The Stata Journal (2016)
16, Number 2, pp. 316–330

Implementing weighted-average estimation of substance concentration using multiple dilutions

Ying Xu
Center for Quantitative Medicine
Duke–NUS Graduate Medical School
Singapore, Singapore
tinayxu@gmail.com

Paul Milligan
Department of Infectious Disease Epidemiology
London School of Hygiene and Tropical Medicine
London, UK
paul.milligan@lshtm.ac.uk

Edmond J. Remarque
Department of Parasitology
Biomedical Primate Research Centre
Rijswijk, The Netherlands
remarque@gmail.com

Yin Bun Cheung
Center for Quantitative Medicine
Duke–NUS Graduate Medical School
Singapore, Singapore
and Department for International Health
University of Tampere
Tampere, Finland
yinbun.cheung@duke-nus.edu.sg

Abstract. In medicine and chemistry, immunoassays are often used to measure substance concentration. These tests use an S-shaped standard curve to map the observed optical responses to the underlying concentration. The enzyme-linked immunosorbent assay is one such test that is commonly used to measure antibody concentration in vaccine and infectious disease research. The enzyme-linked immunosorbent assay and other immunoassays usually involve a series of doubling or tripling dilutions of the test samples so that some of the diluted samples fall within the near-linear range in the center of the standard curve. The dilution that falls within or is nearest to the center of the near-linear range may then be selected for statistical analysis. This common practice of using one dilution does not fully use the information from multiple dilutions and reduces accuracy. We describe a recently proposed weighted-average estimation approach for analyzing multiple-dilution data (Cheung et al. 2015, *Journal of Immunological Methods* 417: 115–123), and we present the new `wavemid` command, which carries out the approach. We also present the new command `midreshape`, which processes raw data in text format exported from some microplate readers into analyzable data format. We use data from an experimental study of malaria vaccine candidates to demonstrate use of the two commands.

Keywords: st0434, wavemid, midreshape, immunoassay, multiple dilutions, weighted-average estimation

1 Introduction

In medicine and chemistry, immunoassays are important tools for detecting and quantifying substance concentration. For example, in vaccine and infectious disease research, the concentration of an antibody or antigen is often measured using the enzyme-linked immunosorbent assay (ELISA). This assay measures the concentration in each sample indirectly with an optical signal generated by an enzyme. For each test sample, the concentration is determined by comparing the observed optical density (OD) with an S-shaped standard curve (see figure 1 for an example). For example, if the observed OD is 2, then the estimated log(concentration) would be 2.56. Typically, the standard curve is a sigmoid curve described by a four-parameter logistic model on the logarithm scale (Ratkowsky and Reedy 1986; O'Connell, Belanger, and Haaland 1993) that characterizes the relationship between the OD and the concentration of the target substance in a set of standard solutions with known concentrations.

A common limitation of immunoassays is that they can obtain an accurate estimate of concentration only if the sample concentration falls within the "optimal" range (that is, the near-linear part in the center of the standard sigmoid curve). When the sample concentration falls outside this range, the assay lacks accuracy. To solve this, one can conduct a series of dilutions of each original test sample, as illustrated in figure 1. For each test sample, the observed response from one dilution that is within or nearest to the center of the optimal range is chosen for subsequent statistical analysis, while data from the other dilutions are ignored. The concentration of the substance in the original sample is then estimated using the inversion of a standard curve to obtain the concentration level of the chosen diluted sample and then multiplying the estimate by the corresponding dilution factor.

One problem with this approach is that selecting one diluted sample may involve some arbitrariness, thus limiting reproducibility. More importantly, using only one data point per original sample for statistical analysis does not fully use the available information, thus reducing accuracy. Remarque (2007) proposed calculating a weighted average of all data points by assigning the weights in a 100:1 ratio for data points in the optimal range versus those outside. This approach uses all the data. However, there is no formal justification to the weights and the ranges to which the weights are applicable, other than knowledge on the technical limits of the laboratory apparatus.

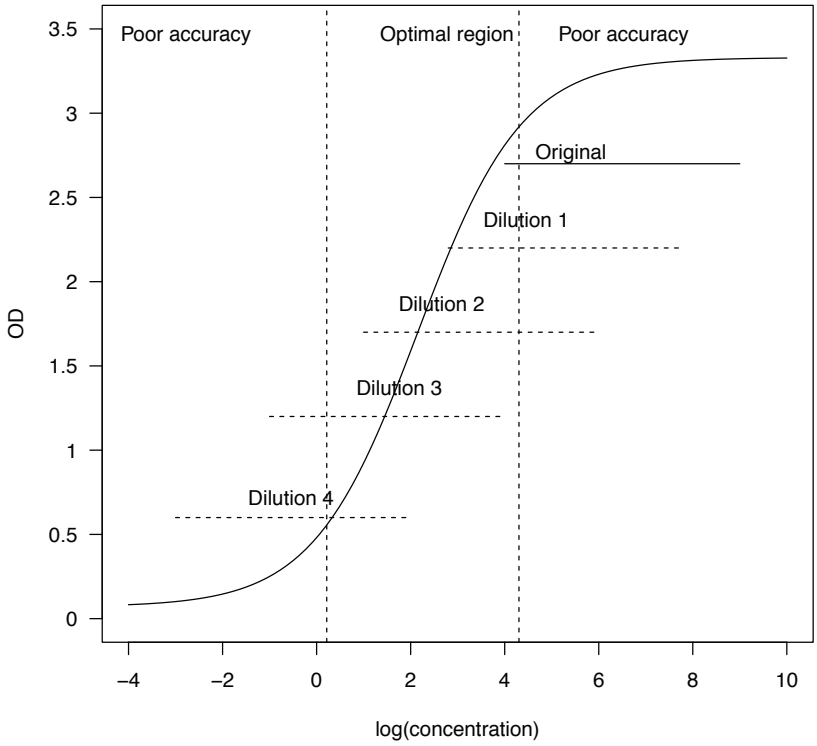

Figure 1. Illustration of a standard curve relating OD to sample concentration with four dilution levels of a set of original samples

To circumvent these problems, Cheung et al. (2015) proposed a weighted-average estimator for using data from multiple dilutions, where the weights are proportional to the inverse of the variance of the individual estimates. They showed by simulation that the proposed estimator yields more-accurate estimates. Using data from a vaccine study, they demonstrated that this method could lead to different practical conclusions. They also proposed a simplified version of the estimator, which is useful if the analyst cannot implement the inverse-variance method. This method is applicable to similar types of assays, not just ELISA.

In this article, we describe the inverse-variance weighted-average estimator proposed by Cheung et al. (2015), and we present a new command, `wavemid`, for implementing this procedure. We also present the new command `midreshape`, which processes raw data in text format exported from some microplate readers into analyzable data format. We use an experimental study of malaria vaccine candidates to illustrate the use of the two commands.

2 Weighted-average estimation approach

A typical microplate has 96 wells for an immunoassay measurement of up to 96 samples at the same time. Some of the wells hold standard samples whose true concentration levels are known. The remaining wells hold a set of (diluted) test samples with unknown concentrations. The outputs of the immunoassay include up to 96 observed OD values. To estimate unknown concentrations in the test samples, one must establish the standard curve that maps the observed response to the known concentrations by using the standard samples.

2.1 Estimation of the standard curve

Suppose one microplate holds n_s unique standard samples with known concentrations denoted by $x_{\text{Standard},1}, x_{\text{Standard},2}, \ldots, x_{\text{Standard},n_s}$. For the ith ($i = 1, \ldots, n_s$) standard sample, there are m_i replicate measurements. Typical laboratory practice is to have $m_i = 2$ or 3 for all i. The total number of replicated standard concentrations is $N_s = \sum_{i=1}^{n_s} m_i$. Let $y_{\text{Standard},i,j}$ denote the observed response for the jth replicate of a standard sample with concentration $x_{\text{Standard},i}$.

A common parameterization of the standard curve is a four-parameter logistic model,

$$E(Y) = Q(x|\mathbf{b}) = b_2 + \frac{b_1 - b_2}{1 + \left(\frac{x}{b_3}\right)^{b_4}} \tag{1}$$

where $E(Y)$ denotes the mean of response Y (for example, OD) at concentration x (O'Connell, Belanger, and Haaland 1993). In the parameter vector $\mathbf{b} = (b_1, b_2, b_3, b_4)^T$, b_1 and b_2 are, respectively, the lower and upper asymptotes of the standard curve as the concentration $x \to 0$ and $x \to \infty$. The parameter b_3 corresponds to the concentration at the midpoint of the two asymptotes, and b_4 is related to the slope of the standard curve. The variance of response Y is often formulated as a power-of-the-mean function, that is,

$$\text{Var}(Y) = \sigma^2 \{Q(x|\mathbf{b})\}^{2\theta} \tag{2}$$

where σ is the scale parameter and θ represents the degree of heteroskedasticity (Carroll and Ruppert 1988; Davidian and Haaland 1990). Expressions (1) and (2) together define the response model.

Given data for the standard samples $\{(y_{\text{Standard},i,j}, x_{\text{Standard},i}) : j = 1, \ldots, m_i; i = 1, \ldots, n_s\}$, a generalized least-squares (GLS) estimation algorithm may be used to fit the model parameters $\mathbf{b}$, σ, and θ. Further details of the GLS estimation algorithm can be found in Carroll and Ruppert (1988), Davidian and Haaland (1990), and O'Connell, Belanger, and Haaland (1993). Briefly, steps for this method are as follows:

1. Initialize the parameter $\mathbf{b}$ by an ordinary least-squares regression of $y_{\text{Standard},i,j}$ on $x_{\text{Standard},i}$ in the dataset. Denote the resultant estimate as $\widehat{\mathbf{b}}^{(s)}$ and $s = 0$.

2. Estimate (θ, σ) using the pseudolikelihood method, that is, by minimizing the following log-likelihood function in (θ, σ):

$$PL(\theta,\sigma) = N_s \log(\sigma) + \theta \sum_{i=1}^{n_s}\sum_{j=1}^{m_i} Q\left(x_{\text{Standard},i}|\widehat{\mathbf{b}}^{(s)}\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{n_s}\sum_{j=1}^{m_i}\left[\frac{y_{\text{Standard},i,j} - Q\left(x_{\text{Standard},i}|\widehat{\mathbf{b}}^{(s)}\right)}{\left\{Q\left(x_{\text{Standard},i}|\widehat{\mathbf{b}}^{(s)}\right)\right\}^{\theta}}\right]^2$$

   Denote the resultant estimate as $(\widehat{\theta}^{(s)}, \widehat{\sigma}^{(s)})$. Form the estimated weights:

$$\widehat{v}_{ij} = \left\{Q\left(x_{\text{Standard},i}|\widehat{\mathbf{b}}^{(s)}\right)\right\}^{-2\widehat{\theta}^{(s)}}$$

3. Use $\widehat{v}_{ij}$ from step 2 to update the estimate $\widehat{\mathbf{b}}^{(s)}$ that minimizes

$$\sum_{i=1}^{n_s}\sum_{j=1}^{m_i}\widehat{v}_{ij}\left\{y_{\text{Standard},i,j} - Q(x_{\text{Standard},i}|\mathbf{b})\right\}^2$$

   Set $s = s + 1$ and return to step 2.

   Iterate between steps 2 and 3 until the parameter estimate $\widehat{\theta}$ converges. Denote the resultant estimates for the model parameters as $\widehat{\mathbf{b}} = (\widehat{b}_1, \widehat{b}_2, \widehat{b}_3, \widehat{b}_4)^T$, $\widehat{\sigma}$, and $\widehat{\theta}$.

Upon convergence, the variance of estimate $\widehat{\mathbf{b}}$ can be readily derived as

$$\text{Var}\left(\widehat{\mathbf{b}}\right) = \widehat{\sigma}^2(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})^{-1}$$

where the matrix $\mathbf{X}$ is an $N_s \times 4$ gradient matrix. For this matrix, the columns are the partial derivative of the four-parameter logistic function $Q(x|b)$ with respect to each of the parameters in $\mathbf{b}$ evaluated at $\mathbf{b} = \widehat{\mathbf{b}}$, denoted by $\{\delta Q(x|\widehat{\mathbf{b}})/\delta\mathbf{b}\}^T$. The rows correspond to each of the N_s standard concentrations. The elements of row vector $\{\delta Q(x|\widehat{\mathbf{b}})/\delta\mathbf{b}\}^T$ are

$$\left\{\delta Q\left(x|\widehat{\mathbf{b}}\right)/\delta\mathbf{b}\right\}^T = \left\{\frac{1}{1+\left(\frac{x}{\widehat{b}_3}\right)^{\widehat{b}_4}}, \frac{\left(\frac{x}{\widehat{b}_3}\right)^{\widehat{b}_4}}{1+\left(\frac{x}{\widehat{b}_3}\right)^{\widehat{b}_4}}, \frac{\left(\frac{x}{\widehat{b}_3}\right)^{\widehat{b}_4}\left(\widehat{b}_1-\widehat{b}_2\right)\frac{\widehat{b}_4}{\widehat{b}_3}}{1+\left(\frac{x}{\widehat{b}_3}\right)^{\widehat{b}_4}}, -\frac{\left(\frac{x}{\widehat{b}_3}\right)^{\widehat{b}_4}\left(\widehat{b}_1-\widehat{b}_2\right)\log\frac{x}{\widehat{b}_3}}{1+\left(\frac{x}{\widehat{b}_3}\right)^{\widehat{b}_4}}\right\}$$

The first m_1 rows of the gradient matrix $\mathbf{X}$ are each $\{\delta Q(x_{\text{Standard},1}|\widehat{\mathbf{b}})/\delta\mathbf{b}\}^T$. The rows $(m_1 + 1)$ to $(m_1 + m_2)$ are each $\{\delta Q(x_{\text{Standard},2}|\widehat{\mathbf{b}})/\delta\mathbf{b}\}^T$, and so on. The matrix $\mathbf{G}$ is an $N_s \times N_s$ diagonal matrix, with the first m_1 diagonal elements equal to $\{Q(x_{\text{Standard},1}|\widehat{\mathbf{b}})\}^{2\widehat{\theta}}$, the $(m_1 + 1)$th to $(m_1 + m_2)$th diagonal elements equal to $\{Q(x_{\text{Standard},2}|\widehat{\mathbf{b}})\}^{2\widehat{\theta}}$, and so on.

2.2 Estimating concentration in test samples

Having established the standard curve, we now estimate the concentration in the test samples on the same microplate. Each original test sample with unknown concentration x is to be diluted in K steps. Typically, in immunoassays, $K = 4$. For the kth $(k = 1, \ldots, K)$ diluted sample, the underlying concentration is x/dil_k, where dil_k denotes the dilution factor. Often, $\text{dil}_{k+1}/\text{dil}_k = 2$ or 3, representing doubling or tripling of the dilutions. Let y_k denote the observed response, which can be the mean of r replicates. On the log scale, the estimate for the unknown concentration x based on its kth diluted sample (that is, the dilution-specific concentration estimate), denoted by $\widehat{x}_k$, is

$$\log(\widehat{x}_k) = \log\left(\widehat{b}_3\right) + \log(\text{dil}_k) + \frac{1}{\widehat{b}_4}\log\left(\frac{\widehat{b}_1 - y_k}{y_k - \widehat{b}_2}\right) \tag{3}$$

Cheung et al. (2015) proposed an inverse-variance weighted-average estimator of the log-transformed concentration, $\log(x)$, for the original serum sample, denoted as $\widehat{\log(x)}$,

$$\widehat{\log(x)} = \frac{1}{\sum_{k=1}^{K} w_k} \sum_{k=1}^{K} w_k \log(\widehat{x}_k) \tag{4}$$

where w_k is the weight assigned to the log-transformed, dilution-specific concentration estimate as follows:

$$w_k = \left[\left(\frac{1}{\widehat{b}_1 - y_k} + \frac{1}{y_k - \widehat{b}_2}\right)^2 y_k^{2\widehat{\theta}}/r + \widehat{b}_4^2 \left\{\frac{\delta \log(\widehat{x}_k)}{\delta \mathbf{b}}\right\}^T \left(\mathbf{X}^T \mathbf{G}^{-1} \mathbf{X}\right)^{-1} \left\{\frac{\delta \log(\widehat{x}_k)}{\delta \mathbf{b}}\right\}\right]^{-1} \tag{5}$$

Details of the derivation and interpretation of (2) are in Cheung et al. (2015). Briefly, (2) contains two additive terms. The first additive term involves a product of two elements. The first element contains the lower and upper asymptotes ($\widehat{b}_1$ and $\widehat{b}_2$). With other factors remaining the same, the closer the OD is to the midpoint of ($\widehat{b}_1$ and $\widehat{b}_2$) (the center of the optimal range), the heavier the weight it receives. The second element contains $\widehat{\theta}$, which reflects the degree of heteroskedasticity in the variance of the response. With other remaining factors the same, an OD value with higher variance receives a lower weight. The second additive term accounts for the uncertainty in estimating parameter vector $\mathbf{b}$ for the standard curve. It involves a column vector, $\{\delta \log(\widehat{x}_k)\}/\delta \mathbf{b}$, and two matrices, $\mathbf{X}$ and $\mathbf{G}$. The matrices $\mathbf{X}$ and $\mathbf{G}$ are both based on the standard samples and were previously defined in section 2.1. The column vector $\{\delta \log(\widehat{x}_k)\}/\delta \mathbf{b}$ is the partial derivative of $\log(\widehat{x}_k)$ for test sample (4) with respect to the parameter vector $\mathbf{b}$ evaluated at $\mathbf{b} = \widehat{\mathbf{b}}$.

$$\frac{\delta \log(\widehat{x}_k)}{\delta \mathbf{b}} = \frac{1}{\widehat{b}_4}\left\{\frac{1}{\widehat{b}_1 - y_k}, \frac{1}{y_k - \widehat{b}_2}, \frac{\widehat{b}_4}{\widehat{b}_3}, -\frac{1}{\widehat{b}_4}\log\left(\frac{\widehat{b}_1 - y_k}{y_k - \widehat{b}_2}\right)\right\}^T$$

The variance for the proposed inverse-variance weighted-average estimate is

$$\begin{aligned}\text{Var}\left\{\widehat{\log(x)}\right\} = &\frac{1}{\sum_{k=1}^{K} 1/\text{Var}\left\{\log\left(\widehat{x}_k\right)\right\}} \\ &+ \frac{2\widehat{\sigma}^2}{\left[\sum_{k=1}^{K} 1/\text{Var}\left\{\log\left(\widehat{x}_k\right)\right\}\right]^2} \sum\sum_{1\le j_1<j_2\le K} \frac{1}{\text{Var}\left\{\log\left(\widehat{x}_{j_1}\right)\right\}} \frac{1}{\text{Var}\left\{\log\left(\widehat{x}_{j_2}\right)\right\}} \\ &\left\{\frac{\delta\log\left(\widehat{x}_{j_1}\right)}{\delta\widehat{\mathbf{b}}}\right\}^T \left(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X}\right)^{-1} \left\{\frac{\delta\log\left(\widehat{x}_{j_2}\right)}{\delta\widehat{\mathbf{b}}}\right\} \end{aligned} \tag{6}$$

In (4), the term Var{log($\widehat{x}_k$)} is the variance of the dilution-specific estimate and can be estimated from a bivariate Taylor-series expansion of (4) with respect to y_k and $\widehat{\mathbf{b}}$, which results in

$$\begin{aligned}\text{Var}\left\{\log\left(\widehat{x}_k\right)\right\} = &\left\{\frac{1}{\widehat{b}_4}\left(\frac{1}{\widehat{b}_1 - y_k} + \frac{1}{y_k - \widehat{b}_2}\right)\right\}^2 \widehat{\sigma}^2 y_k^{2\widehat{\theta}}/r \\ &+ \left\{\frac{\delta\log\left(\widehat{x}_k\right)}{\delta\widehat{\mathbf{b}}}\right\}^T \text{Var}\left(\widehat{\mathbf{b}}\right) \left\{\frac{\delta\log(\widehat{x}_k)}{\delta\widehat{\mathbf{b}}}\right\}\end{aligned}$$

3 The wavemid command

3.1 Syntax

`wavemid` [*if*] [*in*], `testsample(`*varname*`)` `od(`*varname*`)` `standard(`*varname*`)`
`dilutionfactor(`*varname*`)` `id(`*varname*`)` `saving(`*filename*[, `replace`]`)`
[`iterate(`#`)` `tolerance(`#`)` `plot`
`graphexport(`*graphfilename*.*suffix*[, `replace`]`)`]

3.2 Options

`testsample(`*varname*`)` specifies a binary variable in the dataset that takes the value `1` for test samples and `0` for standard samples. `testsample()` is required.

`od(`*varname*`)` specifies a variable corresponding to the OD in the dataset. `od()` is required.

`standard(`*varname*`)` specifies a variable in the dataset that contains the concentration of the standard sample and the missing value for the test sample. `standard()` is required.

`dilutionfactor(`*varname*`)` specifies a variable in the dataset that contains the dilution factor of the test sample and the missing value for the standard sample. `dilutionfactor()` is required.

`id(`*varname*`)` specifies a variable in the dataset that identifies the samples. `id()` is required.

`saving(`*filename*[`, replace`]`)` creates an output data file (*filename*`.dta`) that contains the sample identifier, the concentration estimate, the variance of the log-transformed concentration estimate, the dilution-specific mean ODs among the replicates, and the dilution-specific weight. Use `replace` to overwrite the existing *filename*`.dta`. `saving()` is required.

`iterate(`#`)` specifies the maximum number of iterations for GLS estimation. When the number of iterations equals `iterate()`, estimation stops and presents the current results. If convergence is declared before this threshold is reached, estimation will stop when convergence is declared. The default is `iterate(10)`.

`tolerance(`#`)` specifies the tolerance for the parameter theta in the power-of-the-mean variance function. When the relative change in the parameter theta from one iteration to the next is less than or equal to `tolerance()`, the `tolerance()` convergence criterion is satisfied. The default is `tolerance(10^(-4))`.

`plot` plots the standard curve with the observed data for the standard samples on the logarithm scale for the concentration.

`graphexport(`*graphfilename*`.`*suffix*[`, replace`]`)` exports the graph that is generated after `plot` to *graphfilename*`.`*suffix*. Use `replace` to overwrite the existing *graphfilename*`.`*suffix*.

4 The midreshape command

Immunoassay data can be directly entered into Stata without using `midreshape`. However, some microplate readers export raw data and sample labels as a text file in a popular format. A typical layout of the text file includes the following three panels of data: 1) sample identifiers, 2) ODs, and 3) concentrations for standard samples and dilution factors for test samples. A name is shown in the line above each panel to describe what that panel is. Within each panel, the data values are exhibited as an 8×12 matrix corresponding to the 96-well microplate layout. See section 5 for an example. The `midreshape` command converts the text file into Stata and reshapes the data to the format required by `wavemid`. Users of `midreshape` should carefully examine whether their raw data file is in the format described here.

`midreshape using` *filename*`, template(`*string*`) od(`*string*`) concentration(`*string*`)`
`test(`*string*`) standard(`*string*`)` [`separator(`*list_separator*`)`]

4.1 Options

`template(`*string*`)` specifies the name of the panel corresponding to the sample identifiers. Space characters in the panel name may be safely ignored. This option is not case

sensitive. For instance, if the panel name is `Template Name`, then you may specify `template(Template Name)`, `template(TemplateName)`, `template(Templatename)`, `template(templateName)`, or `template(templatename)`. `template()` is required.

`od(`*string*`)` specifies the name of the panel corresponding to the ODs. Space characters in the panel name may be safely ignored. This option is not case sensitive. `od()` is required.

`concentration(`*string*`)` specifies the name of the panel that corresponds to the concentrations (for the standard samples) and dilution factors (for the test samples). Space characters in the panel name may be safely ignored. This option is not case sensitive. `concentration()` is required.

`test(`*string*`)` specifies the prefix of the sample identifiers to indicate which are the test samples. Multiple prefixes, separated by space characters, may be specified. This option is case sensitive. `test()` is required.

`standard(`*string*`)` specifies the prefix of the sample identifiers to indicate which are the standard samples. Multiple prefixes, separated by space characters, may be specified. This option is case sensitive. `standard()` is required.

`separator(`*list_separator*`)` specifies the character to be used to separate the values. `separator(comma)` specifies that values be comma-separated. `separator(tab)` specifies that values be tab-separated. Users may also specify other separation characters. For instance, if values in the file are separated by a semicolon, the user may specify `separator(";")`. The default is `separator(tab)`.

5 Example

5.1 Preparing an analysis dataset from a raw text file exported from an optical reader

The example dataset we use here is part of an experimental study that immunized rabbits with one of four malaria vaccine candidates. Here we used the data on antibodies to Apical Membrane Antigen 1 from one ELISA microplate. Apical Membrane Antigen 1 is an antigen that plays an important role in the invasion of red blood cells and hepatocytes by the parasites.

The 96-well microplate follows the typical 8-row by 12-column format. Each plate includes 2 blank samples (negative controls), 14 standard samples (seven concentrations with two replicates each), and 80 test samples. Each test sample begins with a starting dilution of 1:24000, followed by tripling dilutions, to obtain four diluted samples per serum sample. The data were stored in a text file, `example_12_tab.txt`, exported by a microplate reader. In the text file, the sample identifiers are shown in the following format:

```
Template
BLK   STD04 SMP05 SMP20 SMP29 SMP13 SMP22 SMP05 SMP20 SMP29 SMP13 SMP22
BLK   STD04 SMP05 SMP20 SMP29 SMP13 SMP22 SMP05 SMP20 SMP29 SMP13 SMP22
STD07 STD03 SMP06 SMP21 SMP30 SMP14 SMP28 SMP06 SMP21 SMP30 SMP14 SMP28
STD07 STD03 SMP06 SMP21 SMP30 SMP14 SMP28 SMP06 SMP21 SMP30 SMP14 SMP28
STD06 STD02 SMP13 SMP22 SMP05 SMP20 SMP29 SMP13 SMP22 SMP05 SMP20 SMP29
STD06 STD02 SMP13 SMP22 SMP05 SMP20 SMP29 SMP13 SMP22 SMP05 SMP20 SMP29
STD05 STD01 SMP14 SMP28 SMP06 SMP21 SMP30 SMP14 SMP28 SMP06 SMP21 SMP30
STD05 STD01 SMP14 SMP28 SMP06 SMP21 SMP30 SMP14 SMP28 SMP06 SMP21 SMP30
```

`BLK` indicates blank samples, and the prefixes `STD` and `SMP` differentiate the standard samples from the test samples, respectively. The duplicated numeric suffixes indicate two replicates per standard sample and per diluted test sample. All values are tab-delimited. The raw ODs corresponding to these samples are displayed in the text file following the same matrix format:

```
Raw Data
0.070 0.740 2.976 2.362 2.229 2.339 1.731 1.103 0.656 0.527 0.825 0.460
0.068 0.824 2.728 2.712 1.792 2.652 1.682 1.279 0.727 0.480 0.816 0.400
0.112 1.389 2.427 2.619 3.165 1.934 1.871 0.982 0.816 1.028 0.570 0.568
0.125 1.360 2.617 2.158 2.942 2.085 1.986 0.931 0.703 1.143 0.458 0.564
0.185 2.176 2.882 2.497 2.130 1.453 1.177 1.809 0.908 0.519 0.359 0.241
0.193 2.277 3.126 2.410 1.958 1.402 1.274 1.400 0.971 0.526 0.334 0.276
0.387 2.798 2.640 2.853 1.719 1.318 1.961 1.251 1.007 0.456 0.321 0.509
0.304 2.775 2.937 2.806 1.826 1.596 2.153 1.166 1.039 0.405 0.341 0.568
```

This is followed by a third data matrix that shows concentrations for the standard samples and dilution factors for the test samples:

```
Concentrations / Dilutions
       1.9235  24000 24000 24000 72000 72000 216000 216000 216000 648000 648000
       1.9235  24000 24000 24000 72000 72000 216000 216000 216000 648000 648000
0.0712 5.7704  24000 24000 24000 72000 72000 216000 216000 216000 648000 648000
0.0712 5.7704  24000 24000 24000 72000 72000 216000 216000 216000 648000 648000
0.2137 17.3111 24000 24000 72000 72000 72000 216000 216000 648000 648000 648000
0.2137 17.3111 24000 24000 72000 72000 72000 216000 216000 648000 648000 648000
0.6411 51.9333 24000 24000 72000 72000 72000 216000 216000 648000 648000 648000
0.6411 51.9333 24000 24000 72000 72000 72000 216000 216000 648000 648000 648000
```

The first two elements in the first column of the above matrix were blank, corresponding to the two blank samples as seen in the `Template` block. Each test sample (for example, `SMP05`) occupied eight wells, that is, two replicates in each of the four dilution levels.

The blank samples on microplates are usually used for quality control purposes, and thus we do not include the two blank samples in our statistical analysis. We now use the `midreshape` command to prepare the dataset for further analysis.

```
. midreshape using example_12_tab.txt, template(Template) od(Raw Data)
> concentration(Concentrations / Dilutions) test(SMP) standard(STD)
```

The command creates five variables: `testSample` (`1` if a test sample and `0` if a standard sample), `ID`, `OD`, `standard` (missing if `testSample` = `1`), and `dilution_factor` (missing if `testSample` = `0`). Data are sorted and shown in ascending order for

`testSample`, `ID`, and `dilution_factor`. The table below shows the actual data for the first 20 rows.

```
. list in 1/20, abbreviate(20) separator(20)
```

	testSample	ID	OD	standard	dilution_factor
1.	0	STD01	2.798	51.9333	.
2.	0	STD01	2.775	51.9333	.
3.	0	STD02	2.176	17.3111	.
4.	0	STD02	2.277	17.3111	.
5.	0	STD03	1.389	5.7704	.
6.	0	STD03	1.36	5.7704	.
7.	0	STD04	.824	1.9235	.
8.	0	STD04	.74	1.9235	.
9.	0	STD05	.387	.6411	.
10.	0	STD05	.304	.6411	.
11.	0	STD06	.185	.2137	.
12.	0	STD06	.193	.2137	.
13.	0	STD07	.125	.0712	.
14.	0	STD07	.112	.0712	.
15.	1	SMP05	2.976	.	24000
16.	1	SMP05	2.728	.	24000
17.	1	SMP05	1.958	.	72000
18.	1	SMP05	2.13	.	72000
19.	1	SMP05	1.279	.	216000
20.	1	SMP05	1.103	.	216000

5.2 Implementing the weighted-average estimation approach using the wavemid command

Now we apply the `wavemid` command to the data described previously.

```
. wavemid, testsample(testSample) od(OD) standard(standard)
> dilutionfactor(dilution_factor) id(ID) saving(plate_estimate, replace)
> plot graphexport(standard_curve.png,replace)
file plate_estimate.dta saved
(file standard_curve.png written in PNG format)
Estimation of Standard Curve Based on Standard Samples.................
```

LOF: Source	SS	df	MS
Pure Error	.012874484	7	.001839212
Lack of Fit	.007700135	3	.002566712
Error	.020574615	10	.002057462

```
Number of obs  =        14
F( 3,    7)    =    1.3955
Prob   >   F   =    0.3215
R-squared      =  .9984546
Root MSE       =  .0014696
```

	Est	Std Err
b1	0.0742918	0.1485931
b2	3.3345455	0.0771453
b3	8.6078653	2.3250908
b4	0.8994575	0.0428863
theta	0.4502218	
sigma	0.0418301	

The regression table above gives the parameter estimates for $\mathbf{b}$, σ, and θ. A lack-of-fit test is also conducted to assess how well this estimated standard curve fit to the observations from standard samples (O'Connell, Belanger, and Haaland 1993). This is a common practice in immunoassays. The residual sum of squares error is partitioned into two components: the sum of squares due to pure error (SSPE) and the sum of squares due to lack of fit (SSLOF). Further details on how to calculate the sum of squares error, SSPE, and SSLOF can be found elsewhere (see, for example, Brook and Arnold [1985, 48–49]).

Then, a variance-ratio F test can be conducted using the test statistic SSLOF/SSPE, which follows an F distribution with degrees of freedom $d_1 = n_s - p$ and $d_2 = N_s - n_s$, where n_s is the number of unique concentrations in the standard samples, p is the number of parameters in the standard curve mean response model, and N_s is the total number of standard samples on the microplate. In this example, $n_s = 7$, $p = 4$ (that is, a four-parameter logistic model), and $N_s = 14$. Thus, $d_1 = 3$ and $d_2 = 7$, hence, SSLOF/SSPE $\sim F_{3,7}$. The lack-of-fit test shown in the above table gave an insignificant p-value (0.3215), which suggests that the null hypothesis of lack of fit could not be rejected at the 5% significance level. The R-squared value being close to 1 also indicated that the estimated standard curve fit the data well on the standard samples. This is further demonstrated graphically in figure 2, which is produced by the `plot` option and depicts the estimated standard curve on the log(concentration) scale together with the scatterplot of the observed data on the standard samples. By specifying the `graphexport()` option, the generated graph is saved to an external file.

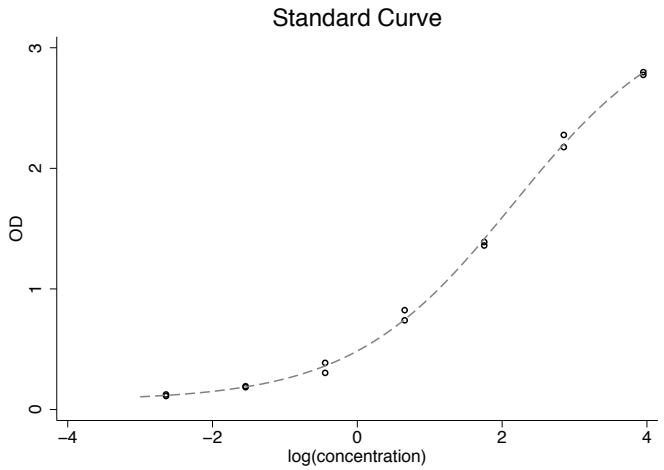

Figure 2. Plot of standard curve and observed data generated by `wavemid`

The results on concentration estimates for the test samples are saved in `plate_estimate.dta`.

```
. use plate_estimate, clear
. describe
Contains data from plate_estimate.dta
  obs:            10
 vars:            15                          16 Nov 2015 11:33
 size:           610
```

variable name	storage type	display format	value label	variable label
ID	str5	%9s		Sample ID
CONC	float	%9.0g		Weighted average estimate
VAR_logCONC	float	%9.0g		Var(ln(CONC))
dilution_fact~1	float	%9.0g		Dilution factor
OD_avg1	float	%9.0g		Mean of responses
Wt_log1	float	%9.0g		Dilution-specific weight
dilution_fact~2	float	%9.0g		Dilution factor
OD_avg2	float	%9.0g		Mean of responses
Wt_log2	float	%9.0g		Dilution-specific weight
dilution_fact~3	float	%9.0g		Dilution factor
OD_avg3	float	%9.0g		Mean of responses
Wt_log3	float	%9.0g		Dilution-specific weight
dilution_fact~4	float	%9.0g		Dilution factor
OD_avg4	float	%9.0g		Mean of responses
Wt_log4	float	%9.0g		Dilution-specific weight

```
Sorted by: ID  CONC  VAR_logCONC
```

The variables `CONC` and `VAR_logCONC` are calculated based on (5) and (4), respectively. The variables `Wt_log1` to `Wt_log4` are the weights at each of the four dilution levels, each calculated based on (2). Below is a list of these data for each of the 10 test samples.

```
. list ID CONC VAR_logCONC Wt_log*, separator(10)
```

	ID	CONC	VAR_lo~C	Wt_log1	Wt_log2	Wt_log3	Wt_log4
1.	SMP05	949363.4	.0056274	.1191221	.1641234	.6275659	.0891886
2.	SMP06	630065.2	.0033245	.0823517	.1337856	.7578192	.0260436
3.	SMP13	1578578	.0030945	.0703939	.0967548	.1928905	.6399609
4.	SMP14	939857.7	.0060724	.1258611	.1737726	.6145999	.0857664
5.	SMP20	433232.4	.0026274	.1522356	.4140148	.411734	.0220156
6.	SMP21	445023.4	.0024643	.1270483	.3108616	.5471681	.0149219
7.	SMP22	602859.1	.0031466	.082595	.1418577	.7500857	.0254616
8.	SMP28	758736.4	.0034011	.0812947	.1216422	.7110945	.0859686
9.	SMP29	305864.5	.0045231	.2056204	.6929231	.0930961	.0083604
10.	SMP30	861425.4	.0037226	.0629233	.1337282	.7196823	.0836661

6 Conclusion

In immunoassays, accurate measurement of concentration using a standard curve requires the use of a dilution that ensures the OD is in the near-linear part of the curve. Analyzing only one dilution wastes useful information and decreases measurement accuracy. We have, therefore, developed an approach that allows data from several dilutions of each sample to be used in a weighted analysis, which gives improved estimates of the substance concentration. The superiority of this approach versus the conventional approach was previously demonstrated by Cheung et al. (2015).

In this article, we demonstrated how to implement inverse-variance weighted-average estimation to analyze data on multiple dilutions by using the `wavemid` command. We also discussed how the `midreshape` command can be used to convert data from the format commonly provided by a microplate reader. These commands are useful when using Stata in laboratory-based biomedical research.

7 Acknowledgments

This work was supported by the National Research Foundation in Singapore, under its Clinician Scientist Award (Award No. NMRC/CSA/039/2012) administered by the Singapore Ministry of Health's National Medical Research Council.

8 References

Brook, R. J., and G. C. Arnold. 1985. *Applied Regression Analysis and Experimental Design*. New York: CRC Press.

Carroll, R. J., and D. Ruppert. 1988. *Transformation and Weighting in Regression*. New York: Chapman & Hall.

Cheung, Y. B., Y. Xu, E. J. Remarque, and P. Milligan. 2015. Statistical estimation of antibody concentration using multiple dilutions. *Journal of Immunological Methods* 417: 115–123.

Davidian, M., and P. D. Haaland. 1990. Regression and calibration with nonconstant error variance. *Chemometrics and Intelligent Laboratory Systems* 9: 231–248.

O'Connell, M. A., B. A. Belanger, and P. D. Haaland. 1993. Calibration and assay development using the four-parameter logistic model. *Chemometrics and Intelligent Laboratory Systems* 20: 97–114.

Ratkowsky, D. A., and T. J. Reedy. 1986. Choosing near-linear parameters in the four-parameter logistic model for radioligand and related assays. *Biometrics* 42: 575–582.

Remarque, E. J. 2007. Auditable data analysis and management system for ELISA (ADAMSEL-v1.1). http://www.transvac.org/sites/default/files/uploads/docs/Projects/OPTIMALVAC/ADAMSEL_30_Manual.pdf.

About the authors

Ying Xu is an assistant professor in the Center for Quantitative Medicine at Duke–NUS Medical School in Singapore. Her research interests include statistical methodology related to infectious disease and human growth as well as design and analysis of clinical trials.

Paul Milligan is a reader in epidemiology and medical statistics at the London School of Hygiene and Tropical Medicine. His research focuses on the epidemiology and control of malaria and statistical methods with application to infectious disease research.

Edmond J. Remarque is a principal investigator at the Biomedical Primate Research Center in the Netherlands. His research interests include immunology, vaccines, and statistics. He is the developer of the Auditable Data Analysis and Management System for ELISA.

Yin Bun Cheung is a professor in the Center for Quantitative Medicine at Duke–NUS Medical School in Singapore, and he is an adjunct professor in the Department for International Health at the University of Tampere in Finland. His research areas include statistical methods for analysis of vaccine efficacy and immunogenicity as well as the impact and interplay of infection and undernutrition on child health in developing countries.