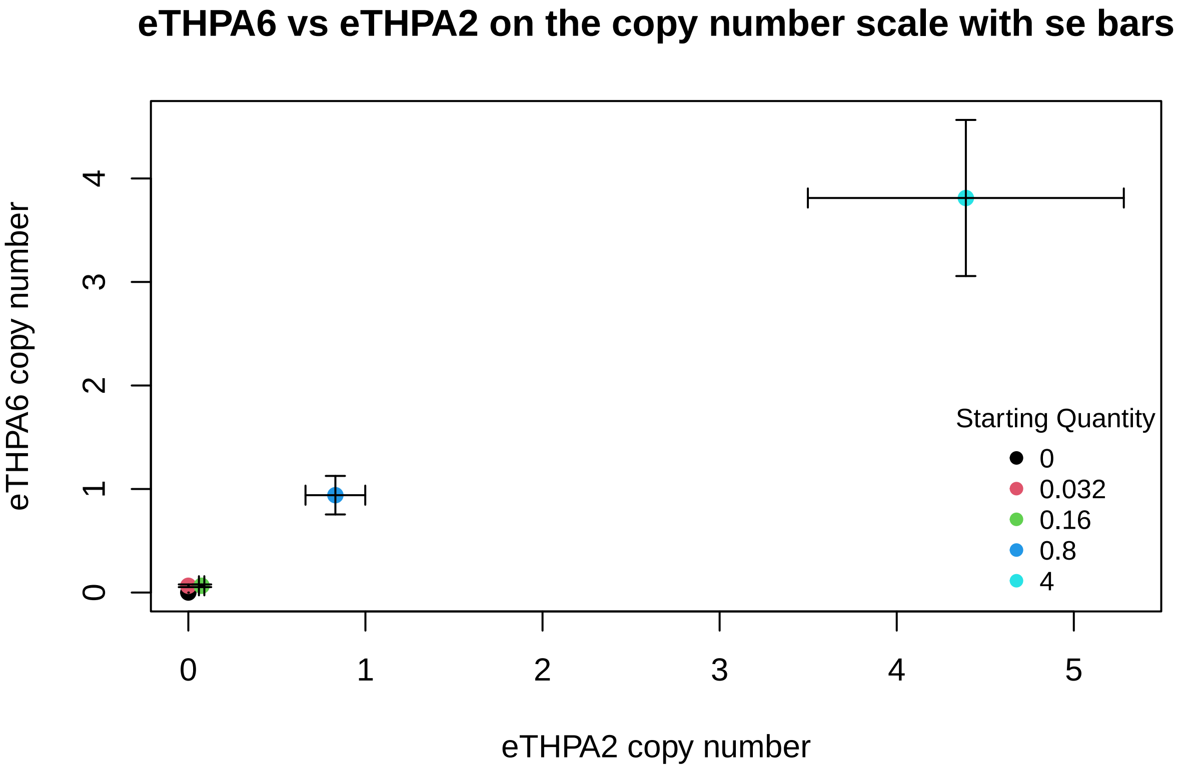


Figure 7: The relationship between eTHPA6 and eTHPA2 on the copy number scale. The copy number estimates are calculated from lambda hat values using the corresponding standard curve for UVic partial detect gblock data. The standard error (se) bars are calculated using the Delta method in section 2.6.

Figure 7 shows the copy number values converted from the $\hat{\lambda}$ from the UVic partial detect data. eTHPA6 and eTHPA2 copy number values display a possible linear relationship with increasing copy number for increasing starting quantity (Figure 7). The standard error (se) of the copy numbers increases moderately with increasing starting quantity (Figure 7).

3.1.3 BVL: full detect data

This portion of the BVL full detect gblock data displays the starting quantity (copy numbers) and the cycle threshold values for eTHPA2 and eTHPA6 (Table 6). Figure 8 displays the cycle threshold values from the full detect gblock data set. eTHPA6 and eTHPA2 display a possible linear relationship with a decrease in variance corresponding to increase in starting quantity (Figure 8). Cycle threshold values increase with decreasing starting quantity (Figure 8). The values naturally group together according to their starting quantity. Note that the BVL full detect gblock data includes starting quantity 4, unlike the UVic full detect gblock data that has starting quantity 20 as the minimum.

Table 6: Sample of full detect BVL assay data: first 10 rows. Starting quantity is quantified by copy number and Ct values for eTHPA2 and eTHPA6.

Starting Quantity	eTHPA2 Ct	eTHPA6 Ct
62500	24.36	23.08
62500	24.31	22.75
62500	24.23	22.96
62500	24.35	23.00
62500	23.94	22.83
62500	24.44	22.82
62500	24.19	22.69
62500	24.32	23.04
12500	26.97	25.68
12500	26.82	25.66

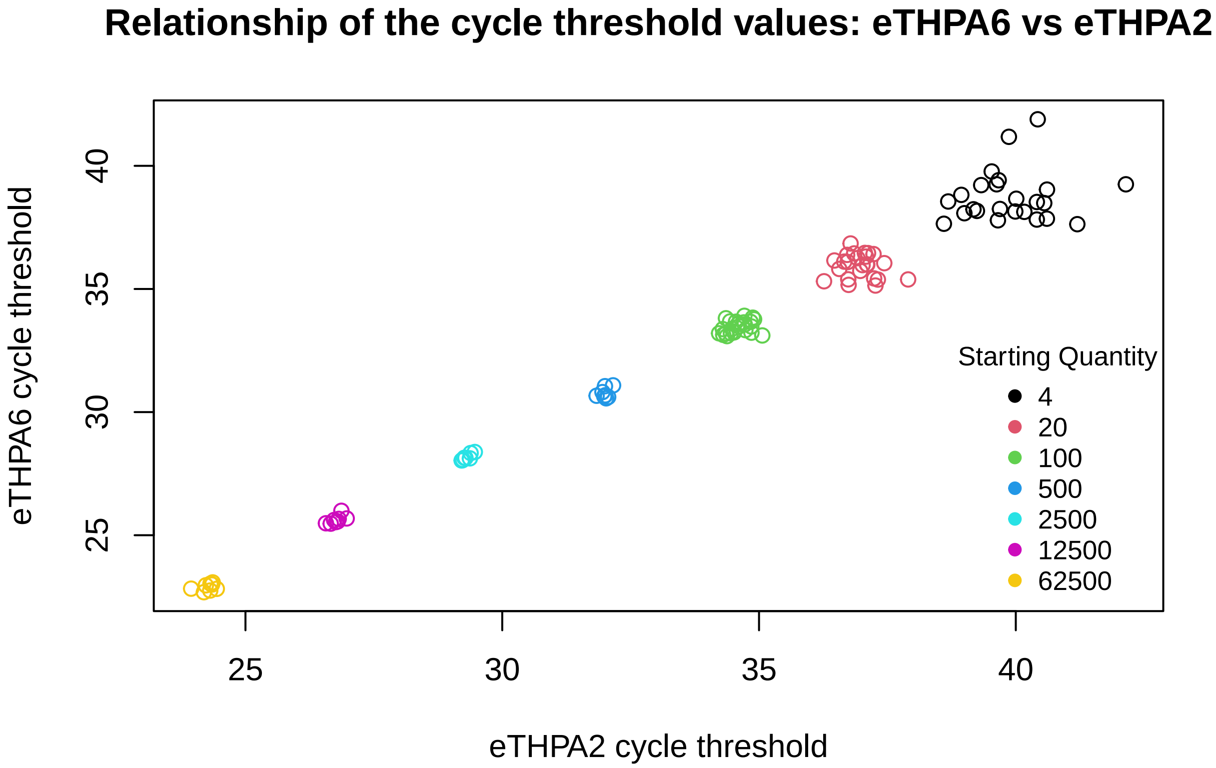


Figure 8: BVL full detect gblock data. The cycle threshold values are grouped by starting quantity (copy numbers).

Table 7: Summary statistics for the BVL full detect gblock data: cycle threshold values. Starting quantity is quantified by copy number, N is the total number of technical replicates in a sample. Estimates for the minimum, maximum, mean, median, sample variance and standard deviation by starting quantity are shown.

Starting quantity	N	Min	Max	Mean	Median	Sample variance	Sample standard deviation
eTHPA2							
4	24	38.60	42.14	39.90	39.78	0.678	0.823
20	24	36.27	37.90	36.97	36.99	0.122	0.350
100	24	34.22	35.06	34.60	34.57	0.052	0.228
500	8	31.84	32.16	32.00	32.00	0.008	0.091
2500	8	29.21	29.47	29.30	29.27	0.009	0.092
12500	8	26.57	26.97	26.77	26.78	0.015	0.124
62500	8	23.94	24.44	24.27	24.32	0.023	0.153
eTHPA6							
4	24	37.63	41.89	38.74	38.51	1.101	1.049
20	24	35.14	36.85	35.95	36.02	0.229	0.479
100	24	33.09	33.91	33.45	33.43	0.065	0.254
500	8	30.56	31.09	30.76	30.67	0.041	0.202
2500	8	28.04	28.38	28.16	28.13	0.017	0.130
12500	8	25.47	25.99	25.62	25.59	0.028	0.167
62500	8	22.69	23.09	22.90	22.90	0.021	0.143

From Table 7 the acceptable range for eTHPA2 and eTHPA6 C_T values are (23-43) and (22-42), respectively. We use the defined acceptable ranges to remove any field data that is out of their corresponding interval in the later analysis. The variance of the cycle threshold values decreases with increasing starting quantity. The construction of the standard curve from the full detect gblock data uses regular least squares, however the variability seen in the standard deviations suggest that a weighted regression should be used to build the standard curves. This issue is left for future investigation.

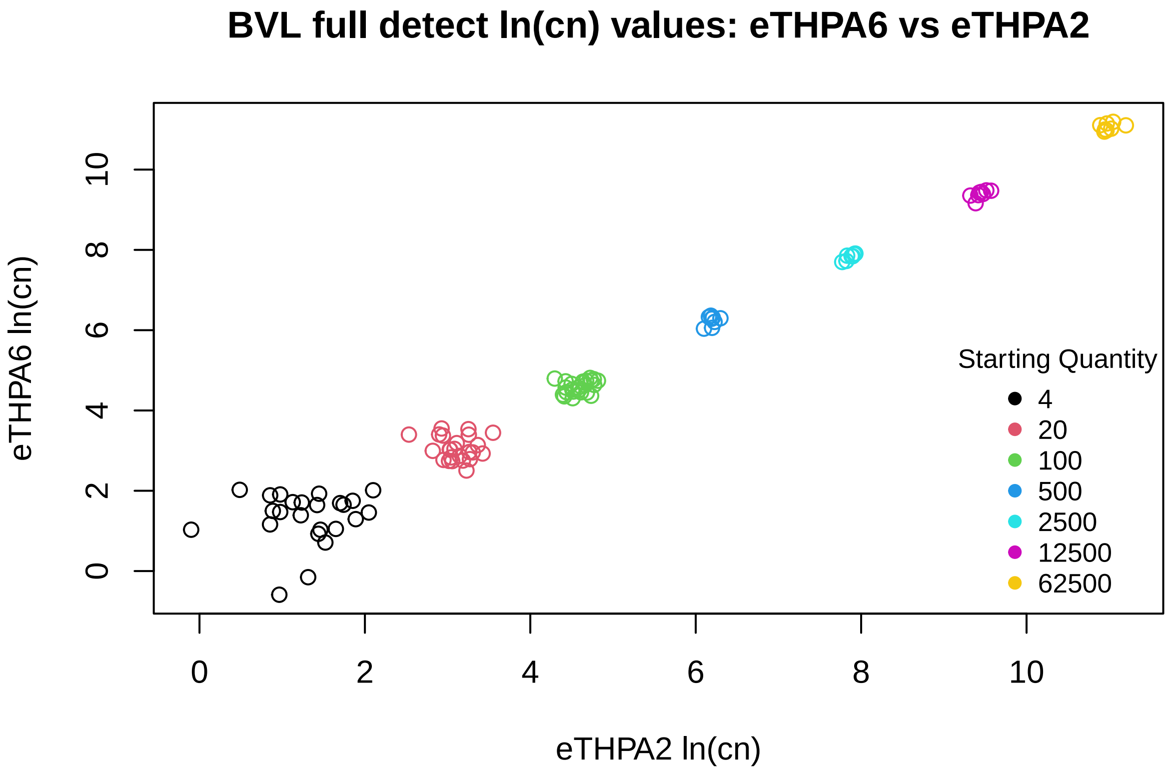


Figure 9: The BVL full detect gblock data. The Ct values are converted to natural log copy number values using the BVL standard curve for full detect data, grouped by starting quantity (copy numbers).

Figure 9 displays the natural log copy number values converted from the BVL full detect C_T values, using the corresponding standard curve for natural log copy number from C_T values in section 2.2 of the methods. eTHPA6 and eTHPA2 natural log copy number values display a possible linear relationship with a decrease in variance corresponding to increasing starting quantity (Figure 9). The natural log copy number values increase with increasing starting quantity and group together according the their starting quantity (Figure 9).

3.1.4 BVL: partial detect data

Table 8 displays a portion of the BVL partial detect gblock data. The partial detect BVL data has 96 observations; 4 included starting quantities with 24 technical replicates. The N/A cycle threshold values within the data set indicate the given replicate did not detect any DNA. These non-detects are represented in the detect vector by a 0, and 1 represents a detect for a given replicate. The BVL partial detect data does not include starting quantity 4, compared to the UVic partial detect data which does include starting quantity 4.

Table 8: Sample of BVL partial detect gblock data: first 10 rows. Starting quantity is quantified by copy number and Ct and detect values for eTHPA2 and eTHPA6.

Starting quantity	eTHPA2 Ct	eTHPA6 Ct	eTHPA2 detect	eTHPA6 detect
0.8	44.02	39.62	1	1
0.8	NA	40.26	0	1
0.8	NA	43.04	0	1
0.8	42.40	41.16	1	1
0.8	41.58	40.70	1	1
0.8	NA	40.61	0	1
0.8	42.95	41.57	1	1
0.8	NA	40.22	0	1
0.8	41.03	NA	1	0
0.8	41.61	41.66	1	1

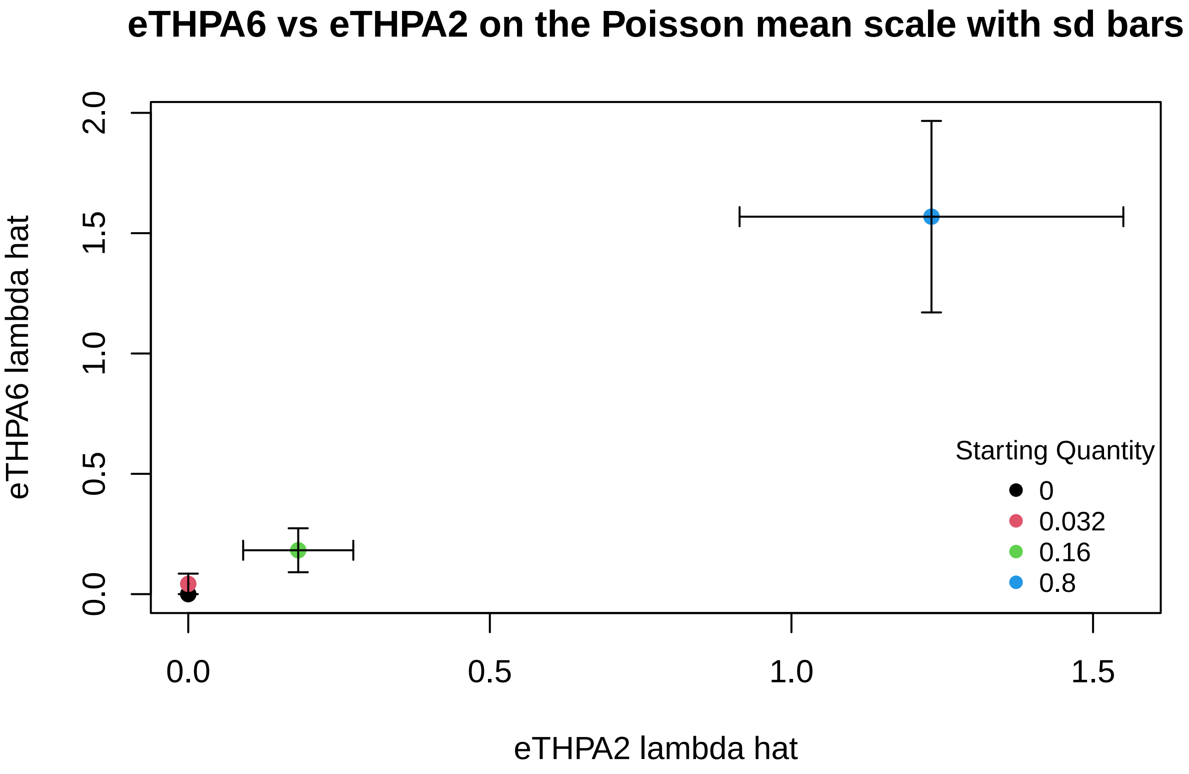


Figure 10: The relationship between eTHPA6 and eTHPA2 on the Poisson mean (lambda hat) scale with standard deviation (sd) bars. Lambda hat estimates are calculated from the proportion of detects per sample, phat, as lambda hat = -ln(1 - phat) from the BVL partial detect gblock data.

The $\hat{\lambda}$ values are calculated from the partial detect gblock data using methodology from (Lesperance et al. 2021). Figure 10 shows a potential linear relationship between eTHPA6 and eTHPA2, with an increase in variance with increase in starting quantity shown by the increasing standard deviation bars around the points. The $\hat{\lambda}$ values for eTHPA6 and eTHPA2 increase with increasing starting quantity (Figure 10).

Table 9: Summary statistics by sample for the BVL partial detect gblock data: eTHPA2 and eTHPA6. Starting quantity is quantified by copy number, n is the number of technical replicates, v indicates variance of variable, sd implies standard deviation of variable, phat is the proportion of detects, and lamhat (lambda hat) = -ln(1 - phat).

Starting quantity	n	Detect	phat	Sample vphat	lamhat	Sample vlamhat	Sample sd lamhat
eTHPA2							
0	24	0	0.0000	0.0000	0.0000	0.0000	0.0000
0.032	24	0	0.0000	0.0000	0.0000	0.0000	0.0000
0.16	24	4	0.1667	0.0058	0.1823	0.0083	0.0913
0.8	24	17	0.7083	0.0086	1.2321	0.1012	0.3181
eTHPA6							
0	24	0	0.0000	0.0000	0.0000	0.0000	0.0000
0.032	24	1	0.0417	0.0017	0.0426	0.0018	0.0426
0.16	24	4	0.1667	0.0058	0.1823	0.0083	0.0913
0.8	24	19	0.7917	0.0069	1.5686	0.1583	0.3979

Table 9 shows the summary statistic for eTHPA2 and eTHPA6 in terms of phat and $\hat{\lambda}$ variables with their variances and or standard deviations. Looking at the number of detects, there is a sufficient amount of data to perform analysis (Table 9). The acceptable range for $\hat{\lambda}$ values for eTHPA2 and eTHPA6 is 0-1.3 and 0-1.6, respectively.

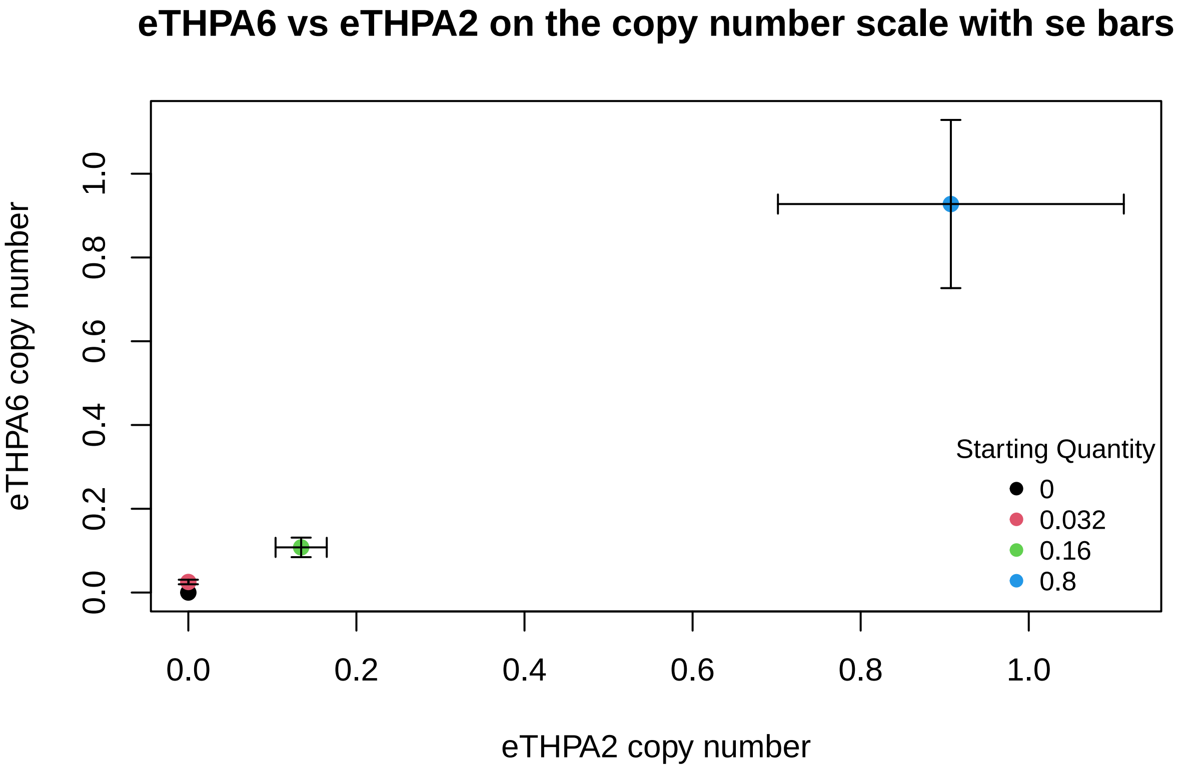


Figure 11: The relationship between eTHPA6 and eTHPA2 on the copy number scale. The copy number estimates are calculated from lambda hat values using the corresponding standard curve for BVL partial detect gblock data. The standard error (se) bars are calculated using the Delta method.

Figure 11 shows the copy number values converted from the $\hat{\lambda}$ from the UVic partial detect data. eTHPA6 and eTHPA2 copy number values display a possible linear relationship with increasing copy number for increasing starting quantity (Figure 11). The standard error for the copy number values does not increase as much when compared to the UVic partial detect copy number values (Figure 7).

3.2 Field data

In this section are the summary statistics and plots for the full and partial detect UVic and BVL field data. UVic's field data is referred to as Kemano and includes other sampling sites. BVL's field data is referred to as Fraser river and includes other sites. All removed and rearranged field data are shown in supplementary excel files.

3.2.1 UVic: full detect data

The UVic full detect field data is from 2021 and does not include any 2023 data. The 2023 full detect field data is currently being processed.

Table 10: Sample of UVic full detect Kemano field data: first 10 rows. sample number and Ct values for eTHPA2 and eTHPA6.

Sample	eTHPA2 Ct	eTHPA6 Ct
1	36.57	32.66
1	35.66	32.40
1	36.03	32.65
1	36.42	32.94
1	35.91	32.52
1	35.48	32.53
1	35.98	32.98
1	36.40	33.32
2	38.06	34.05
2	36.73	33.99

Table 10 displays a portion of the UVic full detect Kemano field data. The Uvic full detect Kemano data set has 80 observations; 10 samples with 8 replicates each. The acceptable range for eTHPA2 (24-40) and eTHPA6 (20-35) values is set by the UVic full detect gblock data. No field data points fell below the lower range limit in either assay, however 12 observation fall above the upper limit and are removed. The analysis for the full detect field data is conducted on 68 observations with added observations from the partial detect Kemano field data set. Observations from the partial detect field data set that do not lie within the acceptable partial detect λ range are added to the full data set if they fall with the full detect acceptable C_T value range.

Table 11: Summary statistics by sample for the UVic full detect Kemano field data: cycle threshold values. N is the total number of acceptable technical replicates in a sample. Estimates for the minimum, maximum, mean, median, variance and standard deviation by sample are shown.

Sample	N	Min	Max	Mean	Median	Sample variance	Sample standard deviation
eTHPA2							
1	8	35.48	36.57	36.06	36.01	0.147	0.384
2	8	36.25	38.06	37.02	37.01	0.318	0.564
3	8	35.10	36.47	35.75	35.74	0.288	0.537
4	2	39.43	39.96	39.70	39.70	0.140	0.375
5	8	32.58	33.36	32.99	32.99	0.065	0.256
6	2	39.55	39.73	39.64	39.64	0.016	0.127
7	8	28.14	29.15	28.66	28.68	0.107	0.327
8	8	31.69	32.33	31.91	31.86	0.038	0.196
9	8	29.96	30.70	30.35	30.41	0.065	0.254
10	8	32.74	33.60	33.26	33.35	0.093	0.305
11	3	39.31	39.82	39.62	39.73	0.074	0.272
eTHPA6							
1	8	32.40	33.32	32.75	32.66	0.094	0.306
2	8	33.28	34.81	33.89	33.89	0.212	0.461
3	8	31.32	32.60	31.88	31.85	0.211	0.460
4	2	34.07	34.93	34.50	34.50	0.370	0.608
5	8	28.42	28.94	28.71	28.76	0.034	0.185
6	2	34.57	34.87	34.72	34.72	0.045	0.212
7	8	25.66	26.05	25.83	25.82	0.019	0.138
8	8	28.32	28.78	28.52	28.52	0.033	0.183
9	8	26.64	27.13	26.86	26.83	0.033	0.182
10	8	29.89	30.45	30.12	30.09	0.044	0.210
11	3	34.48	35.25	34.91	35.01	0.155	0.394

Two samples contained the out of range technical replicate C_T values that are removed (Table 11). Sample 11 is from the 2021 UVic partial detect Kemano field data (Table 11). Sample 11 had full detect (8/8) in eTHPA6 and partial detect (6/8) in eTHPA2. Due to sample 11's full detect in eTHPA6 it was moved to the high copy data set. Sample 11 had three out of range and two non detects removed, leaving 3 technical replicate for the analysis.

3.2.2 UVic: partial detect data

The UVic partial detect field data has two data sets for the years of 2021 and 2023.

Table 12: Sample of 10 observations from the UVic 2021 partial detect Kemano field data: cycle threshold values for eTHPA2 and eTHPA6, sample number, detection indicator, and year of sampling event.

Sample	eTHPA2 Ct	eTHPA2 detect	eTHPA6 Ct	eTHPA6 detect	Year
1	NA	0	36.84	1	2021
1	NA	0	36.63	1	2021
1	48.94	1	36.23	1	2021
1	46.08	1	35.42	1	2021
1	39.08	1	34.56	1	2021
1	39.50	1	35.93	1	2021
1	41.47	1	35.07	1	2021
1	41.52	1	NA	0	2021
2	37.77	1	35.07	1	2021
2	39.44	1	35.26	1	2021

Table 12 displays a portion of the UVic partial detect Kemano field data for 2021. The 2021 partial detect Kemano field data has 40 observations; 5 samples with 8 replicates each. Two of the samples, samples 4 and 5, are removed from the data set. Sample 4 is removed and combined with the UVic full detect field data set. Sample 5 is removed because eTHPA2 had too few detects (4/8) compared to eTHPA6, which had full detection (8/8) (Table 13). This uneven detection ratio leaves us with too few technical replicates to produce a fair $\hat{\lambda}$ estimate for the given sample. The N/A cycle threshold values within the data set indicate the given replicate did not detect any DNA. These non-detects are represented in the detect vector by a 0, and 1 represents a detect for a given replicate.

Table 13: Sample 5 removed from the 2021 partial detect Kemano field data: cycle threshold values for eTHPA2 and eTHPA6, sample number, detect indicator, and year of sampling event.

Sample	eTHPA2 Ct	eTHPA2 detect	eTHPA6 Ct	eTHPA6 detect	Year
5	39.86	1	34.94	1	2021
5	NA	0	37.98	1	2021
5	NA	0	36.97	1	2021
5	NA	0	35.32	1	2021
5	40.20	1	35.59	1	2021
5	NA	0	36.29	1	2021
5	44.10	1	37.11	1	2021
5	40.08	1	37.02	1	2021

Table 14 displays a portion of the UVic partial detect Kemano field data for 2023. The 2023 UVic partial detect kemano field data has 64 observations; 8 samples with 8 technical replicates each. Two of the samples, samples 2 and 5, are removed from the data set (Table 15). Sample 2 and 5 are removed because eTHPA2 lacked enough detects (5/8) compared to eTHPA6, which had full detection (8/8). This uneven detection ratio leaves us with too few technical replicates to produce a fair $\hat{\lambda}$ estimate for the given sample. Samples 2 and 5 cannot have the full detect analysis applied, as all of their C_T values are out of the full detect acceptable C_T value range of (24-40) for eTHPA2 and (20-35) for eTHPA6.

Table 14: Sample of 10 observations of the 2023 partial detect Kemano field data: cycle threshold values for eTHPA2 and eTHPA6, sample number, detect indicator, and year of sampling event.

Sample	eTHPA2 Ct	eTHPA2 detect	eTHPA6 ct	eTHPA6 detect	Year
1	NA	0	45.32	1	2023
1	NA	0	43.66	1	2023
1	NA	0	42.13	1	2023
1	NA	0	41.13	1	2023
1	NA	0	NA	0	2023
1	49.48	1	43.58	1	2023
1	NA	0	NA	0	2023
1	43.11	1	NA	0	2023
3	NA	0	NA	0	2023
3	NA	0	40.19	1	2023

Table 15: Samples 2 and 5 removed from the 2023 partial detect Kemano field data: sample number, cycle threshold values and detect indicator for eTHPA2 and eTHPA6, and year of sampling.

Sample	eTHPA2 Ct	eTHPA2 detect	eTHPA6 Ct	eTHPA6 detect	Year
2	44.74	1	39.51	1	2023
2	NA	0	37.12	1	2023
2	42.69	1	38.15	1	2023
2	43.17	1	37.76	1	2023
2	NA	0	37.31	1	2023
2	42.79	1	38.18	1	2023
2	41.74	1	36.87	1	2023
2	NA	0	37.20	1	2023
5	NA	0	41.01	1	2023
5	45.14	1	40.00	1	2023
5	41.95	1	43.12	1	2023
5	45.76	1	41.59	1	2023
5	43.38	1	43.54	1	2023
5	44.75	1	42.70	1	2023
5	NA	0	42.36	1	2023

Table 16: Summary statistics for UVic partial detect Kemano field data by sample for the number of detects (2021/2023): eTHPA2 and eTHPA6. Here n is the number of technical replicates and detect is the number of detects per sample. A letter v indicates variance of variable, sd implies standard deviation of variable, phat is the proportion of detects, and lamhat (lambda hat) = -ln(1 - phat).

Sample	n	Detect	phat	Sample vphat	lamhat	Sample vlamhat	Sample sd lamhat
eTHPA2: 2021							
1	8	6	0.750	0.0234	1.3863	0.3750	0.6124
2	8	5	0.625	0.0293	0.9808	0.2083	0.4564
3	8	2	0.250	0.0234	0.2877	0.0417	0.2041
eTHPA6: 2021							
1	8	7	0.875	0.0137	2.0794	0.8750	0.9354
2	8	7	0.875	0.0137	2.0794	0.8750	0.9354
3	8	4	0.500	0.0312	0.6931	0.1250	0.3536
eTHPA2: 2023							
1	8	2	0.250	0.0234	0.2877	0.0417	0.2041
3	8	2	0.250	0.0234	0.2877	0.0417	0.2041
4	8	3	0.375	0.0293	0.4700	0.0750	0.2739
6	8	2	0.250	0.0234	0.2877	0.0417	0.2041
7	8	2	0.250	0.0234	0.2877	0.0417	0.2041
8	8	5	0.625	0.0293	0.9808	0.2083	0.4564
eTHPA6: 2023							
1	8	5	0.625	0.0293	0.9808	0.2083	0.4564
3	8	5	0.625	0.0293	0.9808	0.2083	0.4564
4	8	4	0.500	0.0312	0.6931	0.1250	0.3536
6	8	2	0.250	0.0234	0.2877	0.0417	0.2041
7	8	4	0.500	0.0312	0.6931	0.1250	0.3536
8	8	6	0.750	0.0234	1.3863	0.3750	0.6124

The $\hat{\lambda}$ values for the UVic partial detect Kemano field data (2021/2023) do not show any values outside the acceptable range of (0-2.48) for both eTHPA2 and eTHPA6 (Table 16). None of the values from Table 16 look to be outliers.

3.2.3 BVL: full detect data

Table 17 displays a portion of the BVL full detect Fraser river field data. The BVL full detect Fraser river field data has 184 observations; 23 samples with 8 replicates each. The acceptable range for eTHPA2 (23-43) and eTHPA6 (22-42) values is set by the BVL full detect gblock data. No field data points fell below the lower range limit in either assay, observations that fall above the upper limit are removed. The analysis on the field data is conducted on the remaining observations with added observations from the partial detect Fraser river field data set. Partial detect field data that does not fall within the partial detect acceptable λ range are removed and move to the full detect data set if they fall within the acceptable full detect C_T range

Table 17: Sample of BVL full detect Fraser river field data: first 10 rows. Sample number and Ct values for eTHPA2 and eTHPA6.

Sample	eTHPA2 Ct	eTHPA6 Ct
1	34.00	34.34
1	33.75	33.96
1	33.85	33.95
1	33.71	34.12
1	33.69	34.35
1	33.72	34.19
1	33.88	33.60
1	34.04	34.35
2	34.14	34.74
2	34.09	34.61

Table 18: Summary statistics by sample for BVL full detect Fraser river field data: cycle threshold values. N is the total number of acceptable technical replicates in a sample. Estimates for the minimum, maximum, mean, median, sample variance and standard deviation by sample are shown.

Sample	N	Min	Max	Mean	Median	Sample variance	Sample standard deviation
eTHPA2							
1	8	33.69	34.04	33.83	33.80	0.018	0.135
2	8	33.70	37.28	34.47	34.12	1.335	1.155
3	8	31.32	32.35	31.75	31.68	0.123	0.351
4	8	32.65	33.13	32.85	32.80	0.036	0.190
5	8	33.23	34.46	33.92	34.09	0.229	0.478
6	8	28.65	29.00	28.87	28.92	0.017	0.130
7	8	28.79	29.20	28.94	28.91	0.020	0.140
8	8	32.92	33.35	33.10	33.08	0.018	0.133
9	8	28.00	28.28	28.11	28.11	0.007	0.086
10	8	34.36	34.70	34.54	34.55	0.015	0.122
11	8	25.30	25.64	25.44	25.41	0.016	0.126
12	8	25.17	25.36	25.29	25.30	0.003	0.055
13	8	39.43	40.60	39.92	39.86	0.137	0.370
14	8	38.66	41.05	39.67	39.45	0.784	0.886
15	8	37.85	41.10	39.82	40.01	0.953	0.976
16	8	38.96	41.19	39.85	39.63	0.528	0.727
17	8	39.21	42.39	40.78	40.80	1.677	1.295
18	6	39.67	42.84	41.28	41.35	1.388	1.178
19	6	38.82	41.03	40.16	40.19	0.558	0.747
20	8	39.31	41.09	40.38	40.42	0.355	0.596
21	8	38.77	40.41	39.58	39.51	0.318	0.564
22	8	37.60	39.01	38.36	38.48	0.277	0.526
23	6	40.33	42.95	41.62	41.61	0.814	0.902
24	4	40.58	42.51	41.30	41.06	0.764	0.874
25	4	40.86	41.03	40.95	40.95	0.005	0.070
26	4	40.04	41.37	40.73	40.76	0.472	0.687
eTHPA6							
1	8	33.60	34.35	34.11	34.16	0.069	0.263
2	8	33.60	34.74	34.22	34.13	0.126	0.356
3	8	28.14	29.02	28.46	28.41	0.066	0.256
4	8	29.80	30.92	30.22	30.20	0.106	0.325
5	8	29.69	30.75	30.09	30.08	0.102	0.320
6	8	26.54	27.16	26.80	26.75	0.063	0.252
7	8	26.47	27.23	26.71	26.62	0.074	0.272
8	8	29.97	31.31	30.44	30.33	0.188	0.433
9	8	26.71	27.23	26.93	26.90	0.036	0.189
10	8	31.64	32.17	31.87	31.83	0.042	0.205
11	8	23.80	24.18	23.94	23.90	0.023	0.150
12	8	23.86	24.25	23.97	23.95	0.017	0.132
13	8	37.31	39.64	38.34	38.30	0.577	0.759
14	8	37.06	38.49	37.78	37.71	0.222	0.471
15	8	36.86	39.01	38.12	38.29	0.668	0.817
16	8	37.32	40.34	38.57	38.45	0.749	0.866
17	8	37.47	39.73	38.25	37.97	0.564	0.751
18	6	39.71	41.65	40.55	40.45	0.626	0.791
19	6	39.43	40.99	39.99	39.84	0.321	0.566
20	8	38.82	40.87	39.65	39.58	0.417	0.645
21	8	37.41	39.03	38.05	38.01	0.294	0.542
22	8	36.25	38.01	37.05	37.09	0.314	0.560
23	6	38.88	40.14	39.50	39.38	0.261	0.511
24	4	39.34	41.32	40.17	40.02	0.689	0.830
25	4	38.64	39.79	39.36	39.50	0.293	0.541
26	4	38.35	41.49	39.60	39.29	1.817	1.348

From the summary statistics in Table 18, one sample from the full detect Fraser river field data was out of range and removed. Many samples have removed out of range values. Samples 23-26 come from the BVL partial detect Fraser river field data as either eTHPA2 or eTHPA6 had detects greater than or equal to (7/8) with the other having greater than or equal to (5/8) detects. Samples 23-26 produced C_T values that are out of the acceptable range, thus they are removed.

3.2.4 BVL: partial detect data

Table 19 displays a portion of the BVL partial detect Fraser River field data. The BVL partial detect Fraser river field data has 152 observations; 19 samples with 8 replicates each. Samples 1, 13, 15, 16 are removed and added to the full detect data analysis, as either eTHPA2 or eTHPA6 had full detects (8/8) and the other assay had greater than or equal to (5/8) detects (Table 20). Samples 2, 3, 14, and 19 are removed completely as they contain all out of range values and could not have full or partial detect analysis applied. The N/A cycle threshold values within the data set indicate the given replicate did not detect any DNA. These non-detects are represented in the detect vector by a 0, and 1 represents a detect for a given replicate.

Table 19: Sample of 10 observations of the partial detect Fraser river field data: sample number and cycle threshold values and detect indicator for eTHPA2 and eTHPA6.

Sample	eTHPA2 Ct	eTHPA2 detect	eTHPA6 Ct	eTHPA6 detect
4	NA	0	NA	0
4	NA	0	NA	0
4	NA	0	NA	0
4	43.08	1	42.41	1
4	NA	0	NA	0
4	NA	0	NA	0
4	NA	0	NA	0
4	NA	0	NA	0
5	43.15	1	NA	0
5	NA	0	NA	0

Table 20: Samples 1, 13, 15, and 16 of the Fraser river field data: sample number, cycle threshold values and detect indicator for eTHPA2 and eTHPA6.

Sample	eTHPA2 Ct	eTHPA2 detect	eTHPA6 Ct	eTHPA6 detect
1	40.33	1	39.57	1
1	40.56	1	NA	0
1	41.45	1	39.18	1
1	41.07	1	40.14	1
1	42.15	1	38.88	1
1	42.95	1	39.18	1
1	46.30	1	40.43	1
1	41.77	1	40.05	1
13	41.38	1	40.08	1
13	42.51	1	39.34	1
13	40.75	1	41.32	1
13	41.67	1	42.34	1
13	44.32	1	41.24	1
13	NA	0	39.63	1
13	NA	0	39.64	1
13	40.58	1	39.95	1
15	40.86	1	39.79	1
15	43.24	1	41.57	1
15	40.96	1	39.24	1
15	43.40	1	40.21	1
15	40.93	1	38.64	1
15	41.03	1	39.76	1
15	NA	0	38.65	1
15	44.66	1	39.77	1
16	41.25	1	NA	0
16	41.37	1	38.35	1
16	40.24	1	39.54	1
16	41.24	1	42.23	1
16	NA	0	45.08	1
16	40.04	1	39.04	1
16	41.27	1	41.49	1
16	41.39	1	42.43	1

Table 21: Summary statistics for BVL partial detect Fraser river field data by sample for the number of detects: eTHPA2 and eTHPA6. Here n is the number of technical replicates and detect is the number of detects per sample. A letter v indicates variance of variable, sd implies standard deviation of variable, phat is the proportion of detects, and lamhat (lambda hat) = -ln(1 - phat).

Sample	n	Detect	phat	Sample vphat	lamhat	Sample vlamhat	Sample sd lamhat
eTHPA2							
1	8	1	0.125	0.0137	0.1335	0.0179	0.1336
2	8	2	0.250	0.0234	0.2877	0.0417	0.2041
3	8	2	0.250	0.0234	0.2877	0.0417	0.2041
4	8	2	0.250	0.0234	0.2877	0.0417	0.2041
5	8	1	0.125	0.0137	0.1335	0.0179	0.1336
6	8	1	0.125	0.0137	0.1335	0.0179	0.1336
7	8	3	0.375	0.0293	0.4700	0.0750	0.2739
8	8	3	0.375	0.0293	0.4700	0.0750	0.2739
9	8	1	0.125	0.0137	0.1335	0.0179	0.1336
10	8	4	0.500	0.0312	0.6931	0.1250	0.3536
11	8	5	0.625	0.0293	0.9808	0.2083	0.4564
eTHPA6							
1	8	1	0.125	0.0137	0.1335	0.0179	0.1336
2	8	2	0.250	0.0234	0.2877	0.0417	0.2041
3	8	3	0.375	0.0293	0.4700	0.0750	0.2739
4	8	4	0.500	0.0312	0.6931	0.1250	0.3536
5	8	2	0.250	0.0234	0.2877	0.0417	0.2041
6	8	5	0.625	0.0293	0.9808	0.2083	0.4564
7	8	3	0.375	0.0293	0.4700	0.0750	0.2739
8	8	5	0.625	0.0293	0.9808	0.2083	0.4564
9	8	5	0.625	0.0293	0.9808	0.2083	0.4564
10	8	6	0.750	0.0234	1.3863	0.3750	0.6124
11	8	6	0.750	0.0234	1.3863	0.3750	0.6124

Table 21 displays the summary statistics for the remaining partial detect field data. All of the $\hat{\lambda}$ values are within the allowable range for analysis (Table 21).

4 Prediction Curve Modeling

This section details the construction of the prediction curves for each lab's full and partial detect gblock data. The models for the partial detect data sets have intercepts of 0, since the starting quantity of 0 has $\hat{\lambda} = 0$ for both eTHPA2 and eTHPA6. The methods for prediction curve fitting are unweighted and weighted least squares and unweighted and weighted orthogonal (Deming) regression. These models use the gblock data as it is subject only to technical/instrumental error. We compare the recommended orthogonal regression (unweighted and weighted) to the canonical least squares regression (unweighted and weighted), to investigate the difference in estimated parameters, fit of the estimated lines to the data, and decide on the best line. The summaries for each lab's fitted models for full and partial detect gblock data sets are in section 8.1 of the Appendix. Weighted Deming models are built from the field data and their fit is assessed in section 8.4 of the appendix.

4.1 UVic: full detect gblock data

Least squares regression

The top plot in Figure 12 shows the estimated least squares line for eTHPA6 versus eTHPA2 cycle threshold values and appears to fit the data adequately. The starting quantity 62500 group is concentrated however, as the starting quantity decreases to 20, the data becomes more spread out (Figure 12, top). The cycle threshold values decrease with increasing starting quantity (Figure 12, top). Comparing the line of equality to our fitted line, we note how different they are (Figure 12, top). This difference between the lines indicates the intercept for the estimated least squares regression line is different 0 (Figure 12, top). The slopes of the two lines appear parallel, which could indicate the slope is not different from 1 (Figure 12, top). The residual plot of the least squares model indicates a non-constant variance implying the model does not fit the data well (Figure 12, bottom). This non-constant variance is confirmed by the increasing variation within the data as the starting quantity decreases.

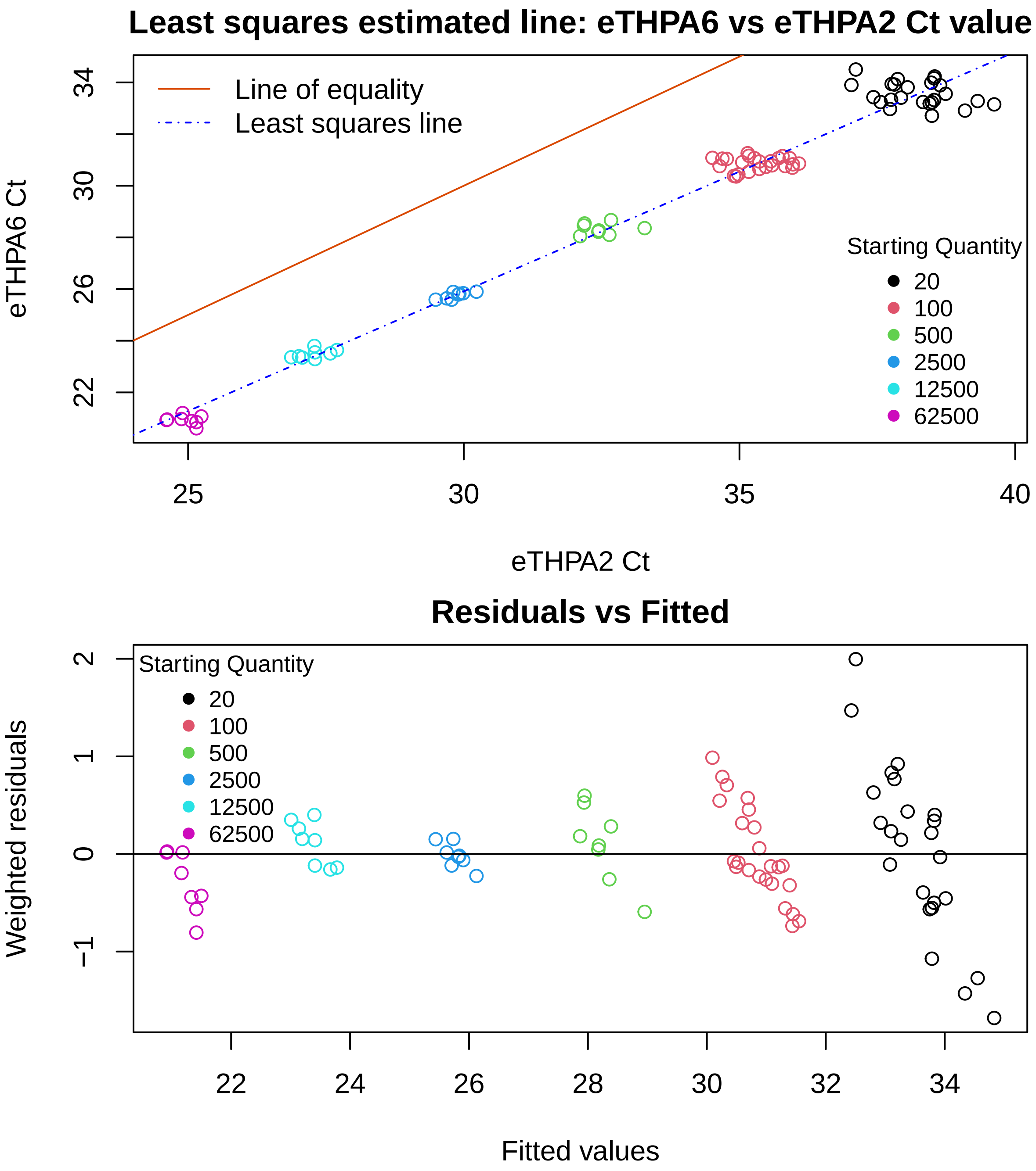


Figure 12: Top: The estimated least squares regression line overlaid with the line of equality for eTHPA6 vs eTHPA2 cycle threshold values. Bottom: the residual plot for the least squares regression on eTHPA6 vs eTHPA2 cycle threshold values. This is used to check the constant variance assumption, the fit of the model, and for outliers.

Weighted least squares regression

Figure 13 (top) shows the estimated weighted least squares line for eTHPA6 versus eTHPA2 cycle threshold values. The model appears to fit the data adequately and it does not show much difference compared to the estimated least squares (Figure 12, top). The weighted least squares line differs greatly from the line of equality, indicating the estimated intercept for the weighted least squares is different from 0 (Figure 13, top). The slopes of the two lines appear parallel, indicating the estimated slope may not be different from 1 (Figure 13, top). The residual plot of the weighted least squares model indicates a constant variance and that the model fits well (Figure 13, bottom). This constant variance is confirmed by the random scatter of data points about y=0 line. The calculation methods for the weights are in section 2.4 of the methods.

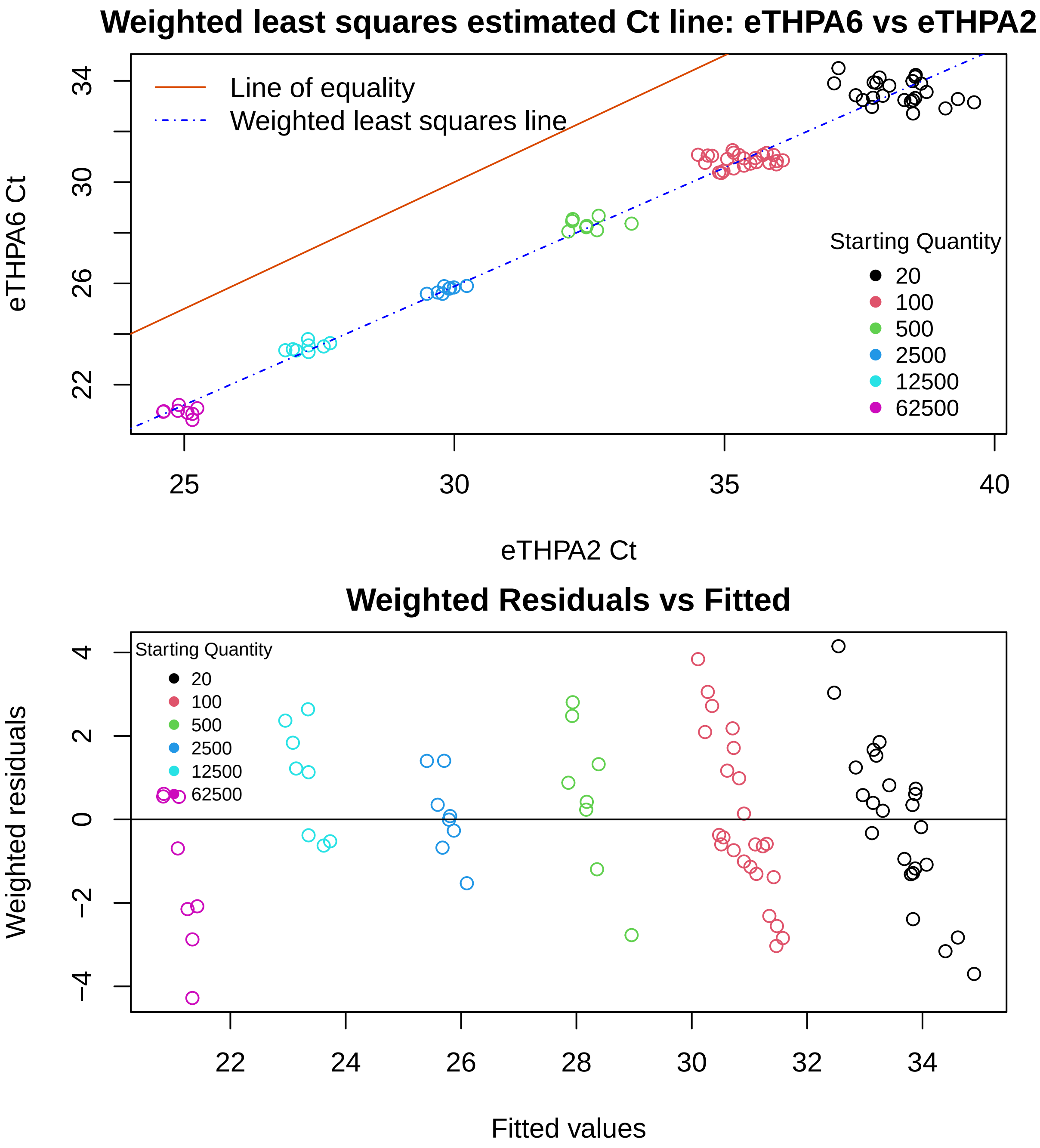


Figure 13: Top: The estimated weighted least squares regression line overlaid with the line of equality for eTHPA6 vs eTHPA2 cycle threshold values. Bottom: the residual plot for the weighted least squares regression on eTHPA6 vs eTHPA2 cycle threshold values. This is used to check the constant variance assumption and the fit of the model.

Deming resgression

Figure 14 (top) shows the estimated Deming line for eTHPA6 versus eTHPA2 cycle threshold values. The model appears to fit the data adequately and does not show much difference compared to the estimated unweighted and weighted least squares (Figure 12 and 13, top). Comparing the line of equality to our fitted line, we note a large difference (Figure 14, top). This difference between the lines indicates the intercept for the estimated Deming line is different from 0 (Figure 14, top). The slopes of the two lines appear parallel, which could indicate the slope may not be different from 1 (Figure 14, top). The residual plot of the Deming model indicates a non-constant variance and implies the model may not fit the data well (Figure 14, bottom). This non-constant variance is confirmed by the increasing variation within the data as the starting quantity decreases (Figure 12, bottom).

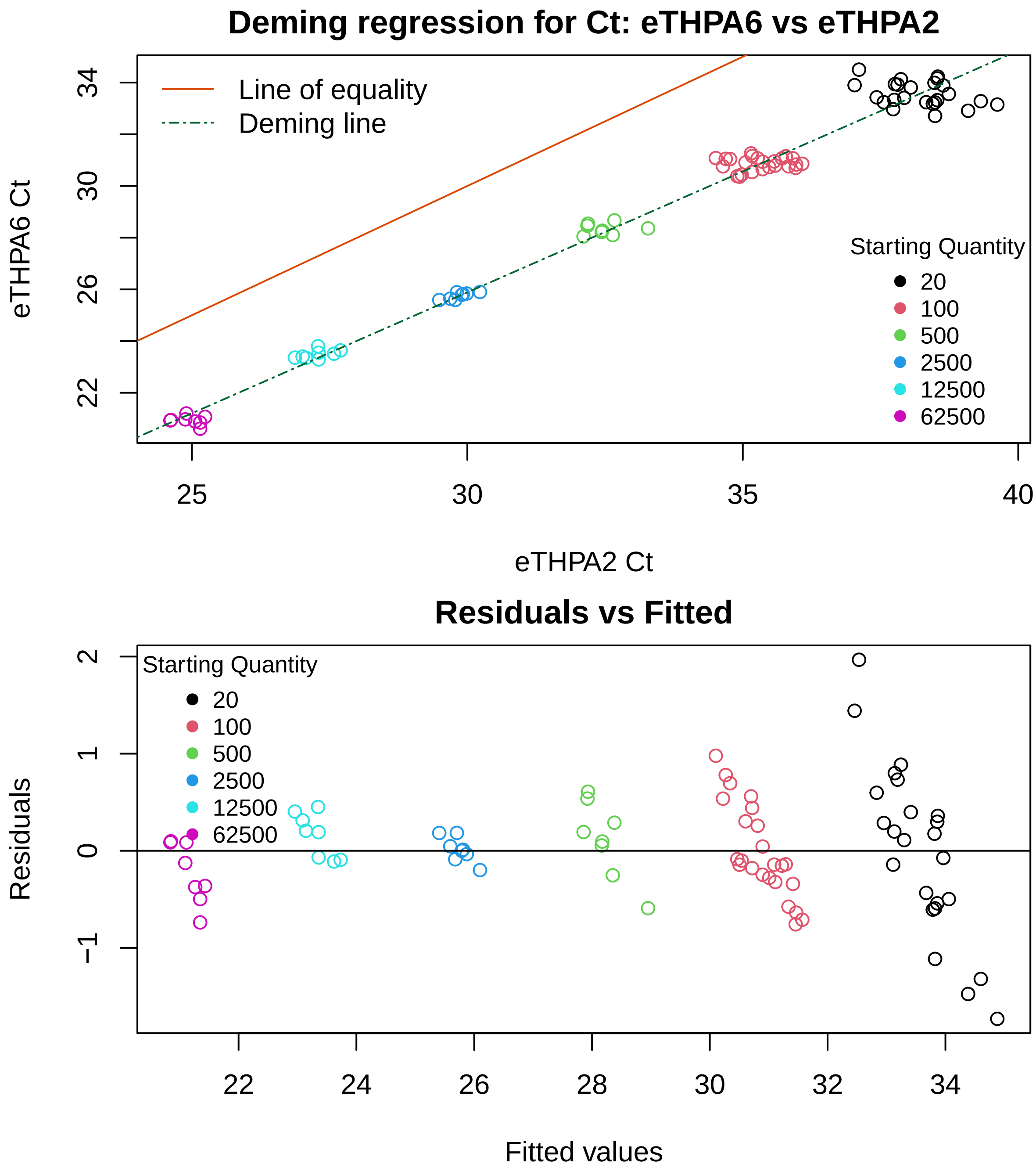


Figure 14: The estimated Deming regression line overlaid with the line of equality for eTHPA6 vs eTHPA2 cycle threshold values. Bottom: the residual plot for the Deming regression on eTHPA6 vs eTHPA2 cycle threshold values. This is used to check the constant variance assumption and the fit of the model.

Weighted Deming regression

Similar to the previous fitted models, the estimated weighted Deming looks to fit the data adequately (Figure 15, top). The cycle threshold values decrease with increasing starting quantity (Figure 15, top). Comparing the line of equality to our fitted line, we note a large difference (Figure 15, top). This difference between the lines indicates the intercept of the weighted Deming regression is different from 0 (Figure 15, bottom). The slopes of the two lines appear parallel, which could indicate the slope may not be different from 1 (Figure 15, bottom). The residual plot of the weighted Deming model indicates a constant variance and that the model fits well (Figure 15, bottom). This constant variance is confirmed by random scatter of data points around the line y=0 (Figure 15, bottom). The calculation methods for the weights are in section 2.4 of the methods.

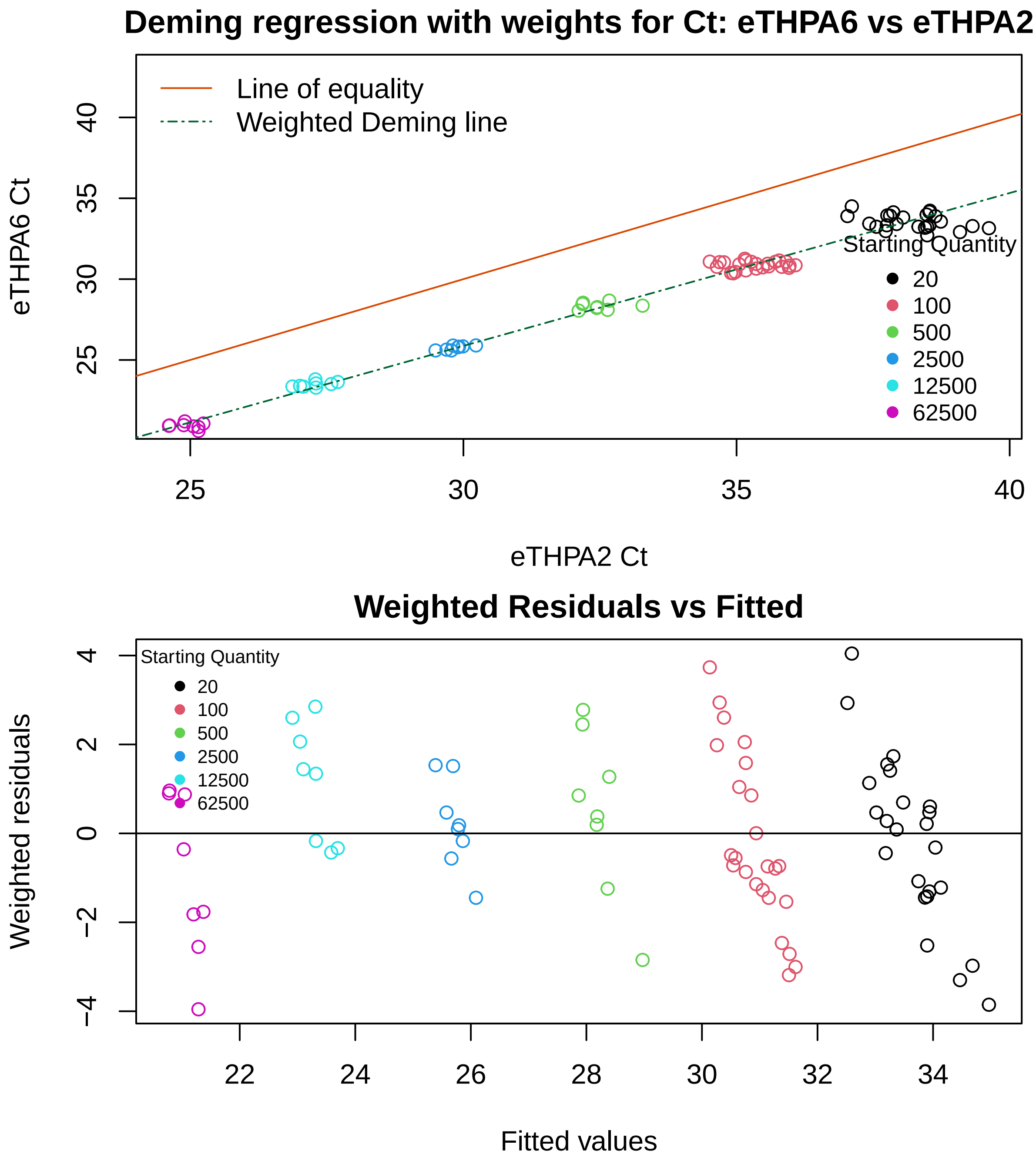


Figure 15: The estimated weighted Deming regression line overlaid with the line of equality for eTHPA6 vs eTHPA2 cycle threshold values. Bottom: the residual plot for the weighted Deming regression on eTHPA6 vs eTHPA2 cycle threshold values. This is used to check the constant variance assumption and the fit of the model.

Table 22 shows the estimated intercepts and slopes for the four models fitted to the UVic full detect gblock data. The slopes of the four models do not differ greatly from each other, however this difference is not statistically proven. The intercept from the least squares model is the only estimate that differs from the other intercepts, this slight difference is not statistically significant (Table 22). Following the recommendations from the literature and knowing our model's estimates do not differ significantly, we choose the weighted Deming model as the best model. The estimated weighted Deming line is: $\text{eTHPA6} = 0.9454 * \text{eTHPA2} \ - 2.490$. The 95% confidence interval (CI) for the slope and intercept are (0.9213,0.9695) and (-3.236,-1.744), respectively. The CI for the slope does not contain 1, indicating the slope is statistically different from 1. The CI for the intercept does not contain 0, indicating the intercept is significantly different from 0. From the estimated line, every eTHPA6 C_T is shifted by -2.333 given an eTHPA2 C_T value. The slope term in our model implies that for every increase in C_T value by 1 for eTHPA2, eTHPA6's C_T values increases by 0.9402. Comparing the slopes from the weighted Deming model (0.9402) and equation (2) (0.9426), we notice that the weighted Deming slope is slightly smaller than the equation (2) slope (Table 22 and 1). The intercepts of the two lines that differ slightly, the weighted Deming line's intercept is -2.333 and equation (2) has an intercept of -2.397. This difference in slopes and intercepts can in the future contribute to any deviations seen when comparing predicted values to "actual" values (Table 22 and 1).

Table 22: Summary of the intercepts and slopes with standard errors for the four regression techniques applied to the Ct values from the UVic full detect gblock data.

	Intercept	Intercept standard error	Slope	Slope standard error
Least squares	-1.900	0.4878	0.9271	0.0144
Weighted least squares	-2.215	0.3382	0.9366	0.0109
Deming regression	-2.173	0.4527	0.9352	0.0148
Weighted Deming regression	-2.490	0.3804	0.9454	0.0123

4.2 UVic: partial detect gblock data

Here we display the fitted least squares and Deming regression with both unweighted and weighted variations, for the $\hat{\lambda}$ values for eTHPA6 versus eTHPA2 calculated from the UVic partial detect data. The calculation methods for the weights are in section 2.4 of the methods.

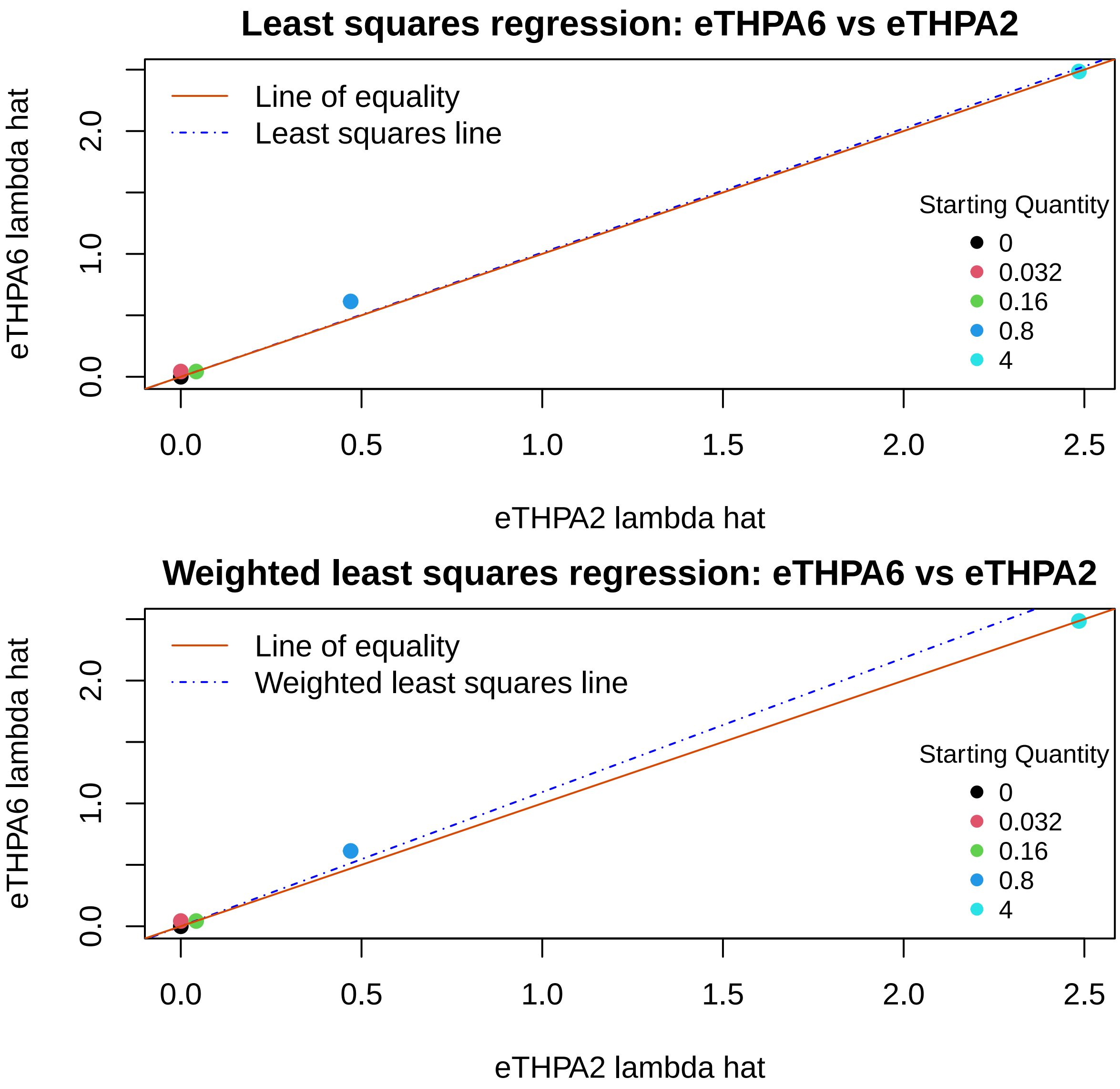


Figure 16: Top: unweighted least squares estimated line overlaid with the line of equality for eTHPA6 vs eTHPA2 on the Poisson mean scale. Bottom: weighted least squares estimated line overlaid with the line of equality for eTHPA6 vs eTHPA2 on the Poisson mean scale.

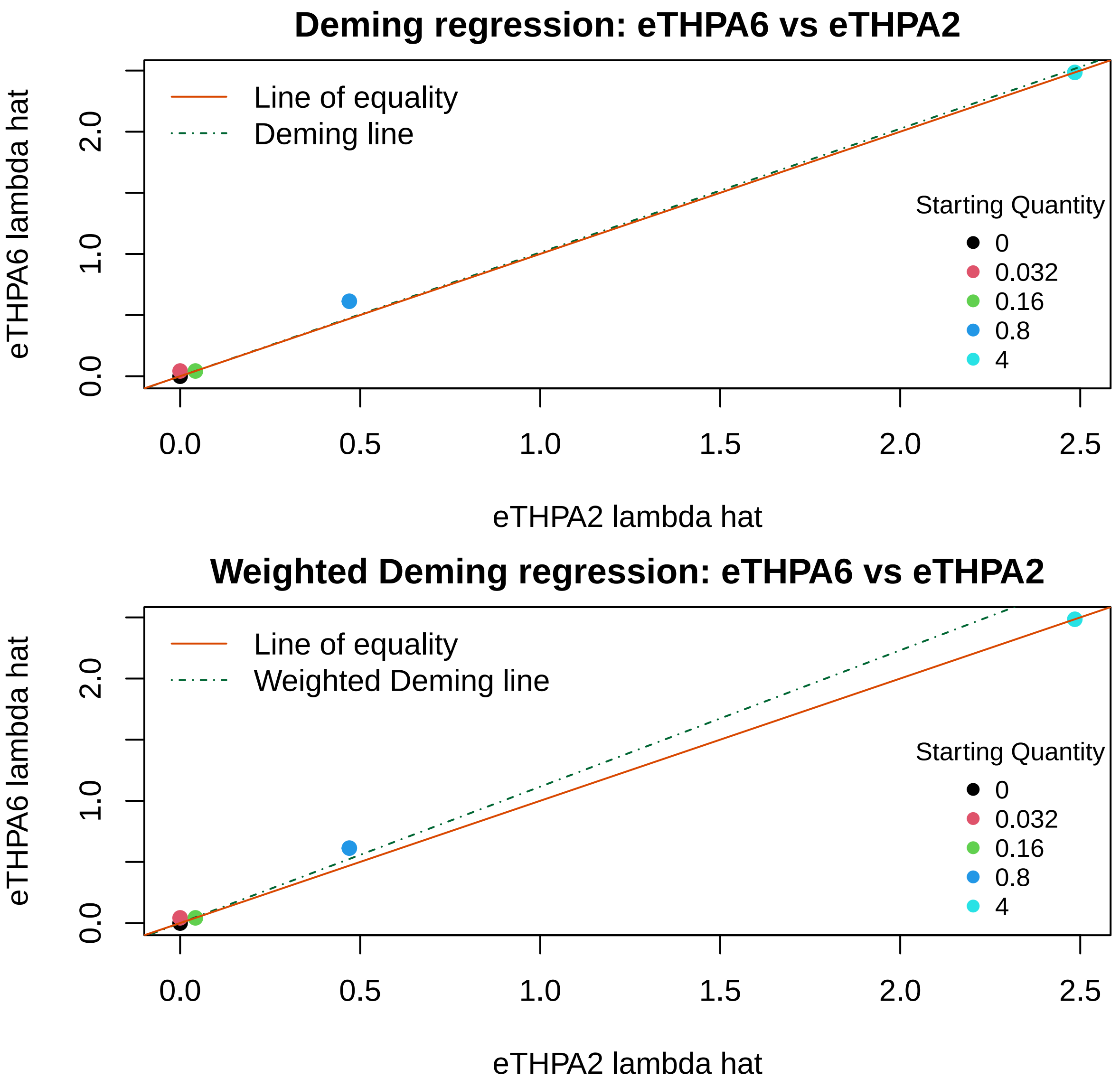


Figure 17: Top: unweighted Deming regression line overlaid with the line of equality for eTHPA6 vs eTHPA2 on the Poisson mean scale. Bottom: weighted Deming estimated line overlaid with the line of equality for eTHPA6 vs eTHPA2 on the Poisson mean scale.

The plots for the least squares and Deming model do not show a large difference between the estimated lines and the line of equality (Figure 16 and 17, top). The plots for the weighted least squares and weighted Deming model pull away from starting quantity 4 due to the addition of weights (Figure 16 and 17, bottom). The movement away from starting quantity 4 is due to the model putting less weight on the largest observation, which is the starting quantity 4 observation. The estimated slopes for the four models do not differ greatly (Table 23). Therefore, we select the weighted Deming model as the best model as per the literature. The estimated weighted

Deming line is: $\hat{\lambda_6} = 1.1156 * \hat{\lambda_2}$. The 95% CI for the slope is (0.7214,1.51), this interval contains 1 which could indicate no significant difference between the line of equality and the weighted Deming line. The slope term implies that for every increase by 1 in eTHPA2 $\hat{\lambda}$ value, eTHPA6's $\hat{\lambda}$ values increases by 1.116.

Table 23: Summary of the slopes with standard errors for the four regression techniques applied to the lambda hat values calculated from UVic partial detect gblock data.

	Slope	Slope standard error
Least squares	1.010	0.0290
Weighted least squares	1.092	0.1302
Deming regression	1.012	0.2943
Weighted Deming regression	1.116	0.2011

4.3 BVL: full detect gblock data

Least squares regression

The top plot in Figure 18 shows the estimated least squares for eTHPA6 versus eTHPA2 C_T values. The estimated line appears to fit the data adequately. The cycle threshold values decrease with increasing starting quantity (Figure 18, top). Comparing the line of equality to our fitted line, we note a moderate difference (Figure 18, top). This difference between the lines indicates the intercept could be different 0 (Figure 18, top). The slopes of the two lines appear parallel, which could indicate the slope may not be different from 1 (Figure 18, top). The residual plot of the least squares model indicates a non-constant variance implying the model may not fit well (Figure 18, bottom). This non-constant variance is confirmed by the increasing variation within the data as the starting quantity decreases. The starting quantity 62500 group is concentrated however, as the starting quantity decreases to 20, the data becomes more spread out.

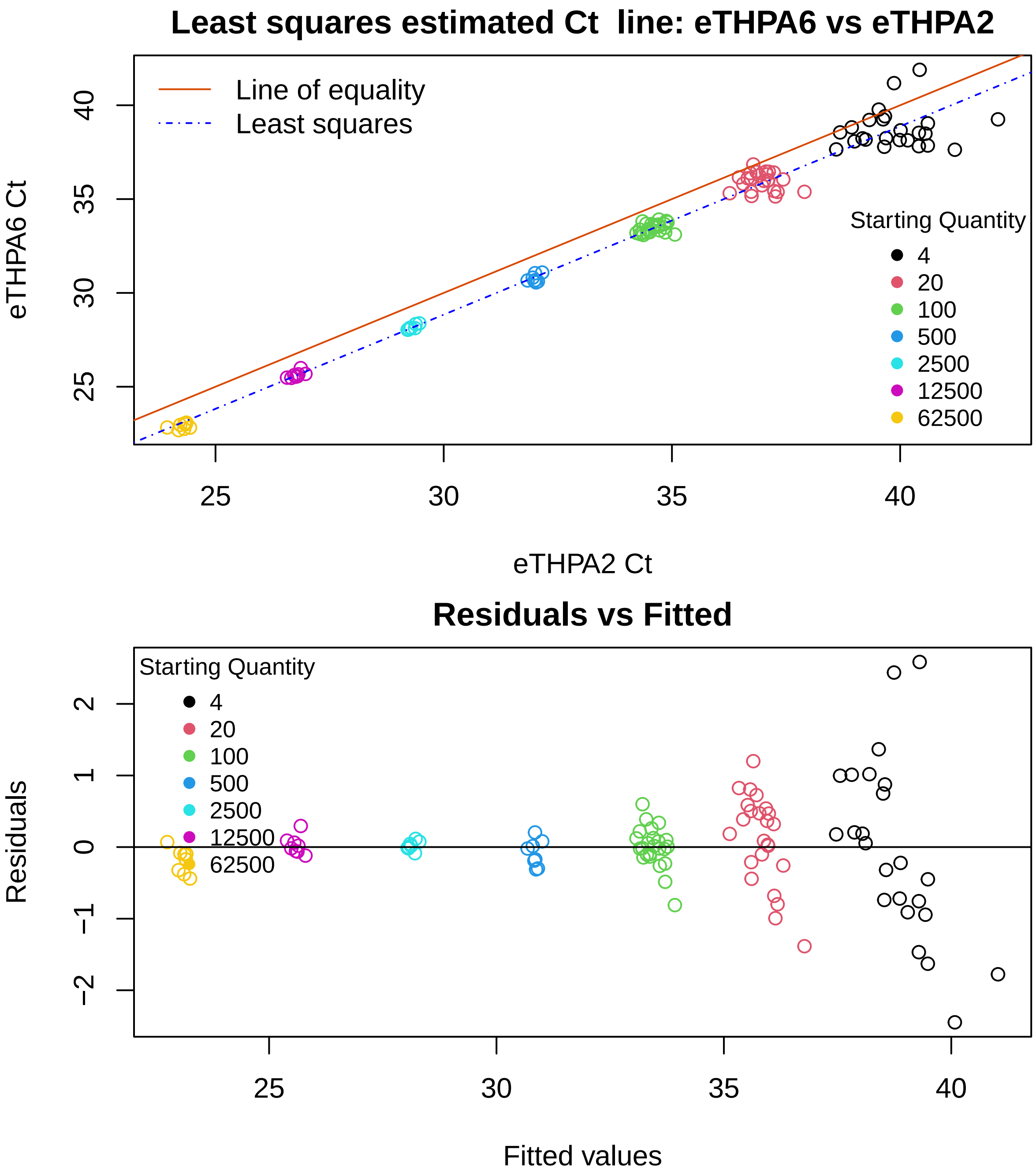


Figure 18: Top: The estimated least squares regression line overlaid with the line of equality for eTHPA6 vs eTHPA2 cycle threshold values. Bottom: the residual plot for the least squares regression on eTHPA6 vs eTHPA2 cycle threshold values. This is used to check the constant variance assumption and the fit of the model.

Weighted least squares regression

Figure 19 (top) shows the estimated weighted least squares line which appears to fit the data adequately, and does not show much difference compared to the estimated least squares (Figure 18, top). The cycle threshold values decrease with increasing starting quantity (Figure 19, top). Comparing the line of equality to our fitted line, we note a moderate difference (Figure 19, top). This difference between the lines indicates intercept could be 0 (Figure 19, top). The slopes of the two lines appear parallel, which could indicate the slope may not be different from 1 (Figure 19, top). The residual plot of the weighted least squares model indicates a constant variance and a well fitting model (Figure 19, bottom). This constant variance is confirmed by the random scatter of data points about y=0 line. The calculation methods for the weights are in section 2.4 of the methods.

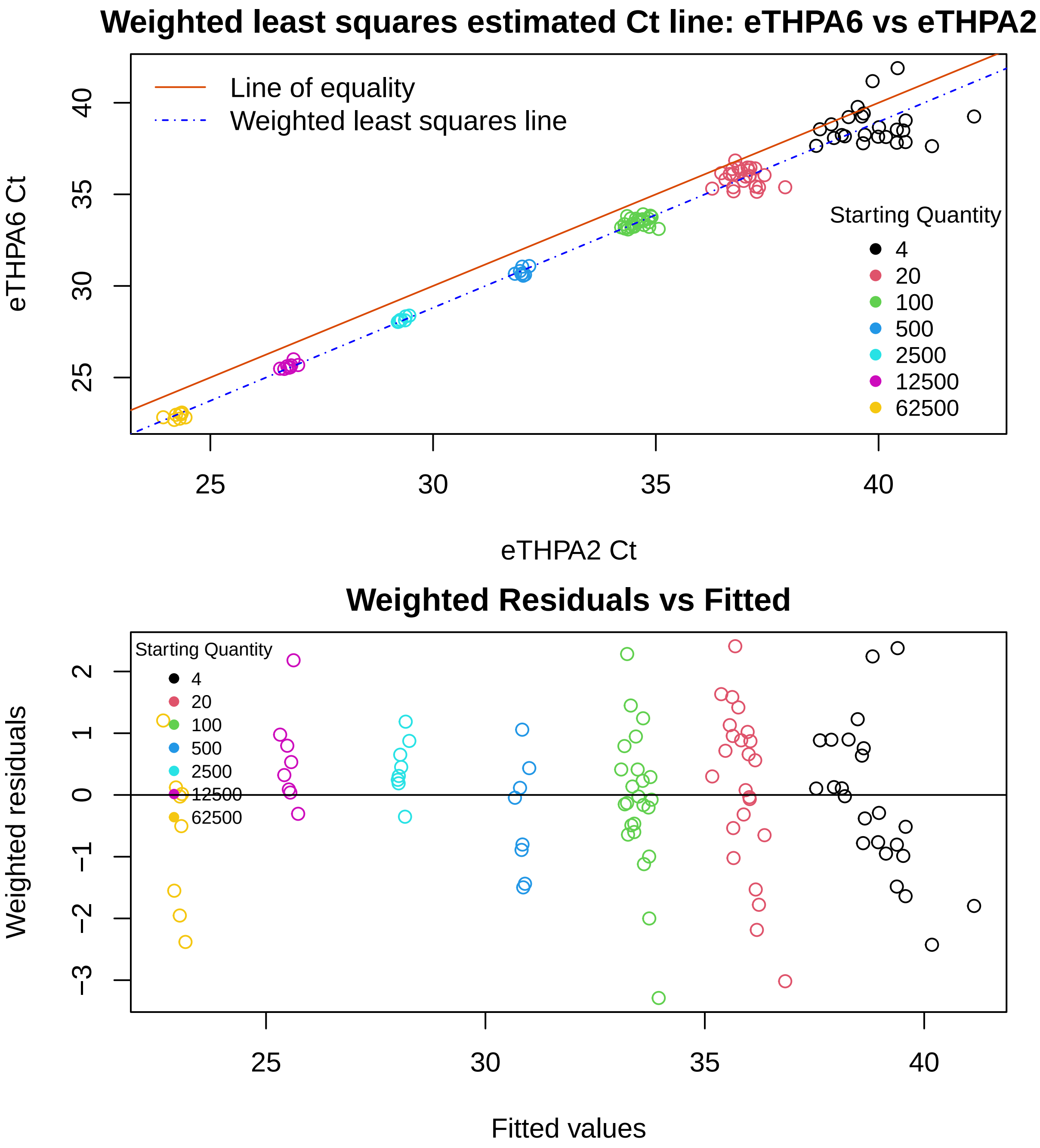


Figure 19: Top: The estimated weighted least squares regression line overlaid with the line of equality for eTHPA6 vs eTHPA2 cycle threshold values. Bottom: the residual plot for the weighted least squares regression on eTHPA6 vs eTHPA2 cycle threshold values. This is used to check the constant variance assumption and the fit of the model.

Deming regression

Figure 20 (top) shows the estimated Deming line which appears to fit the data adequately, and does not show much difference compared to the estimated unweighted and weighted least squares (Figure 18 and 19, top). The cycle threshold values decrease with increasing starting quantity (Figure 20, top). Comparing the line of equality to our fitted line, we note a moderate difference (Figure 20, top). This difference between the lines indicates the intercept could be different from 0 (Figure 20, top). The slopes of the two lines appear parallel, which could indicate the slope may not be different from 1 (Figure 20, top). The residual plot of the Deming model indicates a non-constant variance implying the model does not fit well (Figure 20, bottom). This non-constant variance is confirmed by the increasing variation within the data as the starting quantity decreases (Figure 18, bottom).

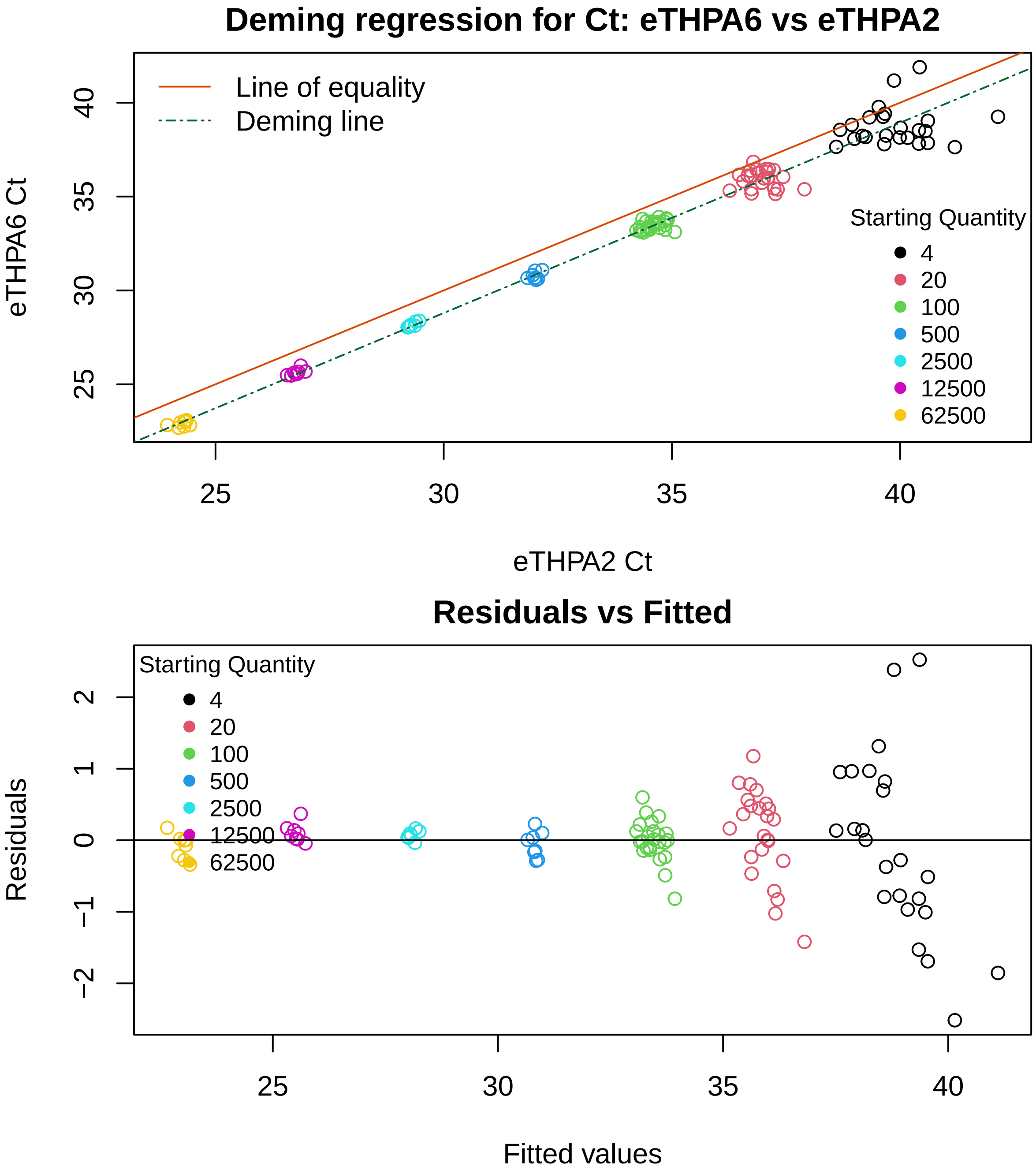


Figure 20: Top: The estimated Deming regression line overlaid with the line of equality for eTHPA6 vs eTHPA2 cycle threshold values. Bottom: the residual plot for the Deming regression on eTHPA6 vs eTHPA2 cycle threshold values. This is used to check the constant variance assumption and the fit of the model.

Weighted Deming regression

Similar to the previous fitted models, the estimated weighted Deming looks to fit the data adequately (Figure 21, top). The cycle threshold values decrease with increasing starting quantity (Figure 21, top). Comparing the line of equality to our fitted line, we note a moderate difference (Figure 21, top). This difference between the lines indicates the intercept could be different 0 (Figure 21, top). The slopes of the two lines appear parallel, which could indicate the slope may not be different from 1 (Figure 21, top). The residual plot of the weighted Deming model indicates a constant variance and that the model fits well (Figure 21, bottom). This constant variance is confirmed by random scatter of data points around the line y=0 (Figure 21, bottom). The calculation methods for the weights are in section 2.4 of the methods.

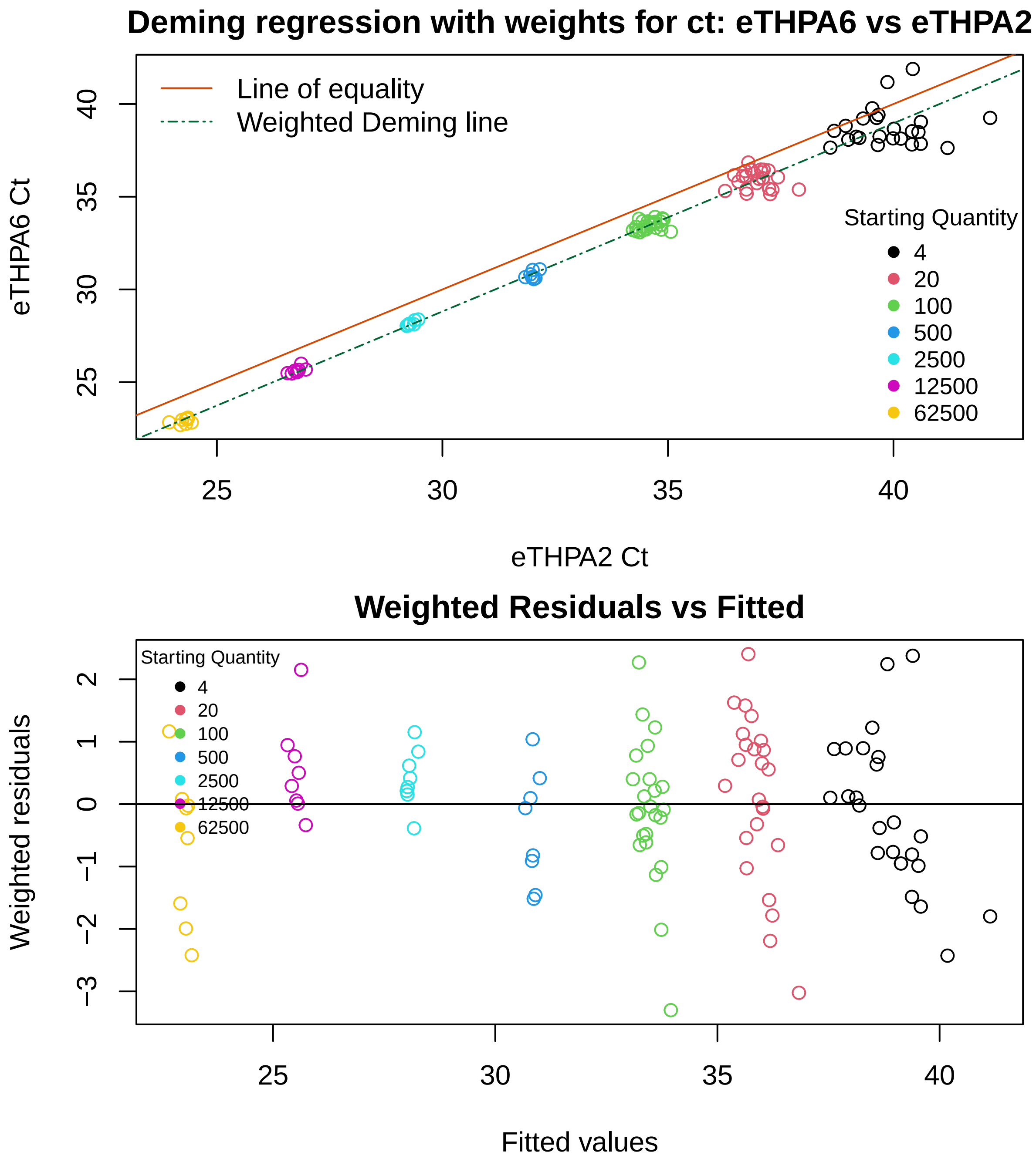


Figure 21: Top: The estimated weighted Deming regression line overlaid with the line of equality for eTHPA6 vs eTHPA2 cycle threshold values. Bottom: the residual plot for the weighted Deming regression on eTHPA6 vs eTHPA2 cycle threshold values. This is used to check the constant variance assumption and the fit of the model.

Table 24 shows the estimated intercepts and slopes for the four models fitted to the BVL full detect gblock data. The slope and intercept from least squares model is the only one that differs from the other slopes and intercepts, however the difference is not statistically significant (Table 24). Following the recommendations from the literature and knowing our model's estimates do not differ considerably, we choose the weighted Deming model as the best model. The estimated weighted Deming line is: $\text{eTHPA6} = 1.015 * \text{eTHPA2}\ - 1.640$. The 95% CI for the slope and intercept are (1.001,1.029) and (-2.052,-1.228). The intercept CI does not contain 0, indicating the intercept is statistically different from 0. The slope CI is on the borderline for including 1. From the estimated line, eTHPA6 C_T values are shifted by -1.632 given eTHPA2 C_T values. The slope term in our model implies that for every increase in C_T value by 1 for eTHPA2, eTHPA6's C_T values increases by 1.015. Comparing the slopes from the weighted Deming model (1.015) and equation (2) (1.0124), we notice that the slope from equation (2) is slightly smaller than the weighted Deming slope (Table 24 and 1). However it is the intercepts of the two lines that differ greatly, the weighted Deming line's intercept is -1.632 and equation (2) has an intercept of -1.571. This difference in slopes and intercepts can contribute to deviations in predicted and "actual" values (Table 24 and 1).

Table 24: Summary of the intercepts and slopes with standard errors for the four regression techniques applied to the Ct values from the BVL full detect gblock data.

	Intercept	Intercept standard error	Slope	Slope standard error
Least squares	-1.268	0.4857	1.004	0.0140
Weighted least squares	-1.652	0.1933	1.015	0.0064
Deming regression	-1.613	0.5048	1.014	0.0165
Weighted Deming regression	-1.640	0.2104	1.015	0.0070

4.4 BVL: partial detect gblock data

Here we display the fitted least square and Deming regression with both unweighted and weighted variations, for the $\hat{\lambda}$ values for eTHPA6 versus eTHPA2 calculated from the BVL partial detect data. The calculation methods for the weights are in section 2.4 of the methods.

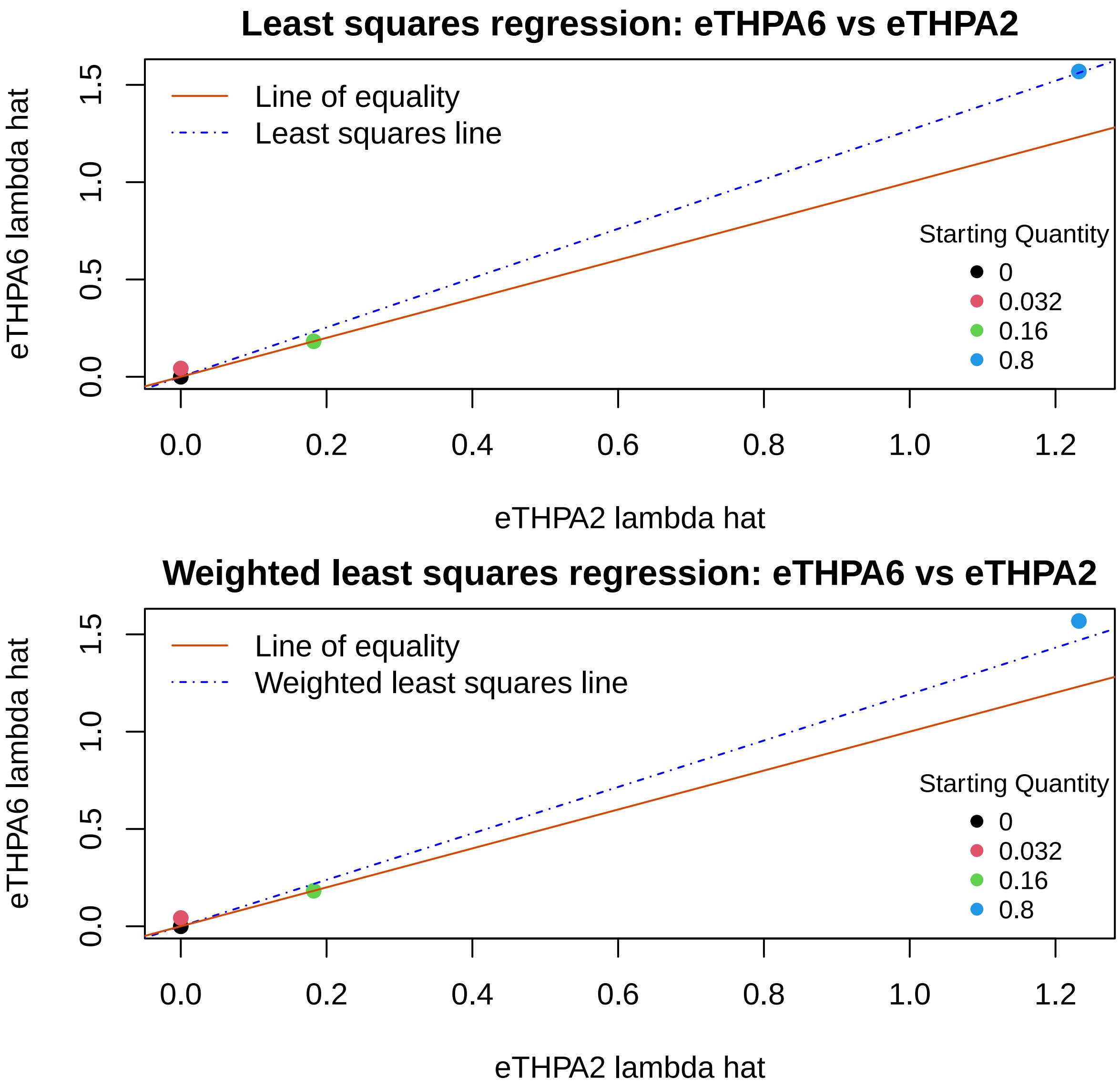


Figure 22: Top: unweighted least squares estimated line overlaid with the line of equality for eTHPA6 vs eTHPA2 on the Poisson mean scale. Bottom: weighted least squares estimated line overlaid with the line of equality for eTHPA6 vs eTHPA2 on the Poisson mean scale.

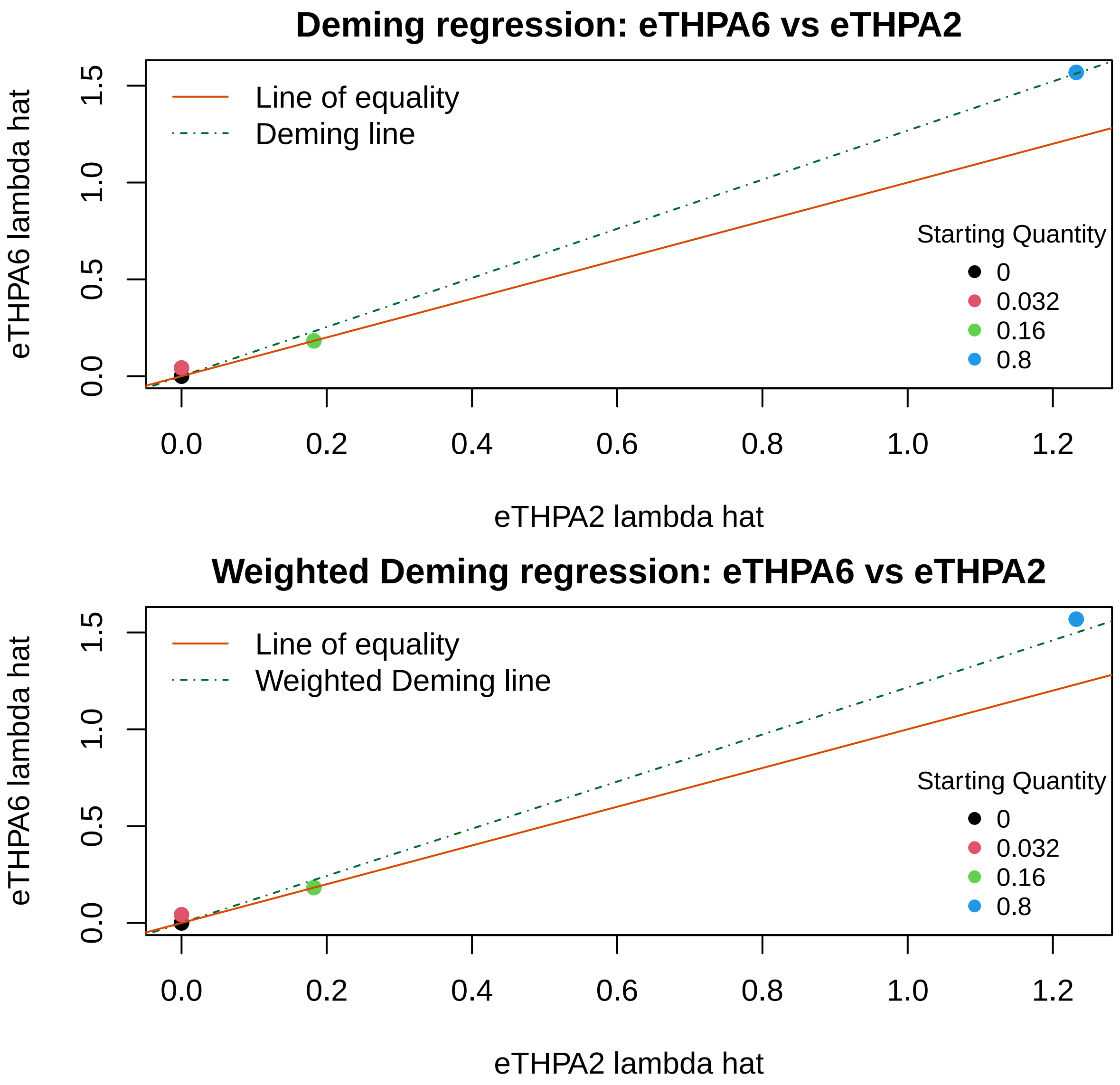


Figure 23: Top: unweighted Deming regression line overlaid with the line of equality for eTHPA6 vs eTHPA2 on the Poisson mean scale. Bottom: weighted Deming estimated line overlaid with the line of equality for eTHPA6 vs eTHPA2 on the Poisson mean scale.

The plots for the least squares and Deming model do not show a large difference between the estimated lines (Figure 22 and 23, top). The plots for the weighted least squares and weighted Deming model pull away from starting quantity 0.8 (Figure 22 and 23). The movement away from starting quantity 0.8 is due to the model putting less weight on the largest observation, which is the starting quantity 0.8 observation. All the estimated lines do not appear parallel to the line of equality. The estimated slope for the weighted least squares regression differs slightly from the other three models, and the other three model slopes are not vastly different,

Table 25: Summary of the slopes with standard errors for the four regression technique applied to the lambda hat values calculated from the BVL partial detect data.

	Slope	Slope standard error
Least squares	1.267	0.0302
Weighted least squares	1.193	0.1723
Deming regression	1.269	0.2410
Weighted Deming regression	1.216	0.2144

not statistically proven (Table 25). Therefore, we select the weighted Deming model as the best model as per the literature. The estimated weighted Deming line is: $\hat{\lambda_6} = 1.2165 * \hat{\lambda_2}$. The 95% CI for the slope is (0.7962,1.637), this interval contains 1 which could indicate no significant difference between the line of equality and the weighted Deming line. The slope term implies that for every increase by 1 in eTHPA2 $\hat{\lambda}$ value, eTHPA6's $\hat{\lambda}$ values increase by 1.216.

5 Field Data Predictions

To determine how well our chosen weighted Deming prediction models predict eTHPA6 from eTHPA2, we apply the models to field data. The prediction models/curves for the field data do not predict copy numbers. The prediction curves predict eTHPA6 values given eTHPA2 values, where the input and output is C_T values for the full detect data and $\hat{\lambda}$ values for the partial detect data. The standard curves estimate (natural log) copy number values from C_T and $\hat{\lambda}$ values. The standard curve calculations for the full partial detect data sets are in section 2.2 of the methods. The standard curves for estimating (natural log) copy numbers for BVL and UVic's full and partial gblock data are as follows:

UVic: full detect data

eTHPA2: $C_{T_2} = -1.654 * ln(c\hat{n}_2) + 42.962$ with inverse $ln(c\hat{n}_2) = \frac{C_{T_2}-42.962}{-1.654}$, where the standard errors for the slope and intercept are 0.0318 and 0.2397, respectively.

eTHPA6: $C_{T_6} = -1.559 * ln(c\hat{n}_6) + 38.097$ with inverse $ln(c\hat{n}_6) = \frac{C_{T_6}-38.097}{-1.559}$ where the standard errors for the slope and intercept are 0.0181 and 0.1364, respectively.

UVic: partial detect data

eTHPA2: $-ln(1-\hat{p_2}) = \hat{\lambda_2} = .566 * c\hat{n}_2$ with inverse $c\hat{n}_2 = \dfrac{\hat{\lambda_2}}{0.566}$, where $\hat{p_2}$ is the proportion of detects and the standard error for $\hat{\beta} = 0.566$ is 0.115 ,

eTHPA6: $-ln(1-\hat{p_6}) = \hat{\lambda_6} = .652 * c\hat{n}_2$ with inverse $c\hat{n}_6 = \dfrac{\hat{\lambda_6}}{0.652}$, where $\hat{p_6}$ is the proportion of detects and the standard error for $\hat{\beta} = 0.652$ is 0.129

BVL: full detect data

The standard curves for eTHPA2 and eTHPA6 from the full detect BVL gblock data are:

eTHPA2: $C_{T_2} = -1.6105 * ln(c\hat{n}_2) + 41.9815$ with inverse $ln(c\hat{n}_2) = \frac{C_{T_2}-41.9815}{-1.6105}$, where the standard errors for the slope and intercept are 0.0139 and 0.0971, respectively.

eTHPA6: $C_{T_6} = -1.6304 * ln(c\hat{n}_6) + 40.9296$ with inverse $ln(c\hat{n}_6) = \frac{C_{T_6}-40.9296}{-1.6304}$, where the standard errors for the slope and intercept are 0.0081 and 0.0570, respectively.

BVL: partial detect data

eTHPA2: $-ln(1-\hat{p_2}) = \hat{\lambda_2} = 1.358 * \hat{cn_2}$ with inverse $\hat{cn_2} = \dfrac{\hat{\lambda}}{1.358}$, where $\hat{p_2}$ is the proportion of detects and the standard error for $\hat{\beta} = 1.358$ is 0.308,

eTHPA6: $-ln(1-\hat{p_6}) = \hat{\lambda_6} = 1.691 * \hat{cn_2}$ with inverse $\hat{cn_6} = \dfrac{\hat{\lambda}}{1.691}$, where $\hat{p_6}$ is the proportion of detects and the standard error for $\hat{\beta} = 1.691$ is 0.366.

From section 4, the prediction curves for C_T and $\hat{\lambda}$ values from eTHPA6 given eTHPA2 from the gblock data models are as follows:

Full detect data (C_T scale):

UVic: $\hat{C_{T_6}} = 0.9454 * C_{T_2} - 2.490$ with the standard error for the slope = 0.0123 and intercept = 0.3804,

BVL: $\hat{C_{T_6}} = 1.015 * C_{T_2} - 1.640$ with the standard error for the slope = 0.0070 and intercept = 0.2104,.

Partial detect data ($\hat{\lambda}$ scale):

UVic: $\hat{\lambda}_6 = 1.1156 * \hat{\lambda}_2$ with the standard error for the slope = 0.2011,

BVL: $\hat{\lambda}_6 = 1.2165 * \hat{\lambda}_2$ with the standard error for the slope = 0.2144.

5.1 UVic: full detect field data

Figure 24 (top) shows the predicted (weighted Deming) versus “actual” natural log copy number values for the full detect Kemano field data from the UVic lab. The majority of data points on the natural log scale appear overestimated, as they fall above the line of equality (Figure 24, top). Figure 24 (bottom) displays the predicted versus “actual” natural log copy number sample means, with 95% CI. We notice approximately six points have their 95% CI overlapping with the line of equality indicating that they have been well predicted (24, bottom). Since 6 of

11 points are well predicted the weighted Deming model is an adequate choice for prediction.

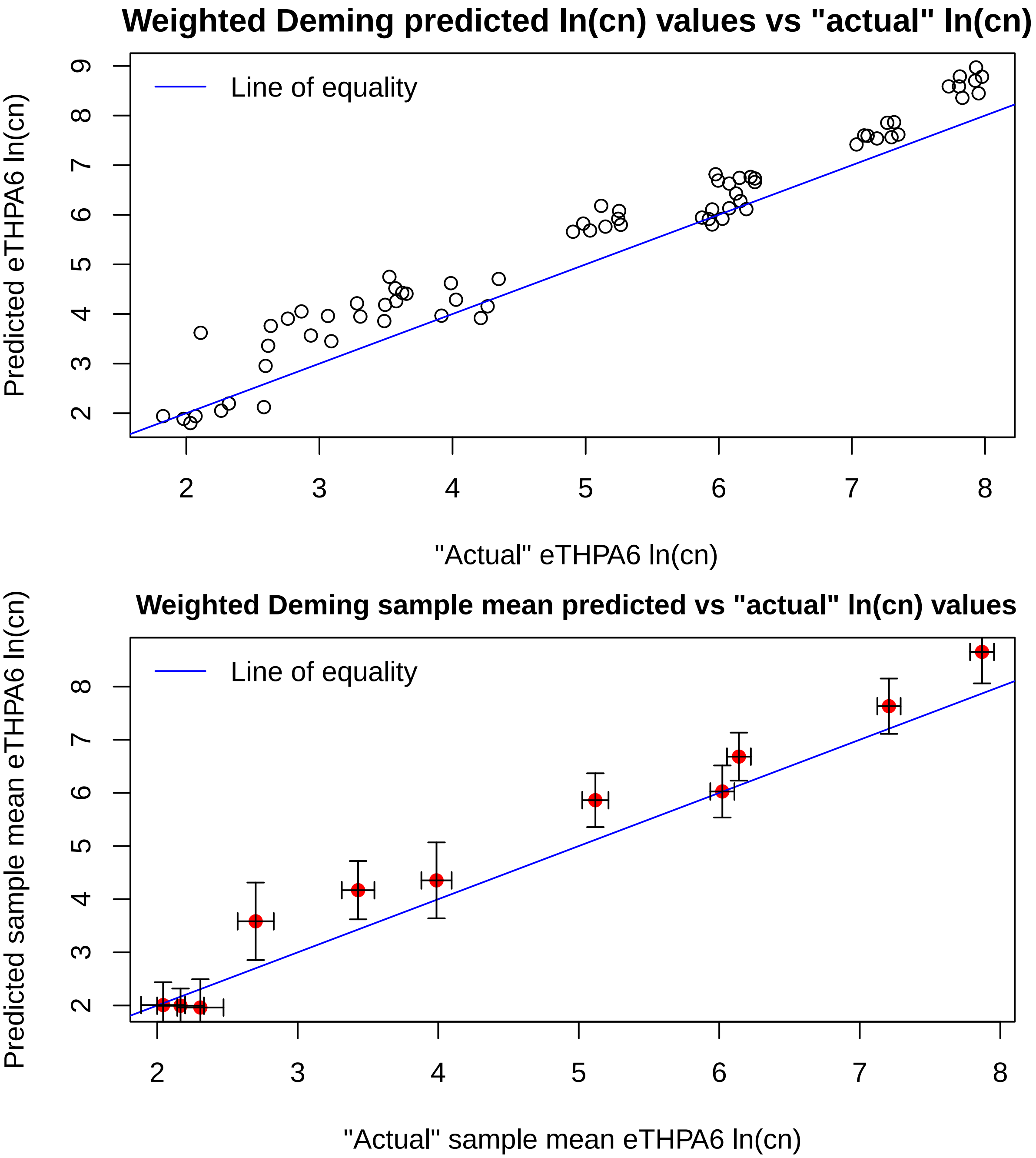


Figure 24: Top: the predicted ln(cn) values from the weighted Deming prediction line plotted against the “actual”, overlaid with the line of equality. Bottom: the predicted sample mean of the ln(cn) values from the weighted Deming prediction line plotted against the “actual” ln(cn) values with 95 percent CIs, overlaid with the line of equality. Both plots use the UVic full detect field data.

Tables 26 and 27 present the “actual” and predicted C_T and natural log copy number values for the full detect Kemano Field data. Comparing the sample means of the “actual” and predicted C_T values, we are underpredicting the majority of predicted C_T values (Table 26). Comparing the sample means of the “actual” and predicted natural log copy number values, we are overpredicting the majority of predicted natural log copy number value (Table 27). This is due to the predicted C_T values being less than the intercept, the larger the gap between the predicted C_T value and the intercept the more the natural log copy number value is overestimated. The intercept of -2.3974 from section 2.2 is similar to the intercept of -1.956 from the weighted Deming regression, but different enough to create deviations between the predicted and “actual” values.

Table 26: Summary statistics for predicted and "actual" eTHPA6 ln(Ct) values by sample. N is the total number of acceptable technical replicates in a sample. Estimates for the minimum, maximum, mean, median, sample variance and standard deviation by sample are shown.

Sample	N	Min	Max	Mean	Median	Sample variance	Sample standard deviation
eTHPA6: "actual" Ct							
1	8	32.40	33.32	32.75	32.66	0.094	0.306
2	8	33.28	34.81	33.89	33.89	0.212	0.461
3	8	31.32	32.60	31.88	31.85	0.211	0.460
4	2	34.07	34.93	34.50	34.50	0.370	0.608
5	8	28.42	28.94	28.71	28.76	0.034	0.185
6	2	34.57	34.87	34.72	34.72	0.045	0.212
7	8	25.66	26.05	25.83	25.82	0.019	0.138
8	8	28.32	28.78	28.52	28.52	0.033	0.183
9	8	26.64	27.13	26.86	26.83	0.033	0.182
10	8	29.89	30.45	30.12	30.09	0.044	0.210
11	3	34.48	35.25	34.91	35.01	0.155	0.394
eTHPA6: predicted Ct							
1	8	31.05	32.08	31.60	31.55	0.132	0.363
2	8	31.78	33.49	32.51	32.49	0.285	0.534
3	8	30.69	31.99	31.31	31.30	0.257	0.507
4	2	34.79	35.29	35.04	35.04	0.126	0.354
5	8	28.31	29.05	28.70	28.70	0.059	0.242
6	2	34.90	35.07	34.99	34.99	0.014	0.120
7	8	24.11	25.07	24.61	24.62	0.096	0.309
8	8	27.47	28.07	27.68	27.64	0.034	0.185
9	8	25.83	26.53	26.20	26.25	0.058	0.240
10	8	28.46	29.27	28.96	29.04	0.083	0.289
11	3	34.67	35.16	34.97	35.07	0.066	0.257

Table 27: Summary statistics for predicted and "actual" eTHPA6 ln(cn) values by sample. N is the total number of acceptable technical replicates in a sample. Estimates for the minimum, maximum, mean, median, sample variance and standard deviation by sample are shown.

Sample	N	Min	Max	Mean	Median	Sample variance	Sample standard deviation
eTHPA6: "actual" ln(cn)							
1	8	3.064	3.654	3.430	3.491	0.039	0.196
2	8	2.108	3.090	2.701	2.699	0.087	0.295
3	8	3.526	4.347	3.988	4.007	0.087	0.295
4	2	2.031	2.583	2.307	2.307	0.152	0.390
5	8	5.874	6.207	6.022	5.989	0.014	0.119
6	2	2.070	2.262	2.166	2.166	0.019	0.136
7	8	7.727	7.978	7.870	7.878	0.008	0.089
8	8	5.976	6.271	6.140	6.143	0.014	0.117
9	8	7.035	7.349	7.208	7.227	0.014	0.116
10	8	4.905	5.264	5.118	5.133	0.018	0.135
11	3	1.826	2.320	2.042	1.980	0.064	0.253
eTHPA6: predicted ln(cn)							
1	8	3.857	4.518	4.169	4.200	0.054	0.233
2	8	2.954	4.051	3.584	3.594	0.117	0.342
3	8	3.918	4.749	4.354	4.358	0.106	0.325
4	2	1.802	2.123	1.962	1.962	0.052	0.227
5	8	5.804	6.277	6.026	6.025	0.024	0.155
6	2	1.941	2.050	1.996	1.996	0.006	0.077
7	8	8.357	8.969	8.653	8.645	0.039	0.198
8	8	6.429	6.817	6.683	6.711	0.014	0.119
9	8	7.417	7.866	7.631	7.596	0.024	0.154
10	8	5.658	6.180	5.862	5.810	0.034	0.185
11	3	1.887	2.196	2.008	1.941	0.027	0.165

5.2 UVic: partial detect field data

Figure 25 shows the predicted (weighted Deming line) values versus the “actual” copy number values. From Figure 25 it can be seen that the majority of the data points fall below the line of equality, however all the 95% CIs for the points overlapping with the line of equality. Since all 95% CI’s overlap with the line of equality, the weighted Deming line is a good choice for the prediction model. Although there are six observation in the 2023 data set, two of them are the same indicated by the opaque point labeled 2. One point between the 2021 and 2023 data set is shared, shown by a purple point with a red diamond shape around it.

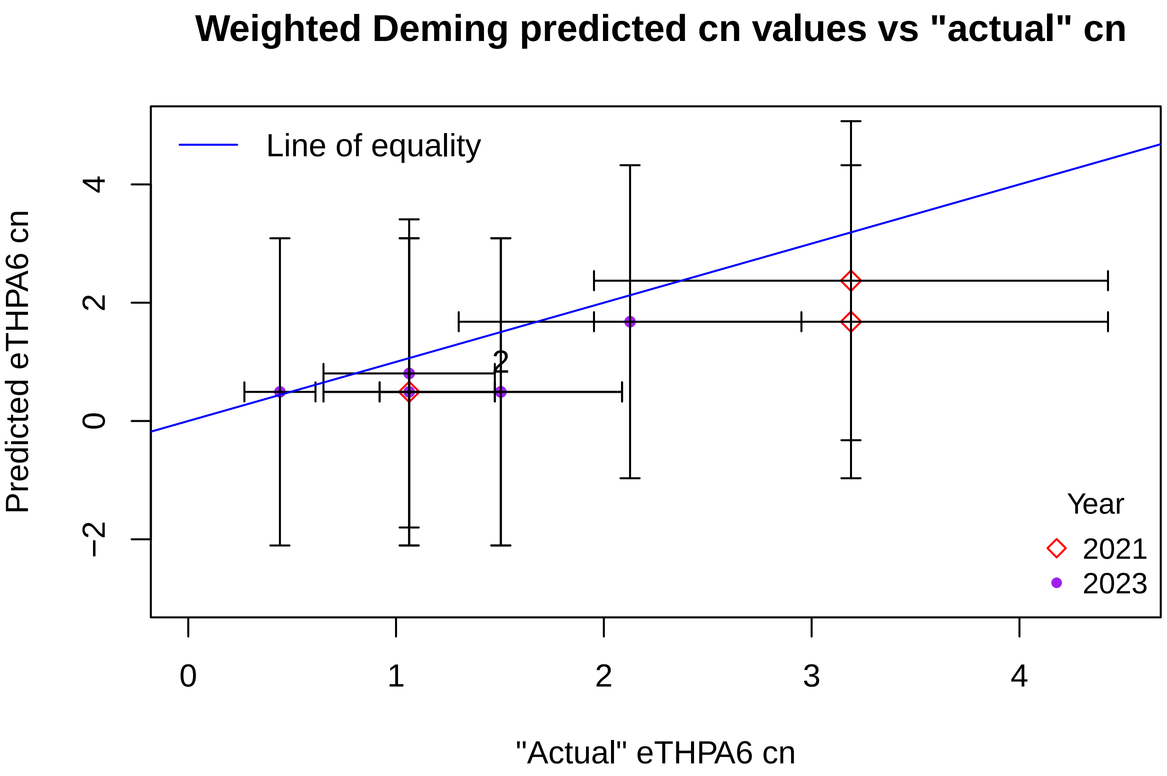


Figure 25: The predicted copy number values from the weighted Deming line plotted against the “actual” copy number values with 95 percent CIs, overlaid with the line of equality. All calculations are done using the UVic partial detect field data.

Table 28 presents the predicted and “actual” $\hat{\lambda}$ and copy number values for eTHPA6. Both the $\hat{\lambda}$ and copy number values display a large difference between the “actual” and predicted values (Table 28). The predicted standard error’s for both the predicted $\hat{\lambda}$ and copy number

Table 28: The predicted and "actual" eTHPA6 lambda hat and copy number values for the UVic partial detect Kemano field data (2021/2023).

Sample	"Actual"	Predicted	Year
Lambda hat values			
1	2.0794	1.5465	2021
2	2.0794	1.0942	2021
3	0.6931	0.3209	2021
1	0.9808	0.3209	2023
2	0.9808	0.3209	2023
3	0.6931	0.5243	2023
4	0.2877	0.3209	2023
5	0.6931	0.3209	2023
6	1.3863	1.0942	2023
Copy number values			
1	3.1893	2.3720	2021
2	3.1893	1.6782	2021
3	1.0631	0.4922	2021
1	1.5043	0.4922	2023
2	1.5043	0.4922	2023
3	1.0631	0.8042	2023
4	0.4412	0.4922	2023
5	1.0631	0.4922	2023
6	2.1262	1.6782	2023

values can be found in section 8.2 and 8.3 of the appendix.

5.3 BVL: full detect field data

Figure 26 (top) shows the predicted (weighted Deming line) values versus “actual” ln(cn) values for the full detect Fraser river field data from the BVL lab. The majority of data points on the natural log scale appear underestimated, as they fall below the line of equality (Figure 24, top). Figure 26 (bottom) displays the predicted versus “actual” natural log copy number sample means, with 95% CIs. We notice many points have their 95% CI overlapping with the line of equality (26, bottom). Therefore, the weighted Deming line is is not majority underestimating the predicted values, but instead making adequate predictions for the majority of the data points.

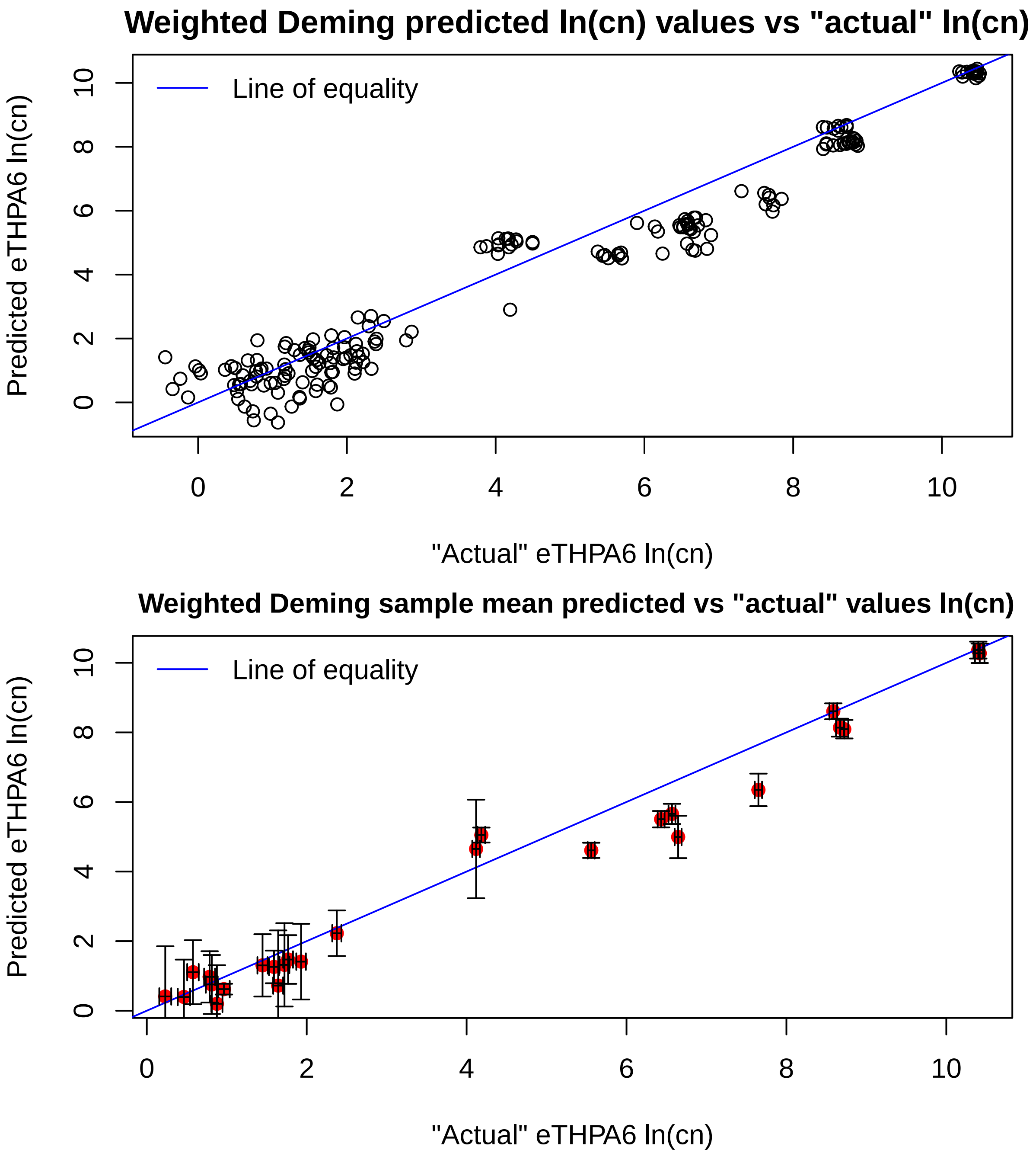


Figure 26: Top: the predicted ln(cn) values from the weighted Deming prediction line plotted against the "actual" ln(cn) values, overlaid with the line of equality. Bottom: the predicted sample mean of the ln(cn) values from the weighted Deming prediction line plotted against the "actual" ln(cn) values with 95 percent CIs, overlaid with the line of equality. Both plots use the BVL full detect field data.

Tables 29 and 30 present the “actual” and predicted C_T and natural log copy number values for the full detect Fraser river field data. Comparing the means of the “actual” and predicted C_T values, we observe that the predicted C_T are majority overestimated (Table 29). Comparing the means of the “actual” and predicted natural log copy number values, we notice the predicted natural log copy number values are mostly underestimated and some are negative (Table 30). This is due the predicted C_T value being close to or more than intercept. The intercept of -1.5706 from section 2.2 is similar to the intercept of -1.445 from the weighted Deming regression, but different enough to create deviations between the predicted and “actual” values.

Table 29: Summary statistics for predicted and "actual" eTHPA6 cycle threshold values by sample. N is the total number of acceptable technical replicates in a sample. Estimates for the minimum, maximum, mean, median, smaple variance and standard deviation by sample are shown.

Sample	N	Min	Max	Mean	Median	Sample variance	Sample standard deviation
eTHPA6: "actual" Ct							
1	8	33.60	34.35	34.11	34.16	0.069	0.263
2	8	33.60	34.74	34.22	34.13	0.126	0.356
3	8	28.14	29.02	28.46	28.41	0.066	0.256
4	8	29.80	30.92	30.22	30.20	0.106	0.325
5	8	29.69	30.75	30.09	30.08	0.102	0.320
6	8	26.54	27.16	26.80	26.75	0.063	0.252
7	8	26.47	27.23	26.71	26.62	0.074	0.272
8	8	29.97	31.31	30.44	30.33	0.188	0.433
9	8	26.71	27.23	26.93	26.90	0.036	0.189
10	8	31.64	32.17	31.87	31.83	0.042	0.205
11	8	23.80	24.18	23.94	23.90	0.023	0.150
12	8	23.86	24.25	23.97	23.95	0.017	0.132
13	8	37.31	39.64	38.34	38.30	0.577	0.759
14	8	37.06	38.49	37.78	37.71	0.222	0.471
15	8	36.86	39.01	38.12	38.29	0.668	0.817
16	8	37.32	40.34	38.57	38.45	0.749	0.866
17	8	37.47	39.73	38.25	37.97	0.564	0.751
18	6	39.71	41.65	40.55	40.45	0.626	0.791
19	6	39.43	40.99	39.99	39.84	0.321	0.566
20	8	38.82	40.87	39.65	39.58	0.417	0.645
21	8	37.41	39.03	38.05	38.01	0.294	0.542
22	8	36.25	38.01	37.05	37.09	0.314	0.560
23	6	38.88	40.14	39.50	39.38	0.261	0.511
24	4	39.34	41.32	40.17	40.02	0.689	0.830
25	4	38.64	39.79	39.36	39.50	0.293	0.541
26	4	38.35	41.49	39.60	39.29	1.817	1.348
eTHPA6: predicted Ct							
1	8	32.55	32.91	32.70	32.67	0.019	0.137
2	8	32.57	36.20	33.35	32.99	1.375	1.173
3	8	30.15	31.20	30.58	30.52	0.127	0.356
4	8	31.50	31.99	31.70	31.65	0.037	0.193
5	8	32.09	33.34	32.79	32.96	0.235	0.485
6	8	27.44	27.80	27.66	27.71	0.017	0.132
7	8	27.58	28.00	27.74	27.70	0.020	0.142
8	8	31.77	32.21	31.95	31.94	0.018	0.135
9	8	26.78	27.06	26.90	26.89	0.008	0.087
10	8	33.23	33.58	33.41	33.42	0.015	0.123
11	8	24.04	24.39	24.18	24.15	0.016	0.128
12	8	23.91	24.10	24.03	24.04	0.003	0.056
13	8	38.38	39.57	38.88	38.82	0.141	0.376
14	8	37.60	40.03	38.63	38.40	0.808	0.899
15	8	36.78	40.08	38.78	38.97	0.981	0.991
16	8	37.90	40.17	38.80	38.59	0.544	0.738
17	8	38.16	41.39	39.75	39.78	1.728	1.314
18	6	38.62	41.84	40.26	40.34	1.430	1.196
19	6	37.76	40.01	39.13	39.15	0.575	0.758
20	8	38.26	40.07	39.34	39.39	0.365	0.605
21	8	37.71	39.38	38.53	38.46	0.327	0.572
22	8	36.52	37.95	37.30	37.42	0.285	0.534
23	6	39.30	41.95	40.60	40.59	0.838	0.916
24	4	39.55	41.51	40.28	40.04	0.787	0.887
25	4	39.83	40.01	39.92	39.92	0.005	0.072
26	4	39.00	40.35	39.70	39.73	0.487	0.698

Table 30: Summary statistics for predicted and "actual" eTHPA6 ln(cn) values by sample. N is the total number of acceptable technical replicates in a sample. Estimates for the minimum, maximum, mean, median, sample variance and standard deviation for sample are shown.

Sample	N	Min	Max	Mean	Median	Sample variance	Sample standard deviation
eThPA6: "actual" ln(cn)							
1	8	4.036	4.496	4.184	4.155	0.026	0.161
2	8	3.796	4.496	4.118	4.171	0.048	0.218
3	8	7.305	7.845	7.650	7.676	0.025	0.157
4	8	6.140	6.826	6.569	6.578	0.040	0.199
5	8	6.244	6.896	6.646	6.654	0.039	0.196
6	8	8.446	8.826	8.670	8.697	0.024	0.154
7	8	8.403	8.869	8.724	8.777	0.028	0.167
8	8	5.900	6.720	6.433	6.501	0.071	0.266
9	8	8.400	8.719	8.587	8.605	0.013	0.116
10	8	5.373	5.698	5.559	5.581	0.016	0.126
11	8	10.271	10.508	10.418	10.447	0.008	0.092
12	8	10.233	10.473	10.400	10.414	0.007	0.081
13	8	0.791	2.220	1.591	1.613	0.217	0.466
14	8	1.496	2.373	1.930	1.975	0.083	0.289
15	8	1.177	2.496	1.722	1.619	0.251	0.501
16	8	0.362	2.214	1.447	1.524	0.282	0.531
17	8	0.736	2.122	1.643	1.818	0.212	0.461
18	6	-0.442	0.748	0.231	0.294	0.235	0.485
19	6	-0.037	0.920	0.577	0.668	0.121	0.347
20	8	0.037	1.294	0.786	0.831	0.157	0.396
21	8	1.165	2.159	1.767	1.794	0.111	0.333
22	8	1.791	2.870	2.377	2.358	0.118	0.344
23	6	0.484	1.257	0.877	0.954	0.098	0.314
24	4	-0.239	0.975	0.464	0.561	0.259	0.509
25	4	0.699	1.404	0.964	0.877	0.110	0.332
26	4	-0.344	1.582	0.812	1.006	0.684	0.827
eTHPA6: predicted ln(cn)							
1	8	4.919	5.136	5.049	5.068	0.007	0.084
2	8	2.901	5.130	4.650	4.872	0.517	0.719
3	8	5.971	6.612	6.347	6.388	0.048	0.218
4	8	5.485	5.784	5.658	5.691	0.014	0.118
5	8	4.657	5.423	4.995	4.887	0.089	0.298
6	8	8.056	8.274	8.136	8.106	0.007	0.081
7	8	7.932	8.187	8.092	8.112	0.008	0.087
8	8	5.348	5.616	5.505	5.516	0.007	0.083
9	8	8.505	8.679	8.607	8.613	0.003	0.053
10	8	4.508	4.719	4.610	4.604	0.006	0.076
11	8	10.148	10.360	10.274	10.291	0.006	0.079
12	8	10.322	10.441	10.366	10.360	0.001	0.034
13	8	0.835	1.563	1.259	1.295	0.053	0.231
14	8	0.554	2.042	1.410	1.554	0.304	0.551
15	8	0.523	2.547	1.319	1.205	0.369	0.608
16	8	0.467	1.856	1.303	1.435	0.205	0.452
17	8	-0.280	1.700	0.722	0.707	0.650	0.806
18	6	-0.560	1.414	0.410	0.364	0.538	0.733
19	6	0.567	1.943	1.106	1.093	0.216	0.465
20	8	0.529	1.638	0.973	0.943	0.137	0.371
21	8	0.953	1.974	1.472	1.516	0.123	0.351
22	8	1.824	2.702	2.227	2.154	0.107	0.328
23	6	-0.629	1.003	0.199	0.206	0.315	0.562
24	4	-0.355	0.847	0.396	0.545	0.296	0.544
25	4	0.567	0.673	0.620	0.620	0.002	0.044
26	4	0.355	1.183	0.754	0.738	0.183	0.428

5.4 BVL: partial detect field data

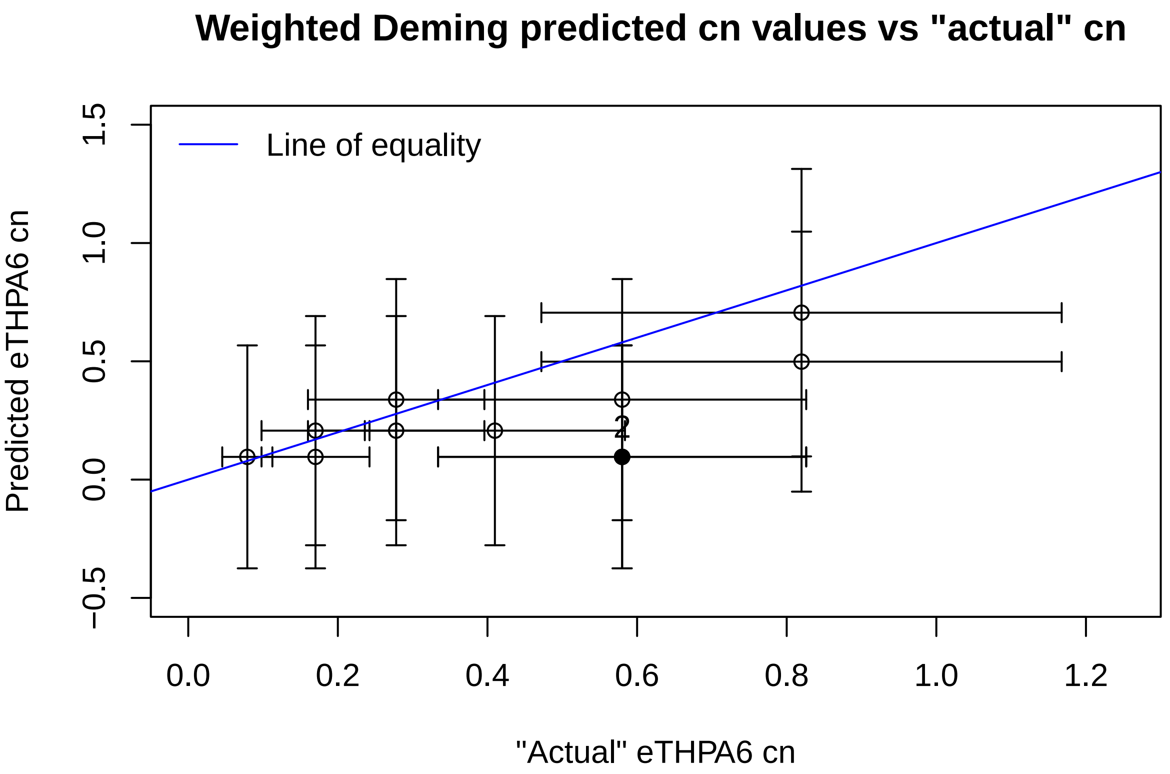


Figure 27: The predicted copy number values from the weighted Deming line plotted against the "actual" copy number values with 95 percent CIs, overlaid with the line of equality. All calculations use the BVL partial detect field data.

Figure 27 shows the predicted (weighted Deming line) values versus the "actual" copy number values with 95% CI's. From Figure 27 it can be seen that the majority of the data points fall below the line of equality, however all the 95% CI's that overlap with the line of equality. This overlap with 95% CI's and the line of equality indicates the weighted Deming model is making satisfactory predictions. Two data points are the same indicated by the opaque point labeled 2 (Figure 27.

Table 31 presents the predicted and "actual" $\hat{\lambda}$ and copy number values for eTHPA6. Both the $\hat{\lambda}$ and copy number values display a moderate difference between the "actual" and predicted values (Table 31).The predicted standard error's for both the predicted $\hat{\lambda}$ and copy number values can be found in section 8.2 and 8.3 of the appendix.

Table 31: The predicted and "actual" eTHPA6 lambda hat and copy number values for the BVL partial detect Fraser river field data.

Sample	"Actual"	Predicted
Lambda hat values		
1	0.1335	0.1624
2	0.2877	0.3500
3	0.4700	0.3500
4	0.6931	0.3500
5	0.2877	0.1624
6	0.9808	0.1624
7	0.4700	0.5718
8	0.9808	0.5718
9	0.9808	0.1624
10	1.3863	0.8432
11	1.3863	1.1932
Copy number values		
1	0.0790	0.0961
2	0.1701	0.2070
3	0.2779	0.2070
4	0.4099	0.2070
5	0.1701	0.0961
6	0.5800	0.0961
7	0.2779	0.3381
8	0.5800	0.3381
9	0.5800	0.0961
10	0.8198	0.4986
11	0.8198	0.7056

6 Conclusion

Prediction curves built by weighted Deming regression led to good predictions in the final eTHPA6 (natural log) copy number values from eTHPA2 C_T or $\hat{\lambda}$ values for all data sets. In the UVic full detect data case, 6 of the 11 eTHPA6 predicted natural log sample means were well predicted. The prediction curve built from the BVL full detect gblock data resulted in a majority of well predicted eTHPA6 natural log sample means. Both partial detect data sets had all their final copy number values well predicted. The UVic full detect data saw an approximate 50/50 split of well and poorly predicted eTHPA6 natural log copy number values. This approximate 50/50 split of well and poorly predicted values, when compared to the majority well predicted values from the BVL full detect field data could be due to the smaller data set or what starting quantities that were included in the full detect gblock data sets. The UVic full detect data included starting quantity 20 and up, whereas the BVL full detect data included starting quantity 4 and up. Thus, the BVL full detect data had 24 more observations than the UVic full detect data. The impact of included/excluded starting quantities is investigated in section 8.5 of the appendix.

Prediction curves built by the recommended weighted Deming regression produced majority well predicted C_T and $\hat{\lambda}$ values. Therefore, we can say that a weighted Deming model is a good choice for building prediction curves between both assays. With the weighted Deming model being chosen, other questions still remain to be addressed in the future. Should a weighted regression model be used for the standard curve estimation and how would it impact the accuracy of the final predicted values? Can we adjust the weighted Deming model to produce more accurate predicted values? Could the inclusion of a parameter for non-technical error in the weighted Deming model produce more accurate predicted values and how would one include this extra error parameter? Should partial detect data with more than (6/8) detects always be analyzed as full detect data?

7 References

Allison, Michael. J., Rene. L. Warren, M. Louie. Lopez, Neha. Acharya-Patel, Jacob. J. Imbery, Lauren. Coombe, Cecilia. L. Yang, Inanc Birol, and Caren. C. Helbing. 2023. "Enabling Robust Environmental DNA Assay Design with "Unikseq" for the Identification of Taxon-Specific Regions Within Whole Mitochondrial Genomes." *Environmental DNA* 5 (5): 1032–47. https://doi.org/10.1002/edn3.438.

Helbing, Caren. 2023. "Helbing Lab Environmental DNA Tools." 2023. https://onlineacademiccommunity.uvic.ca/helbinglab/wp-content/uploads/sites/6036/2021/05/Helbing-lab-eDNA-validated-taxa-list-brief-051121MJA-New.pdf.

Hocking, Morgan D., Jeffrey C. MacAdams, Michael J. Allison, Lauren C. Bergman, Robert Sneiderman, Ben F. Koop, Brian M. Starzomski, Mary L. Lesperance, and Caren C. Helbing. 2022. "Establishing the Signal Above the Noise: Accounting for an Environmental Background in the Detection and Quantification of Salmonid Environmental DNA." *Fishes* 7 (5). https://doi.org/10.3390/fishes7050266.

Kane, Michael. T., and Andrew. A. Mroch. 2020. "Orthogonal Regression, the Cleary Criterion and the Lord's Paradox: Asking the Right Questions." https://doi.org/https://doi.org/10.1002/ets2.12298.

Lavagnini, I., and F. Magno. 2007. "A Statistical Overview on Univariate Calibration, Inverse Regression, and Detection Limits: Application to Gas Chromatography/Mass Spectrometry Technique." *Mass Spectrometry Reviews* 26 (1): 1–18. https://doi.org/10.1002/mas.20100.

Lesperance, Mary L., Michael J. Allison, Lauren C. Bergman, Morgan D. Hocking, and Caren C. Helbing. 2021. "A Statistical Model for Calibration and Computation of Detection and Quantification Limits for Low Copy Number Environmental DNA Samples." *Environmental DNA* 3 (5): 970–81. https://doi.org/10.1002/edn3.220.

Lisy, J. M., A. Cholvadova, and J. Kutej. 1990. "Multiple Straight-Line Least-Squares Analysis with Uncertainties in All Variables." *Computers & Chemistry* 14 (3): 189–92. https://doi.org/10.1016/0097-8485(90)80045-4.

Markovsky, I., and S. Van Huffel. 2007. "Overview of Total Least-Squares Methods." *Signal*

Processing 87 (10): 2283–2302. https://doi.org/10.1016/j.sigpro.2007.04.004.

Nash, J. C. 1990. *Compact Numerical Methods for Computers: Linear Algebra and Function Minimization-Second Edition.* Adam Hilger, Bristol; New York.

Noblitt, S. D., K. E. Berg, D. M. Cate, and C. S. Henry. 2016. "Characterizing Nonconstant Instrumental Variance in Emerging Miniaturized Analytical Techniques." *Analytica Chimica Acta* 915: 64–73. https://doi.org/10.1016/j.aca.2016.02.023.

Rio, F. J. del, J. Riu, and F. X. Rius. 2001. "Prediction Intervals in Linear Regression Taking into Account Errors on Both Axes." *Journal of Chemometrics* 15 (10): 773–88. https://doi.org/10.1002/cem.663.

Ripley, B. D., and M. Thompson. 1987. "Regression Techniques Fro the Detection of Analytical Bias." *Analyst* 112 (4): 377–83. https://doi.org/10.1039/an9871200377.

Souza, Rcsnp, S. C. Leite, W. Meira, and E. R. Hruschka. 2018. "Online Orthogonal Regression Based on a Regularized Squared Loss." *2018 17th Ieee International Conference on Machine Learning and Applications (Icmla)*, 925–30. https://doi.org/10.1109/icmla.2018.00150.

Taberner, D. A., and J. M. K. Dufty. 1995. "An Easier Alternative to Orthogonal Regression for Calculation of International Sensitivity Indexes." *Journal of Clinical Pathology* 48 (10): 901–3. https://doi.org/10.1136/jcp.48.10.901.

Tellinghuisen, J. 2001. "Statistical Error Propagation." *Journal of Physical Chemistry A* 105 (15): 3917–21. https://doi.org/10.1021/jp003484u.

Therneau, T. 2018. "Total Least Squares: Deming, Theil-Sen, and Passing-Bablock Regression." https://cran.r-project.org/web/packages/deming/index.html.

Wu, Jianxin. 2021. *Essentials of Pattern Recognition: An Accessible Approach.* Cambridge University Press. https://doi.org/10.1017/9781108650212.

8 Appendix

8.1 Model Summaries

The summary output of the four fitted prediction models for each labs full and partial detect gblock data sets are displayed in this section.

UVic: full detect

The summary output for the four fitted prediction models using UVic full detect gblock data.

```
## [1] "Least squares"

##
## Call:
## lm(formula = e6 ~ e2, data = gblock)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.681 -0.273 -0.024  0.316  1.996
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.9004     0.4878    -3.9  0.00021 ***
## e2            0.9271     0.0144    64.3  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.584 on 78 degrees of freedom
## Multiple R-squared:  0.981,  Adjusted R-squared:  0.981
## F-statistic: 4.13e+03 on 1 and 78 DF,  p-value: <2e-16

## [1] "Weighted least squares"

##
## Call:
## lm(formula = e6 ~ e2, data = gblock, weights = 1/e6.ct.var)
##
## Weighted Residuals:
##    Min     1Q Median     3Q    Max
## -4.278 -1.145  0.037  1.226  4.150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.2147     0.3382   -6.55  5.6e-09 ***
```

```
## e2            0.9366     0.0109   85.70  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.81 on 78 degrees of freedom
## Multiple R-squared:  0.989,  Adjusted R-squared:  0.989
## F-statistic: 7.34e+03 on 1 and 78 DF,  p-value: <2e-16
```

```
## [1] "Deming regression"
```

```
##
## Call:
## deming(formula = e6 ~ e2, data = gblock)
##
## n= 80
##              Coef se(coef) lower 0.95 upper 0.95
## Intercept -2.1728  0.45269    -3.0601    -1.2855
## Slope      0.9352  0.01476     0.9063     0.9641
##
##    Scale= 0.4276
```

```
## [1] "Weighted Deming regression"
```

```
##
## Call:
## deming(formula = e6 ~ e2, data = gblock, xstd = e2.ct.std, ystd = e6.ct.std)
##
## n= 80
##              Coef se(coef) lower 0.95 upper 0.95
## Intercept -2.4901  0.38040    -3.2357    -1.7445
## Slope      0.9454  0.01229     0.9213     0.9695
##
##    Scale= 1.011
```

UVic: partial detect

The summary output for the four fitted prediction models using UVic partial detect gblock data.

```
## [1] "Least squares"
```

```
##
## Call:
## lm(formula = tablow6$lamhat ~ tablow2$lamhat - 1)
##
## Residuals:
##         1         2         3         4         5
##  2.74e-16  4.26e-02 -4.47e-04  1.38e-01 -2.61e-02
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## tablow2$lamhat    1.010      0.029    34.8  4.1e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.0735 on 4 degrees of freedom
## Multiple R-squared:  0.997,  Adjusted R-squared:  0.996
## F-statistic: 1.21e+03 on 1 and 4 DF,  p-value: 4.07e-06
```

```
## [1] "Weighted least squares"
```

```
##
## Call:
## lm(formula = tablow6$lamhat ~ tablow2$lamhat - 1, weights = 1/(tablow6$vlamhat +
##     epsilon))
##
## Weighted Residuals:
##         1         2         3         4         5
##  5.55e-17  1.00e+00 -9.20e-02  5.32e-01 -3.38e-01
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## tablow2$lamhat     1.09       0.13    8.39   0.0011 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.593 on 4 degrees of freedom
## Multiple R-squared:  0.946,  Adjusted R-squared:  0.933
## F-statistic: 70.4 on 1 and 4 DF,  p-value: 0.0011
```

```
## [1] "Deming regression"
```

```
##
## Call:
## deming(formula = tablow6$lamhat ~ tablow2$lamhat - 1)
##
## n= 5
##            Coef se(coef) lower 0.95 upper 0.95
## Intercept 0.000   0.0000     0.0000      0.000
## Slope     1.012   0.2943     0.4353      1.589
##
##    Scale= 0.05964
```

```
## [1] "Weighted Deming regression"
```

```
##
## Call:
```

```
## deming(formula = tablow6$lamhat ~ tablow2$lamhat - 1, xstd = (tablow2$‘sd lamhat‘ +
##
## n= 5
##           Coef se(coef) lower 0.95 upper 0.95
## Intercept 0.000   0.0000     0.0000       0.00
## Slope     1.116   0.2011     0.7214       1.51
##
##    Scale= 0.5267
```

BVL: full detect
The summary output for the four fitted prediction models using BVL full detect gblock data.

```
## [1] "Least squares"
```

```
##
## Call:
## lm(formula = e6 ~ e2, data = gblockb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4478 -0.2234  0.0016  0.2087  2.5851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.268      0.486   -2.61     0.01 *
## e2             1.004      0.014   71.71   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.692 on 102 degrees of freedom
## Multiple R-squared:  0.981,  Adjusted R-squared:  0.98
## F-statistic: 5.14e+03 on 1 and 102 DF,  p-value: <2e-16
```

```
## [1] "Weighted least squares"
```

```
##
## Call:
## lm(formula = e6 ~ e2, data = gblockb, weights = 1/e6.ct.varb)
##
## Weighted Residuals:
##    Min     1Q Median     3Q    Max
## -3.289 -0.645  0.084  0.793  2.409
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.65185    0.19329   -8.55  1.3e-13 ***
## e2           1.01530    0.00643  157.84  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.15 on 102 degrees of freedom
## Multiple R-squared:  0.996,  Adjusted R-squared:  0.996
## F-statistic: 2.49e+04 on 1 and 102 DF,  p-value: <2e-16
```

```
## [1] "Deming regression"
```

```
##
## Call:
## deming(formula = e6 ~ e2, data = gblockb)
##
## n= 104
##             Coef se(coef) lower 0.95 upper 0.95
## Intercept -1.613  0.50484    -2.6027    -0.6238
## Slope      1.014  0.01648     0.9814     1.0459
##
##    Scale= 0.4872
```

```
## [1] "Weighted Deming regression"
```

```
##
## Call:
## deming(formula = e6 ~ e2, data = gblockb, xstd = e2.ct.stdb,     ystd = e6.ct.stdb)
##
## n= 104
##             Coef se(coef) lower 0.95 upper 0.95
## Intercept -1.640 0.210392     -2.053     -1.228
## Slope      1.015 0.007017      1.001      1.029
##
##    Scale= 0.8925
```

BVL: partial detect

The summary output for the four fitted prediction models using BVL partial detect gblock data.

```
## [1] "Least squares"
```

```
##
## Call:
## lm(formula = tablow6b$lamhat ~ tablow2b$lamhat - 1)
##
## Residuals:
##         1         2         3         4
## -3.48e-16  4.26e-02 -4.87e-02  7.21e-03
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## tablow2b$lamhat   1.2672     0.0302      42    3e-05 ***
```

```
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.0376 on 3 degrees of freedom
## Multiple R-squared:  0.998,  Adjusted R-squared:  0.998
## F-statistic: 1.76e+03 on 1 and 3 DF,  p-value: 2.97e-05
```

```
## [1] "Weighted least squares"
```

```
##
## Call:
## lm(formula = tablow6b$lamhat ~ tablow2b$lamhat - 1, weights = 1/(tablow6b$vlamhat +
##     epsilon))
##
## Weighted Residuals:
##         1         2         3         4
##  8.05e-16  1.00e+00 -3.85e-01  2.48e-01
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## tablow2b$lamhat    1.193      0.172    6.92   0.0062 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.635 on 3 degrees of freedom
## Multiple R-squared:  0.941,  Adjusted R-squared:  0.921
## F-statistic: 47.9 on 1 and 3 DF,  p-value: 0.00618
```

```
## [1] "Deming regression"
```

```
##
## Call:
## deming(formula = tablow6b$lamhat ~ tablow2b$lamhat - 1)
##
## n= 4
##            Coef se(coef) lower 0.95 upper 0.95
## Intercept 0.000    0.000     0.0000      0.000
## Slope     1.269    0.241     0.7962      1.741
##
##    Scale= 0.0285
```

```
## [1] "Weighted Deming regression"
```

```
##
## Call:
## deming(formula = tablow6b$lamhat ~ tablow2b$lamhat - 1, xstd = (tablow2b$‘sd lamhat‘
##
## n= 4
```

```
##            Coef se(coef) lower 0.95 upper 0.95
## Intercept 0.000   0.0000     0.0000      0.000
## Slope     1.216   0.2144     0.7962      1.637
##
##    Scale= 0.5973
```

8.2 Standard error calculations for the predicted values

Here we calculate the standard error from the variance of the predicted eTHPA6 C_T and $\hat{\lambda}$ values for the full and partial detect field data.

8.2.1 UVic full detect data: predicted cycle threshold values

Using equation (18) from Rio, Riu, and Rius (2001), we calculate the variance and the standard error for the mean of each sample of predicted eTHPA6 C_T value. Next, we calculate a 95% CI using the estimated standard error. Table 32 shows the 95% CI for each sample mean of the predicted eTHPA6 C_T values, alongside the "actual" sample mean C_T values. We note that 9 95% CI for the "actual" and predicted sample means overlap (Table 32). The CI's are calculated using the estimated sample mean (μ) and the estimated corresponding standard error (se_{μ}), $\mu \pm 1.96 * se_{\mu}$.

Table 32: The predicted standard errors (se) for the mean of the eTHPA6 Ct values by sample from the UVic full detect Kemano field data. The "actual" and predicted sample mean Ct values and N the number of acceptable technical replicates per sample. Lower and upper bounds for the 95 percent CI for the sample mean of the "actual" and predicted Ct values.

Sample	N	"Actual" lower 95	"Actual" upper 95	Sample "actual" se	Sample mean "actual"	Predicted lower 95	Predicted upper 95	Rio se	Sample mean predicted
1	8	32.15	33.35	0.3061	32.75	30.87	32.33	0.3712	31.60
2	8	32.98	34.79	0.4606	33.89	31.45	33.57	0.5408	32.51
3	8	30.98	32.78	0.4598	31.88	30.30	32.31	0.5129	31.31
4	2	33.31	35.69	0.6081	34.50	34.31	35.77	0.3735	35.04
5	8	28.35	29.07	0.1853	28.71	28.22	29.19	0.2473	28.70
6	2	34.30	35.14	0.2121	34.72	34.66	35.32	0.1682	34.99
7	8	25.56	26.10	0.1383	25.83	23.99	25.22	0.3133	24.61
8	8	28.17	28.88	0.1827	28.52	27.31	28.05	0.1904	27.68
9	8	26.50	27.22	0.1816	26.86	25.72	26.68	0.2442	26.20
10	8	29.71	30.53	0.2102	30.12	28.38	29.53	0.2934	28.96
11	3	34.14	35.69	0.3940	34.91	34.41	35.52	0.2828	34.97

8.2.2 UVic partial detect data: predicted lambda hat values

Using equation (18) from Rio, Riu, and Rius (2001), we calculate the variance and the standard error for each predicted eTHPA6 $\hat{\lambda}$ value. Table 33 shows the standard errors for the predicted $\hat{\lambda}$ values. The all of the “actual” and predicted 95% CIs overlap (Table 33). The CI’s are calculated using the estimated $\hat{\lambda}$ and the estimated corresponding standard error ($se_{\hat{\lambda}}$), $\hat{\lambda} \pm 1.96 * se_{\hat{\lambda}}$.

Table 33: The predicted and "actual" eTHPA6 lambda hat values, and their standard errors (se) for the UVic partial detect Kemano field data (2021/2023). Lower and upper bounds for 95 percent CI for the "actual" and predicted lambda hat values.

Sample	"Actual" lower 95	"Actual" upper 95	Sample "actual" se	Sample mean "actual"	Predicted lower 95	Predicted upper 95	Rio se	Sample mean predicted	Year
1	0.2460	3.9129	0.9354	2.0794	0.1003	2.993	0.7379	1.5465	2021
2	0.2460	3.9129	0.9354	2.0794	0.0000	2.488	0.7111	1.0942	2021
3	0.0002	1.3861	0.3536	0.6931	0.0000	1.665	0.6856	0.3209	2021
1	0.0862	1.8754	0.4564	0.9808	0.0000	1.665	0.6856	0.3209	2023
2	0.0862	1.8754	0.4564	0.9808	0.0000	1.665	0.6856	0.3209	2023
3	0.0002	1.3861	0.3536	0.6931	0.0000	1.876	0.6897	0.5243	2023
4	0.0000	0.6878	0.2041	0.2877	0.0000	1.665	0.6856	0.3209	2023
5	0.0002	1.3861	0.3536	0.6931	0.0000	1.665	0.6856	0.3209	2023
6	0.1860	2.5865	0.6124	1.3863	0.0000	2.488	0.7111	1.0942	2023

8.2.3 BVL full detect data: predicted cycle threshold values

Using equation (18) from Rio, Riu, and Rius (2001), we calculate the variance and the standard error for each sample] mean of predicted eTHPA6 C_T values. Next, we calculate a 95% CI using the standard error. Table 34 shows the 95% CI for each sample mean predicted eTHPA6 C_T value, alongside the "actual" sample mean C_T values. We note that 19 95% CI for the "actual" and predicted sample means overlap (Table 34). The CI's are calculated using the estimated sample mean (μ) and the estimated corresponding standard error (se_μ), $\mu \pm 1.96 * se_\mu$.

Table 34: The predicted standard errors (se) for the mean of the eTHPA6 Ct values by sample from the BVL full detect Fraser river field data. The "actual" and predicted sample mean Ct values and N the number of acceptable technical replicates per sample. Lower and upper bounds for the 95 percent CI for the sample mean of the "actual" and predicted Ct values.

Sample	N	"Actual" lower 95	"Actual" upper 95	Sample "actual" se	Sample mean "actual"	Predicted lower 95	Predicted upper 95	Rio se	Sample mean predicted
1	8	33.59	34.62	0.2626	34.11	32.42	32.98	0.1423	32.70
2	8	33.52	34.91	0.3556	34.22	31.05	35.65	1.1734	33.35
3	8	27.96	28.96	0.2559	28.46	29.88	31.28	0.3570	30.58
4	8	29.58	30.86	0.3249	30.22	31.32	32.09	0.1956	31.71
5	8	29.47	30.72	0.3201	30.09	31.83	33.74	0.4867	32.79
6	8	26.30	27.29	0.2515	26.80	27.40	27.93	0.1339	27.66
7	8	26.17	27.24	0.2718	26.71	27.45	28.02	0.1446	27.74
8	8	29.59	31.29	0.4332	30.44	31.68	32.23	0.1395	31.96
9	8	26.56	27.30	0.1885	26.93	26.72	27.07	0.0907	26.90
10	8	31.46	32.27	0.2054	31.87	33.16	33.67	0.1300	33.41
11	8	23.65	24.24	0.1500	23.94	23.92	24.44	0.1339	24.18
12	8	23.72	24.23	0.1321	23.97	23.90	24.16	0.0684	24.03
13	8	36.85	39.82	0.7593	38.34	38.13	39.63	0.3832	38.88
14	8	36.86	38.71	0.4708	37.78	36.86	40.40	0.9018	38.63
15	8	36.52	39.72	0.8174	38.12	36.83	40.73	0.9935	38.78
16	8	36.87	40.27	0.8655	38.57	37.35	40.26	0.7414	38.81
17	8	36.78	39.72	0.7510	38.25	37.17	42.33	1.3169	39.75
18	6	39.00	42.10	0.7910	40.55	37.91	42.61	1.1988	40.26
19	6	38.88	41.10	0.5665	39.99	37.63	40.62	0.7623	39.13
20	8	38.38	40.91	0.6454	39.65	38.15	40.54	0.6095	39.34
21	8	36.99	39.11	0.5421	38.05	37.40	39.66	0.5767	38.53
22	8	35.96	38.15	0.5603	37.05	36.24	38.35	0.5382	37.30
23	6	38.50	40.50	0.5111	39.50	38.80	42.41	0.9196	40.60
24	4	38.55	41.80	0.8302	40.17	38.54	42.03	0.8911	40.28
25	4	38.30	40.42	0.5409	39.36	39.71	40.13	0.1086	39.92
26	4	36.96	42.25	1.3480	39.60	38.32	41.08	0.7023	39.70

8.2.4 BVL partial detect data: predicted lambda hat values

Using equation (18) from Rio, Riu, and Rius (2001), we calculate the variance and the standard error for each predicted eTHPA6 $\hat{\lambda}$ values. Table 35 shows the standard errors for the predicted $\hat{\lambda}$ values. The all of the "actual" and predicted 95% CIs overlap (Table 35). The CI's are calculated using the estimated $\hat{\lambda}$ and the estimated corresponding standard error ($se_{\hat{\lambda}}$), $\hat{\lambda} \pm 1.96 * se_{\hat{\lambda}}$.

Table 35: The predicted and "actual" eTHPA6 lambda hat and copy number values, and the standard error for the predicted lambda hat and copy number values for the BVL partial detect Fraser river field data.

Sample	"Actual" lower 95	"Actual" upper 95	Sample "actual" se	Sample mean "actual"	Predicted lower 95	Predicted upper 95	Rio se	Sample mean predicted
1	0.0000	0.3954	0.1336	0.1335	0.0000	0.4860	0.1651	0.1624
2	0.0000	0.6878	0.2041	0.2877	0.0092	0.6907	0.1739	0.3500
3	0.0000	1.0068	0.2739	0.4700	0.0092	0.6907	0.1739	0.3500
4	0.0002	1.3861	0.3536	0.6931	0.0092	0.6907	0.1739	0.3500
5	0.0000	0.6878	0.2041	0.2877	0.0000	0.4860	0.1651	0.1624
6	0.0862	1.8754	0.4564	0.9808	0.0000	0.4860	0.1651	0.1624
7	0.0000	1.0068	0.2739	0.4700	0.1969	0.9466	0.1913	0.5718
8	0.0862	1.8754	0.4564	0.9808	0.1969	0.9466	0.1913	0.5718
9	0.0862	1.8754	0.4564	0.9808	0.0000	0.4860	0.1651	0.1624
10	0.1860	2.5865	0.6124	1.3863	0.4115	1.2749	0.2203	0.8432
11	0.1860	2.5865	0.6124	1.3863	0.6722	1.7141	0.2658	1.1932

8.3 Error propagation

Here, we calculate the propagated error using equation (23) for the prediction and standard curves from the UVic and BVL partial and full detect data sets. Using equation (19) from Lavagnini and Magno (2007), we calculate the Lavagnini standard error for the "actual" and predicted C_T values sample mean. The CI's are calculated using the estimated sample mean (μ) and the estimated corresponding standard error (se_μ), $\mu \pm 1.96 * se_\mu$.

8.3.1 UVic: full detect data

The propagated standard errors from the standard curve for the UVic full detect Kemano field data are between (0.25-0.50) (Table 36). The propagated standard errors for the predicted sample means in Table 36 are majority larger than the predicted Lavagnini standard errors. We expect the propagated standard errors for the means of the predicted values to be large, as it is accounting for the error introduced by the prediction and standard curve. Comparing the 95% CI's for the "actual" and predicted sample means, we note 6 95% CI's overlap (Table 36).

Table 36: The propagated standard error (se) and Lavagnini standard error by sample from the standard curve for the UVic full detect Kemano field data. The lower and upper bounds from the 95 percent CI for the sample mean of the "actual" and predicted copy number values. The mean of the "actual" and predicted copy number values by sample for eTHPA6 and N the number of acceptable technical replicates per sample.

Sample	N	"actual" lower 95	"actual" upper 95	"Actual" Lavagnini se	Sample mean "actual"	Propagated lower 95	Propagated upper 95	Propagated se	Predicted Lavagnini se	Sample mean predicted
1	8	3.313	3.546	0.0593	3.430	3.620	4.718	0.2798	0.0536	4.169
2	8	2.572	2.829	0.0655	2.701	2.855	4.313	0.3719	0.0581	3.584
3	8	3.880	4.096	0.0549	3.988	3.640	5.068	0.3645	0.0523	4.354
4	2	2.143	2.472	0.0840	2.307	1.431	2.494	0.2713	0.0867	1.962
5	8	5.936	6.108	0.0438	6.022	5.536	6.516	0.2499	0.0437	6.026
6	2	1.999	2.333	0.0851	2.166	1.672	2.319	0.1650	0.0864	1.996
7	8	7.785	7.955	0.0433	7.870	8.060	9.246	0.3025	0.0463	8.653
8	8	6.055	6.225	0.0434	6.140	6.231	7.134	0.2302	0.0424	6.683
9	8	7.125	7.291	0.0423	7.208	7.111	8.151	0.2654	0.0428	7.631
10	8	5.025	5.212	0.0476	5.118	5.356	6.369	0.2583	0.0443	5.862
11	3	1.885	2.199	0.0799	2.042	1.579	2.437	0.2190	0.0802	2.008

8.3.2 UVic: partial detect data

The propagated standard errors from the standard curve for the UVic partial detect Kemano field data (2021/2023) lie between (1.30-1.50) (Table 37). The propagated standard errors for the predicted sample means from Table 37 are larger than the Delta method predicted standard errors. Comparing the 95% CI’s for the “actual” and predicted sample means, we note all 95% CI’s overlap (Table 37).

Table 37: The propagated standard error (se) and Lavagnini standard error by sample from the standard curve for the UVic partial detect Kemano field data (2021/2023). The lower and upper bounds from the 95 percent CI for the sample mean of the "actual" and predicted copy number values. The mean of the "actual" and predicted copy number values by sample and year for eTHPA6.

Sample	"actual" lower 95	"actual" upper 95	"Actual" Lavagnini se	"Actual" cn	Propagated lower 95	Propagated upper 95	Propagated se	Predicted Lavagnini se	Predicted cn	Year
1	1.9525	4.4261	0.6310	3.1893	0	5.069	1.376	0.4693	2.3720	2021
2	1.9525	4.4261	0.6310	3.1893	0	4.324	1.350	0.3320	1.6782	2021
3	0.6508	1.4754	0.2103	1.0631	0	3.089	1.325	0.0974	0.4922	2021
1	0.9210	2.0877	0.2976	1.5043	0	3.089	1.325	0.0974	0.4922	2023
2	0.9210	2.0877	0.2976	1.5043	0	3.089	1.325	0.0974	0.4922	2023
3	0.6508	1.4754	0.2103	1.0631	0	3.409	1.329	0.1591	0.8042	2023
4	0.2701	0.6123	0.0873	0.4412	0	3.089	1.325	0.0974	0.4922	2023
5	0.6508	1.4754	0.2103	1.0631	0	3.089	1.325	0.0974	0.4922	2023
6	1.3017	2.9507	0.4207	2.1262	0	4.324	1.350	0.3320	1.6782	2023

8.3.3 BVL: full detect data

The propagated standard errors from the standard curve for the BVL full detect Fraser river field data lie between (0.22-0.71) (Table 38). The propagated standard errors for the predicted sample means in Table 38 are majority larger than the predicted Lavagnini standard errors. We expect the propagated standard errors for the means of the predicted values to be large, as it is accounting for the error introduced by the prediction and standard curve. Comparing the 95% CI's for the “actual” and predicted sample means, we note 17 95% CI's overlap (Table 38).

Table 38: The propagated standard error (se) and Lavagnini standard error by sample from the standard curve for the BVL full detect Fraser river field data. The lower and upper bounds from the 95 percent CI for the sample mean of the "actual" and predicted copy number values. The mean of the "actual" and predicted copy number values by sample for eTHPA6 and N the number of acceptable technical replicates per sample.

Sample	N	"actual" lower 95	"actual" upper 95	"Actual" Lavagnini se	Sample mean "actual"	Propagated lower 95	Propagated upper 95	Propagated se	Predicted Lavagnini se	Sample mean predicted
1	8	4.1369	4.2319	0.0242	4.1844	4.8336	5.2650	0.1101	0.0228	5.0493
2	8	4.0707	4.1662	0.0244	4.1185	3.2335	6.0666	0.7227	0.0233	4.6501
3	8	7.6046	7.6953	0.0231	7.6499	5.8798	6.8149	0.2385	0.0220	6.3473
4	8	6.5256	6.6121	0.0221	6.5689	5.3685	5.9472	0.1476	0.0222	5.6578
5	8	6.6027	6.6894	0.0221	6.6460	4.3856	5.6040	0.3108	0.0228	4.9948
6	8	8.6203	8.7190	0.0252	8.6696	7.8785	8.3943	0.1316	0.0240	8.1364
7	8	8.6745	8.7737	0.0253	8.7241	7.8253	8.3588	0.1361	0.0239	8.0921
8	8	6.3900	6.4764	0.0220	6.4332	5.2683	5.7407	0.1205	0.0223	5.5045
9	8	8.5377	8.6357	0.0250	8.5867	8.3792	8.8353	0.1163	0.0250	8.6072
10	8	5.5155	5.6027	0.0222	5.5591	4.3910	4.8282	0.1115	0.0234	4.6096
11	8	10.3586	10.4777	0.0304	10.4182	9.9941	10.5542	0.1429	0.0299	10.2742
12	8	10.3404	10.4594	0.0303	10.3999	10.1228	10.6092	0.1241	0.0302	10.3660
13	8	1.5281	1.6531	0.0319	1.5906	0.7897	1.7292	0.2397	0.0331	1.2594
14	8	1.8694	1.9897	0.0307	1.9295	0.3219	2.4988	0.5553	0.0325	1.4104
15	8	1.6602	1.7833	0.0314	1.7217	0.1205	2.5166	0.6112	0.0329	1.3186
16	8	1.3837	1.5108	0.0324	1.4473	0.4073	2.1987	0.4570	0.0329	1.3030
17	8	1.5807	1.7049	0.0317	1.6428	0.0000	2.3083	0.8091	0.0351	0.7225
18	6	0.1561	0.3054	0.0381	0.2308	0.0000	1.8531	0.7362	0.0374	0.4101
19	6	0.5053	0.6494	0.0368	0.5773	0.1867	2.0260	0.4692	0.0348	1.1064
20	8	0.7172	0.8540	0.0349	0.7856	0.2359	1.7101	0.3761	0.0342	0.9730
21	8	1.7057	1.8282	0.0313	1.7670	0.7722	2.1715	0.3570	0.0323	1.4719
22	8	2.3193	2.4337	0.0292	2.3765	1.5715	2.8820	0.3343	0.0297	2.2267
23	6	0.8070	0.9467	0.0356	0.8768	0.0000	1.3080	0.5656	0.0382	0.1995
24	4	0.3876	0.5411	0.0392	0.4643	0.0000	1.4695	0.5479	0.0394	0.3956
25	4	0.8910	1.0375	0.0374	0.9642	0.4653	0.7741	0.0788	0.0386	0.6197
26	4	0.7381	0.8867	0.0379	0.8124	0.0000	1.6017	0.4327	0.0381	0.7536

8.3.4 BVL: partial detect data

The propagated standard errors from the standard curve for the BVL partial detect Fraser river field data lie between (0.24-0.35) (Table 39). The propagated standard errors for the predicted sample means from Table 39 are larger than the Delta method predicted standard errors. Comparing the 95% CI's for the "actual" and predicted sample means, we note all 95% CI's overlap (Table 37).

Table 39: The propagated standard error (se) by sample from the standard curve for the BVL partial detect Fraser river field data. The lower and upper bounds from the 95 percent CI for the predicted copy number values for each sample. The "actual" and predicted copy number (cn) values by sample for eTHPA6.

Sample	"actual" lower 95	"actual" upper 95	"Actual" Lavagnini se	"Actual" cn	Propagated lower 95	Propagated upper 95	Propagated se	Predicted Lavagnini se	Predicted cn
1	0.04	0.1125	0.0171	0.0790	0.0000	0.5672	0.2404	0.0208	0.0961
2	0.0980	0.2423	0.0368	0.1701	0.0000	0.6913	0.2471	0.0448	0.2070
3	0.16	0.3959	0.0602	0.2779	0.0000	0.6913	0.2471	0.0448	0.2070
4	0.2360	0.5838	0.0887	0.4099	0.0000	0.6913	0.2471	0.0448	0.2070
5	0.09	0.2423	0.0368	0.1701	0.0000	0.5672	0.2404	0.0208	0.0961
6	0.3340	0.8261	0.1255	0.5800	0.0000	0.5672	0.2404	0.0208	0.0961
7	0.16	0.3959	0.0602	0.2779	0.0000	0.8477	0.2600	0.0732	0.3381
8	0.3340	0.8261	0.1255	0.5800	0.0000	0.8477	0.2600	0.0732	0.3381
9	0.33	0.8261	0.1255	0.5800	0.0000	0.5672	0.2404	0.0208	0.0961
10	0.4720	1.1676	0.1774	0.8198	0.0000	1.0481	0.2803	0.1079	0.4986
11	0.47	1.1676	0.1774	0.8198	0.0981	1.3131	0.3100	0.1527	0.7056

8.4 Field data models and prediction

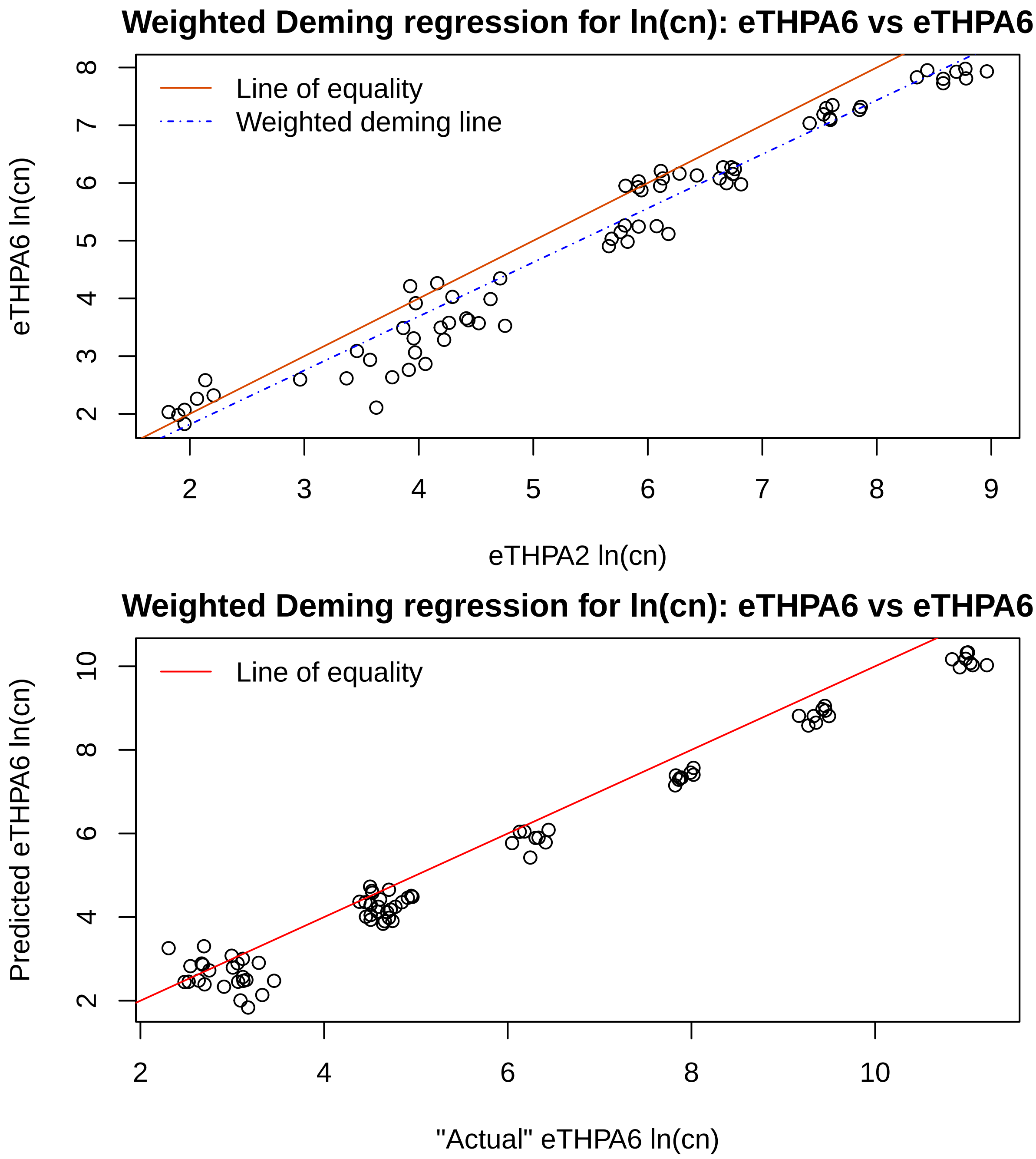
In this section we use the recommended weighted Deming regression to fit models to the four field data sets. We convert the C_T and $\hat{\lambda}$ values of the field data to the (natural log) copy number scale by applying the standard curves. The (natural log) copy number values of the field data have weighted Deming models fit to them to produce prediction curves. We then use the gblock data to validate the prediction curve.

8.4.1 UVic: full detect data

```
##
## Call:
## deming(formula = uvic.cn6 ~ uvic.cn2, xstd = e2var.u, ystd = e6var.u)
##
## n= 71
##                Coef se(coef) lower 0.95 upper 0.95
## Intercept -0.05507  0.17394    -0.3960     0.2858
## Slope      0.93606  0.02542     0.8862     0.9859
##
##    Scale= 1.505
```

Above is the model for the eTHPA6 ln(cn) values as a fuction of eTHPA2 ln(cn) values. From the model summary for the UVic full detect field data, we notice that 0 is included in the

CI for the intercept, but 1 is not included in the CI for the slope. This indicates that the slope is significantly different from 1, but the intercept is not significantly different from 0. Figure 28 (top) displays the estimated weighted Deming prediction line alongside the line of equality, we observe that the weighted Deming prediction line fits the data adequately (Figure 28 ,top). The estimated prediction line from the model summary is $c\hat{n}_6 = 0.936 * c\hat{n}_2 - 0.055$. Figure 28 (bottom) displays the predicted versus “actual” natural log copy number values, which appear to be underestimated.


Weighted Deming regression for ln(cn): eTHPA6 vs eTHPA6

Weighted Deming regression for ln(cn): eTHPA6 vs eTHPA6

Figure 28: Top: the predicted ln(cn) values from the weighted Deming prediction line plotted against the “actual” ln(cn) values, overlaid with the line of equality. Bottom: the predicted ln(cn) values from the weighted Deming prediction line plotted against the “actual” ln(cn) values, overlaid with the line of equality. Both plots used the UVic full detect field data.

8.4.2 UVic: partial detect data

```
##
## Call:
## deming(formula = uvic.cn6l ~ uvic.cn2l - 1)
##
## n= 9
##             Coef se(coef) lower 0.95 upper 0.95
## Intercept 0.000     0.00      0.000      0.000
## Slope     1.558     0.22      1.126      1.989
##
##    Scale= 0.3096
```

Above is the model for the eTHPA6 ln(cn) values as a fuction of eTHPA2 ln(cn) values. From the model summary for the UVic partial detect field data, we notice that 1 is not included in the CI for the slope. This indicates that the slope is significantly different from 1. Figure 29 (top) displays the estimated Deming prediction line alongside the line of equality, comparing the fits of the Deming prediction line appears to fit the data better than the line of equality (Figure 29 ,top). The estimated prediction line from the model summary is $c\hat{n}_6 = 1.558 * c\hat{n}_2$. Figure 29 (bottom) displays the predicted versus "actual" copy number values, which lie on the line of equality or are overestimated. Despite the fact that the literature recommends a weighted regression over an unweighted one, we cannot use one here. The reason is that weighting the variables removes influence from the largest observations. The removal of influence from the largest points leads to a highly skewed final model towards the smaller points. For a data set with five different observations, we cannot lose any information for any points. Therefore, the unweighted and no intercept Deming model will suffice as it still considers the error accompanying the observations.

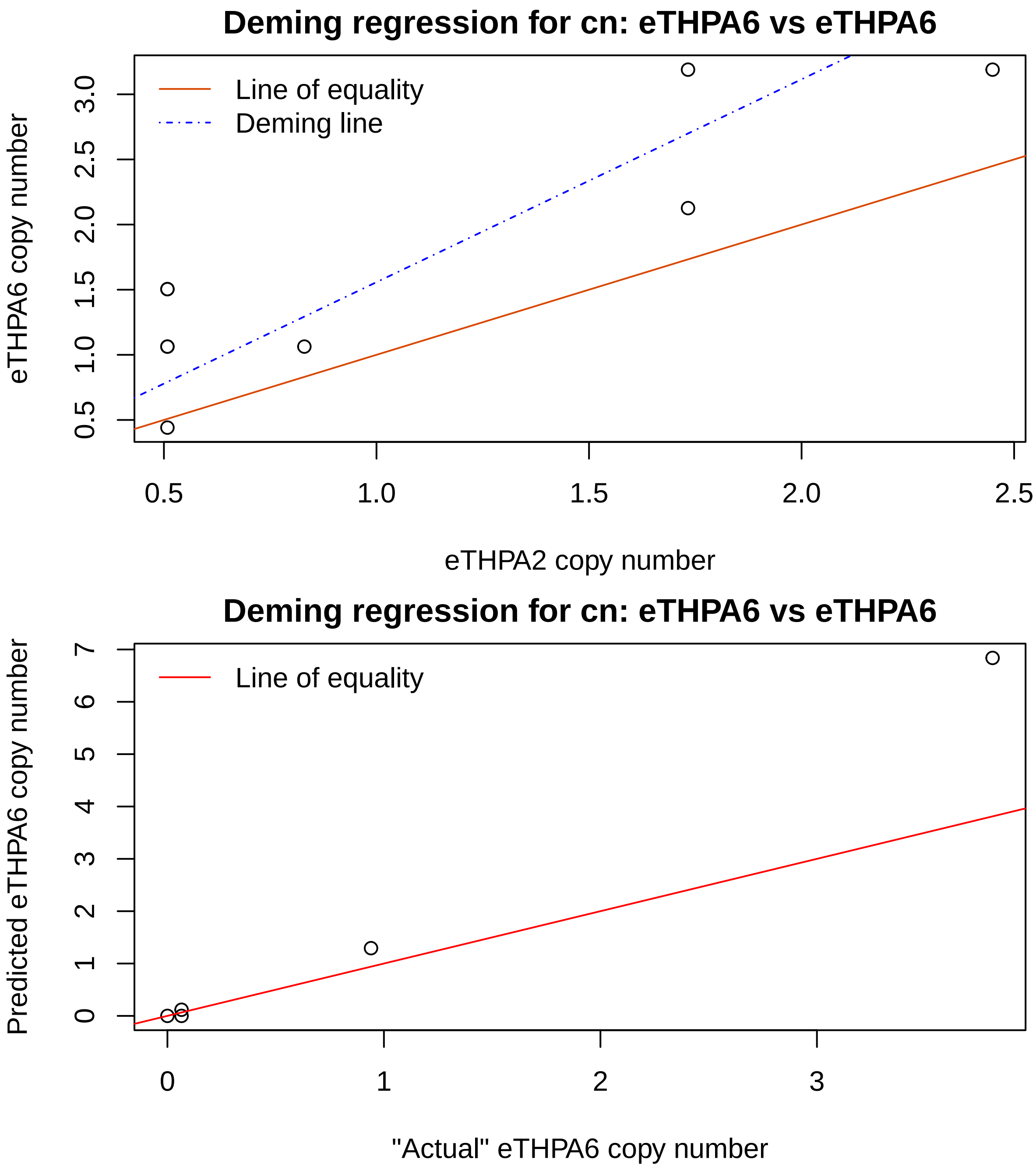


Figure 29: Top: the predicted copy number values from the Deming prediction line plotted against the “actual” copy number values, overlaid with the line of equality. Bottom: the predicted copy number values from the Deming prediction line plotted against the “actual” copy number values, overlaid with the line of equality. Both plots used the UVic partial detect field data.

8.4.3 BVL: full detect data

```
##
## Call:
## deming(formula = bvl.cn6 ~ bvl.cn2, xstd = e2var.b, ystd = e6var.b)
##
## n= 190
##             Coef se(coef) lower 0.95 upper 0.95
## Intercept 0.5636  0.15334     0.2631     0.8642
## Slope     0.9622  0.01558     0.9317     0.9928
##
##    Scale= 2.442
```

Above is the model for the eTHPA6 ln(cn) values as a fuction of eTHPA2 ln(cn) values. From the model summary for BVL full detect field data, we notice that 0 is not included in the CI for the intercept, and 1 is not included in the CI for the slope. This indicates that the slope is significantly different from 1, and the intercept is significantly different from 0. Figure 30 (top) displays the estimated weighted Deming prediction line alongside the line of equality, we observe the weighted Deming fits the data better than the line of equality line (Figure 30 (top)). The estimated prediction line from the model summary is $c\hat{n}_6 = 0.962 * c\hat{n}_2 + 0.564$. Figure 30 (bottom) displays the predicted versus "actual" natural log copy number values, which look to be majority underestimated.

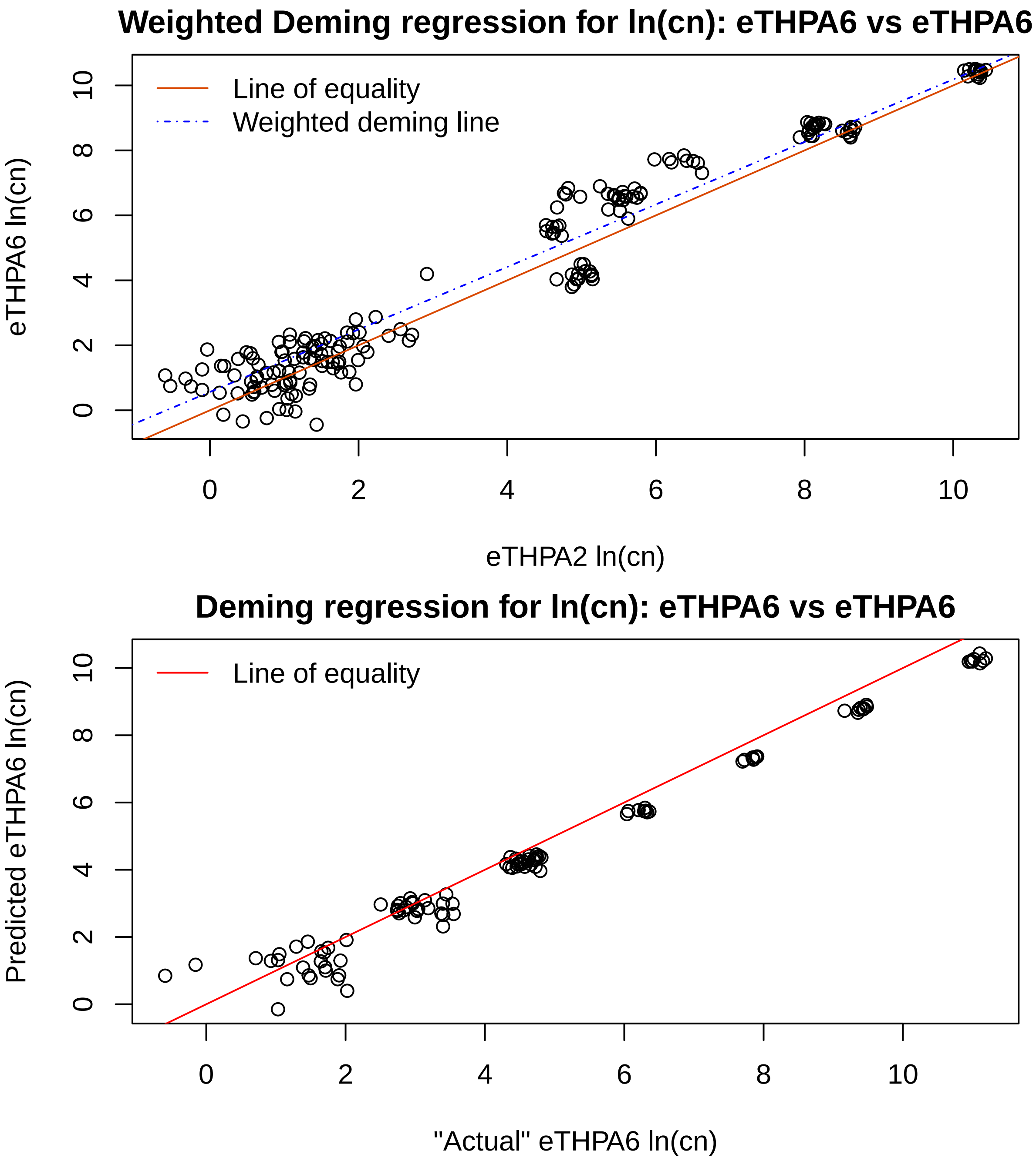


Figure 30: Top: the predicted ln(cn) values from the weighted Deming prediction line plotted against the “actual” ln(cn) values, overlaid with the line of equality. Bottom: the predicted ln(cn) values from the weighted Deming prediction line plotted against the “actual” ln(cn) values, overlaid with the line of equality. Both plots used the BVL full detect field data.

8.4.4 BVL: partial detect data

```
##
## Call:
## deming(formula = bvl.cn6l ~ bvl.cn2l - 1)
##
## n= 11
##            Coef se(coef) lower 0.95 upper 0.95
## Intercept 0.000   0.0000       0.00       0.00
## Slope     2.005   0.4515       1.12       2.89
##
##    Scale= 0.1431
```

Above is the model for the eTHPA6 ln(cn) values as a fuction of eTHPA2 ln(cn) values. From the model summary for the BVL partial detect field data, we notice that 1 is not included in the CI for the slope. Thus, the slope is significantly different from 1. Figure 31 (top) displays the estimated Deming prediction line alongside the line of equality, the estimated Deming line appears to fit the data better (Figure 31 ,top). The estimated prediction line from the model summary is $\hat{cn}_6 = 2.01 * \hat{cn}_2$. Figure 31 (bottom) displays the predicted versus "actual" copy number values, which are mostly underestimated. Despite the fact that the literature recommends a weighted regression over an unweighted one, we cannot use one here. The reason is that weighting variables removes influence from the largest observations. The removal of influence from the largest points leads to a highly skewed final model towards the smaller points. For a data set with six different observation, we cannot lose any information for any points. Therefore, an unweighted and no intercept Deming model will suffice as it still considers the error accompanying the observations.

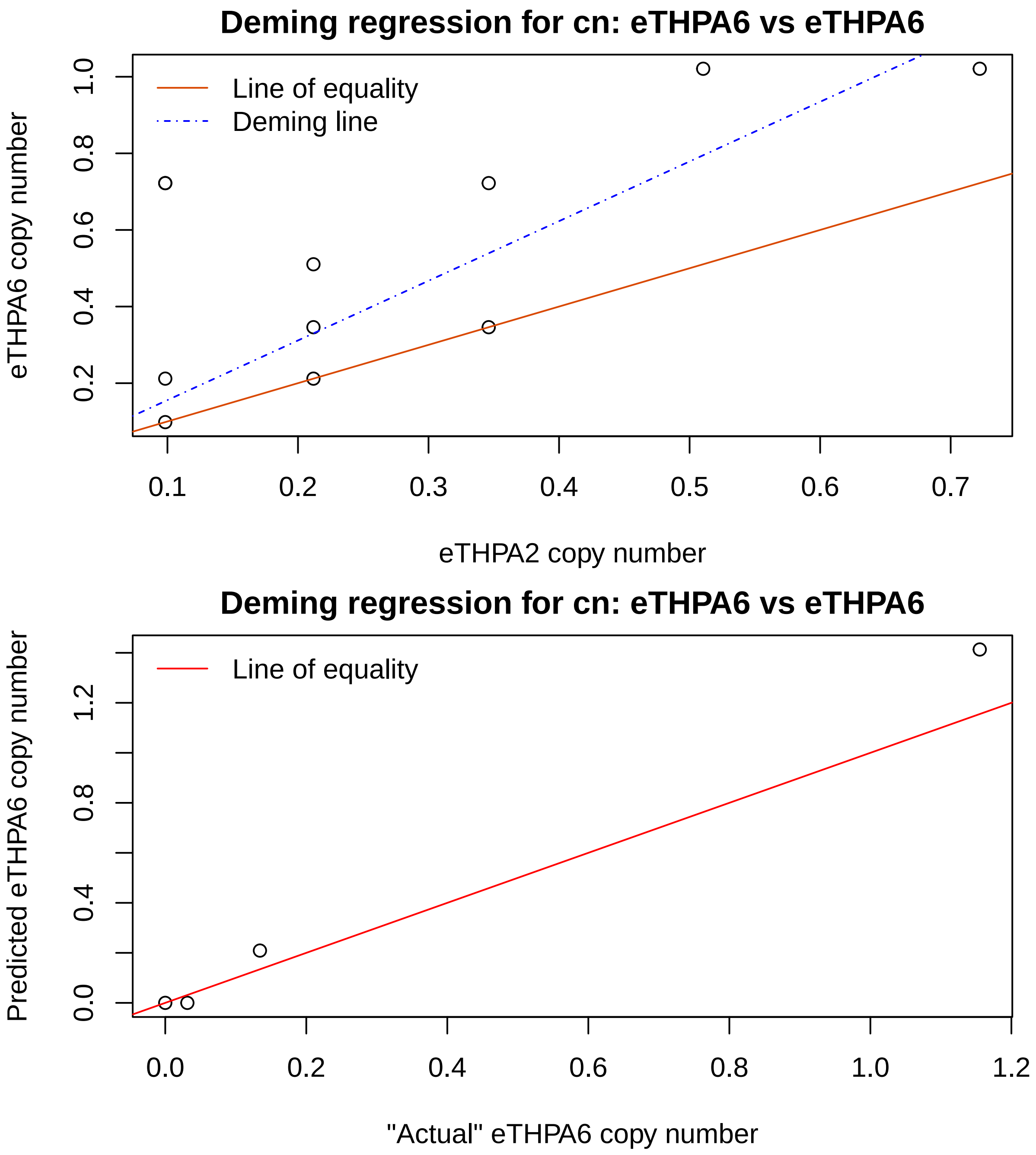


Figure 31: Top: the predicted copy number values from the Deming prediction line plotted against the "actual" copy number values, overlaid with the line of equality. Bottom: the predicted copy number values from the Deming prediction line plotted against the "actual" copy number values, overlaid with the line of equality. Both plots used the BVL partial detect field data.

Overall, using estimated (natural log) copy number values calculated from field data does not produce prediction models that are better than the ones derived from gblock data. In fact, it introduces more error that we are not able to account for in the prediction model, leading to worse overestimates or underestimates for (natural log) copy number values. The partial detect field data sets, have too few data points to ensure our regression is a good fit.

8.5 Starting quantity inclusion/exclusion

In this section are the model summaries for the prediction and standard curve for the BVL full detect gblock data without starting quantity 4. A plot for the predicted versus “actual” natural log copy number values using the new equations is shown.

```
##
## Call:
## lm(formula = gb6.ct$Mean ~ log(sq)[-1])
##
## Residuals:
##       1       2       3       4       5       6
## -0.0511  0.0634 -0.0144 -0.0103  0.0657 -0.0533
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.85385    0.06649     614  4.2e-11 ***
## log(sq)[-1] -1.62129    0.00882    -184  5.3e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.0594 on 4 degrees of freedom
## Multiple R-squared:     1,   Adjusted R-squared:     1
## F-statistic: 3.38e+04 on 1 and 4 DF,  p-value: 5.25e-09
```

Above is the output for the the standard curve model from the full detect BVL gblock data without starting quantity 4. The new standard curve does not differ greatly from the old standard curve for the BVL full detect gblock data that included starting quantity 4.

```
##
## Call:
```

```
## deming(formula = e6 ~ e2, data = gbvl, xstd = b2w, ystd = b6w)
##
## n= 80
##             Coef se(coef) lower 0.95 upper 0.95
## Intercept -1.666 0.216886     -2.091     -1.241
## Slope      1.016 0.007261      1.002      1.030
##
##    Scale= 0.8809
```

Above is the weighted Deming model for the BVL full detect gblock data without starting quantity 4. The new prediction equation is similar to the old prediction equation.

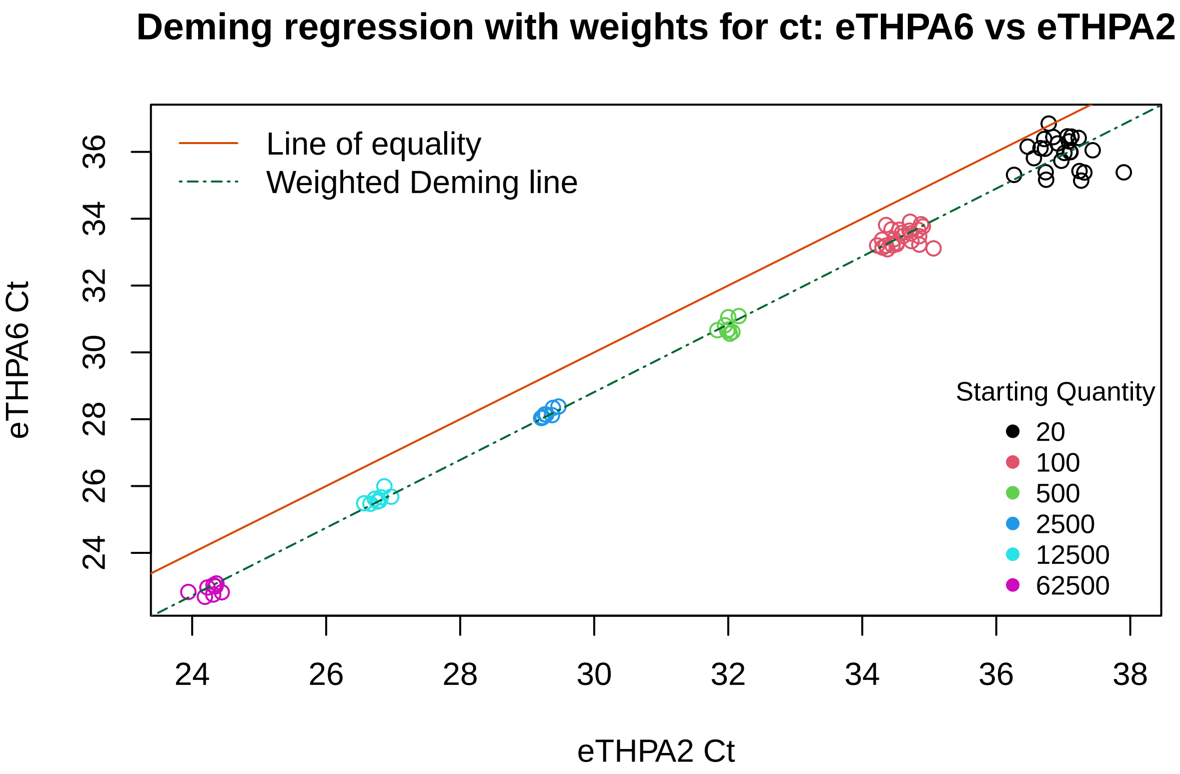


Figure 32: The new estimated weighted Deming regression line overlaid with the line of equality for eTHPA6 vs eTHPA2 cycle threshold values.

Figure 32 is the plot of the new prediction equation for the BVL full detect gblock data without starting quantity 4. The new model does not differ greatly from the old model in Figure 21 (top).

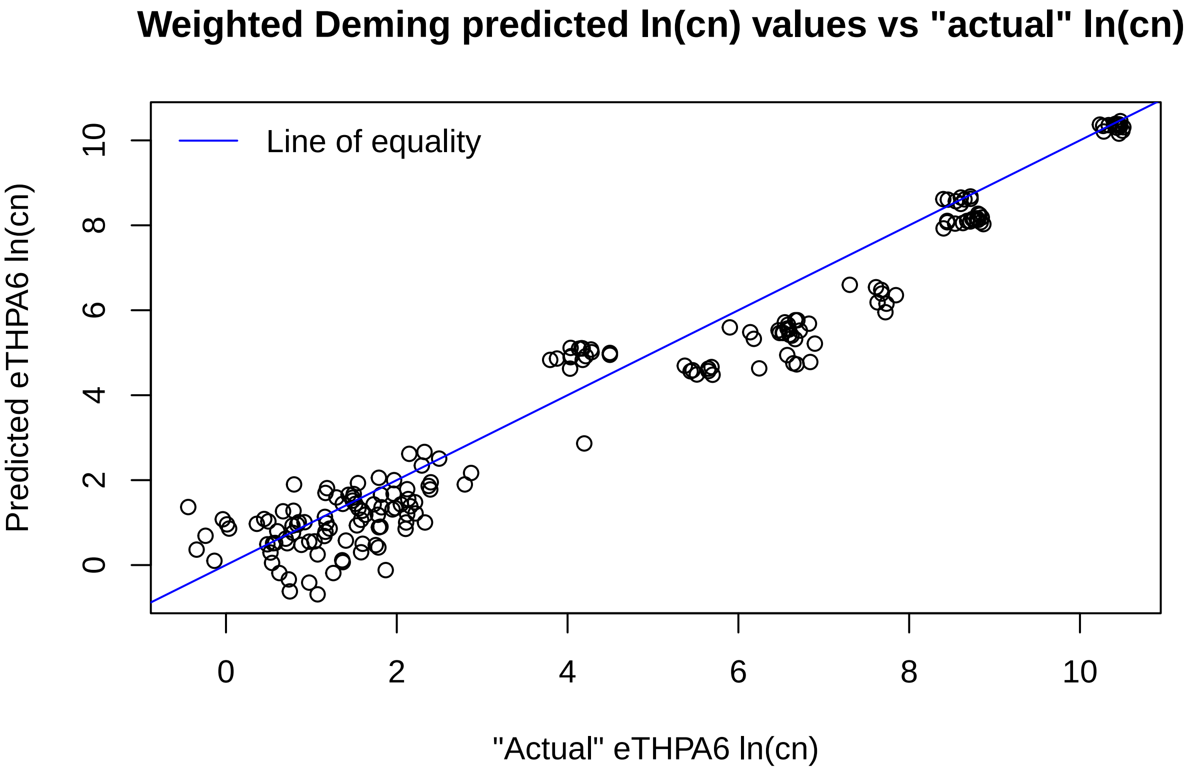


Figure 33: Top: the new predicted ln(cn) values from the new weighted Deming prediction line plotted against the "actual" ln(cn) values, overlaid with the line of equality.

Figure 33 displays the predicted versus “actual” natural log copy number values from the new prediction and standard curves. Comparing Figure 33 with Figure 26, we do not notice any major difference between the two plots. The similarities between the two plots is attributed to the lack of differences between the new and old prediction and standard curve equations. Therefore, the inclusion/exclusion of starting quantities in the BVL full detect gblock data set does not appear to impact the final natural log copy number values.