# Contents

# Chapter 1

# Formal Models and Tests

## 1.1 Formal Models and Tests

The Bland-Altman plot is a simple tool for inspection of data, and Kinsella (1986) comments on the lack of formal testing offered by that approach, instead relying on the practitioner's opinion to judge the outcome of the technique. A test proposed by Altman and Bland (1983) is a formal test on the Pearson correlation coefficient of case-wise differences and means. According to the authors, this test is equivalent to the 'Pitman-Morgan Test' a key contribution on the subject which shall discussed shortly (Morgan, 1939; Pitman, 1939). There has been no further mention of this particular test in Bland and Altman (1986), although Bland and Altman (1999) refers to Spearman's rank correlation coefficient. Bland and Altman (1999) commented 'we do not see a place for methods of analysis based on hypothesis testing', while also stating that they consider structural equation models to be inappropriate.

### Kinsella's Model

Kinsella (1986) presented a simple model to describe a measurement by method $m$, describing the relationship with its real value. Only the non-replicate case is considered, as it is the context of the Bland-Altman plots. Other authors, such as Carstensen

(2004); Carstensen et al. (2008), present similar formulations of the same model, as well as modified models to account for multiple measurements by each methods on each item, known as replicate measurements.

Kinsella (1986) formulates a model for single measurement observations for a method comparison study as a linear mixed effects model, i.e. model that additively combine fixed effects and random effects.

$$Y_{ij} = \quad \mu + \beta_j + u_i + \epsilon_{ij} \qquad i = 1, \dots, n \qquad j = 1, 2$$

The true value of the measurement is represented by $\mu$ while the fixed effect due to method $j$ is $\beta_j$. For simplicity these terms can be combined into single terms; $\mu_1 = \mu + \beta_1$ and $\mu_2 = \mu + \beta_2$. The inter-method bias is the difference of the two fixed effect terms, $\beta_1 - \beta_2$. Each of the $i$ individuals are assumed to give rise to random error, represented by $u_i$. This random effects terms is assumed to have mean zero and be normally distributed with variance $\sigma^2$. There is assumed to be an attendant error for each measurement on each individual, denoted $\epsilon_{ij}$. This is also assumed to have mean zero. The variance of measurement error for both methods are not assumed to be identical for both methods variance, hence it is denoted $\sigma^2_j$. The set of observations $(x_i, y_i)$ by methods $X$ and $Y$ are assumed to follow the bivariate normal distribution with expected values $E(x_i) = \mu_1$ and $E(y_i) = \mu_2$ respectively. The variance covariance of the observations $\Sigma$ is given by

$$\Sigma = \begin{bmatrix} \sigma^2 + \sigma^2_1 & \sigma^2 \\ \sigma^2 & \sigma^2 + \sigma^2_2 \end{bmatrix}.$$

The case-wise differences and means are calculated as $d_i = x_i - y_i$ and $a_i = (x_i + y_i)/2$ respectively. Both $d_i$ and $a_i$ are assumed to follow a bivariate normal distribution with $E(d_i) = \mu_d = \mu_1 - \mu_2$ and $E(a_i) = \mu_a = (\mu_1 + \mu_2)/2$. The variance matrix $\Sigma_{(a,d)}$ is

$$\Sigma_{(a,d)} = \begin{bmatrix} \sigma^2_1 + \sigma^2_2 & \frac{1}{2}(\sigma^2_1 - \sigma^2_2) \\ \frac{1}{2}(\sigma^2_1 - \sigma^2_2) & \sigma^2 + \frac{1}{4}(\sigma^2_1 + \sigma^2_2) \end{bmatrix}. \tag{1.1}$$

Likewise the separate $\alpha$ can not be estimated, only their difference can be estimated as $\bar{d}$ (i.e. the inter-method bias). This model implies that the difference between the paired measurements can be expressed as

$$d_i = y_{1i} - y_{2i} \sim \mathcal{N}(\alpha_1 - \alpha_2, \sigma_1^2 + \sigma_2^2).$$

Importantly, this is independent of the item levels $\mu_i$. As the case-wise differences are of interest, the parameters of interest are the fixed effects for methods $\alpha_m$.

Kinsella (1986) demonstrates the estimation of the variance terms and relative precisions relevant to a method comparison study, with attendant confidence intervals for both. The measurement model introduced by Grubbs (1948, 1973) provides a formal procedure for estimate the variances $\sigma^2$, $\sigma_1^2$ and $\sigma_2^2$ devices. Grubbs (1948) offers maximum likelihood estimates, commonly known as Grubbs estimators, for the various variance components,

$$\hat{\sigma^2} = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} = Sxy,$$
$$\hat{\sigma_1^2} = \sum \frac{(x_i - \bar{x})^2}{n-1} = S^2x - Sxy,$$
$$\hat{\sigma_2^2} = \sum \frac{(y_i - \bar{y})^2}{n-1} = S^2y - Sxy.$$

Thompson (1963) presents confidence intervals for the relative precisions of the measurement methods, $\Delta_j = \sigma_S^2/\sigma_j^2$ (where $j = 1,2$), as well as the variances $\sigma_S^2, \sigma_1^2$ and $\sigma_2^2$,

$$\Delta_1 > \quad \frac{C_{xy} - t(|A|/n-2))^{\frac{1}{2}}}{C_x - C_{xy} + t(|A|/n-2))^{\frac{1}{2}}}. \tag{1.2}$$

Thompson (1963) defines $\Delta_j$ to be a measure of the relative precision of the measurement methods, with $\Delta_j = \sigma^2/\sigma_j^2$. Thompson also demonstrates how to make statistical inferences about $\Delta_j$. Based on the following identities,

$$\begin{aligned} C_x &= (n-1)S_x^2, \\ C_{xy} &= (n-1)S_{xy}, \\ C_y &= (n-1)S_y^2, \\ |A| &= C_x \times C_y - (C_{xy})^2, \end{aligned}$$

the confidence interval limits of $\Delta_1$ are

$$\Delta_1 > \frac{C_{xy} - t(\frac{|A|}{n-2}))^{\frac{1}{2}}}{C_x - C_{xy} + t(\frac{|A|}{n-2}))^{\frac{1}{2}}} \tag{1.3}$$

$$\Delta_1 > \frac{C_{xy} + t(\frac{|A|}{n-2}))^{\frac{1}{2}}}{C_x - C_{xy} - t(\frac{|A|}{n-1}))^{\frac{1}{2}}}$$

The value $t$ is the $100(1-\alpha/2)\%$ upper quantile of Student's $t$ distribution with $n-2$ degrees of freedom (Kinsella, 1986). The confidence limits for $\Delta_2$ are found by substituting $C_y$ for $C_x$ in (1.3). Negative lower limits are replaced by the value 0.

### 1.1.1 Conversion Problems

In some types of analysis, such as the conversion problems described by Lewis et al. (1991), an estimate for the scaling factor $\beta$ may also be sought. For the time being, we will restrict ourselves to problems where $\beta$ is assumed to be 1.

$$X_i = \tau_i + \delta_i, \qquad \delta_i \sim \mathcal{N}(0, \sigma^2_\delta) \tag{1.4}$$

$$Y_i = \alpha + \beta\tau_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2_\epsilon) \tag{1.5}$$

In this formulation, $\alpha$ represents the inter-method bias, and can be estimated as $E(X-Y)$. That is to say, a simple estimate of the inter-method bias is given by the differences between pairs of measurements.

### 1.1.2 Pitman-Morgan Testing

An early contribution to formal testing in method comparison was made by both Morgan (1939) and Pitman (1939), in separate contributions. Correlation between differences and means is a test statistics for the null hypothesis of equal variances given bivariate normality.

This test assess the equality of population variances by testing for zero correlation between the sums and products. The basis of this approach is that the distribution of the original measurements is bivariate normal.

These authors noted that the correlation coefficient depends upon the difference $\sigma_1^2 - \sigma_2^2$, being zero if and only if $\sigma_1^2 = \sigma_2^2$. Therefore a test of the hypothesis $H : \sigma_1^2 = \sigma_2^2$ is equivalent to a test of the hypothesis $H : \rho(a, d) = 0$. This corresponds to the well-known $t-$test for a correlation coefficient with $n - 2$ degrees of freedom.

The Pitman-Morgan test for equal variances is based on the correlation of $D$ with $S$. The correlation coefficient is zero if, and only if, the variances are equal. The test statistic is the familiar t-test with $n - 2$ degree of freedom.

The test of the hypothesis that the variances $\sigma_1^2$ and $\sigma_2^2$ are equal, which was devised concurrently by Pitman (1939) and Morgan (1939), is based on the correlation of the casewise-differences and sums, $d$ with $s$, the coefficient being $\rho_{(d,s)} = (\sigma_1^2 - \sigma_2^2)/(\sigma_D \sigma_S)$, which is zero if, and only if, $\sigma_1^2 = \sigma_2^2$.

The classical Pitman-Morgan test can be adapted as a hypothesis test of equal variance for both methods, based on the correlation value $\rho_{a,d}$;

$$\rho(a, d) = \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(4\sigma_S^2 + \sigma_1^2 + \sigma_2^2)}}. \tag{1.6}$$

Bartko (1994) describes the Pitman-Morgan test as identical to the test of the slope equal to zero in the regression of $Y_{i1}$ on $Y_{12}$, a result that can be derived using straightforward algebra. The Pitman-Morgan test is equivalent to the marginal test of the slope estimate in Bradley and Blackwood (1989).

Bartko (1994) discusses the use of the well known paired sample $t$ test to test for inter-method bias; $H : \mu_d = 0$. The test statistic is distributed a $t$ random variable with $n - 1$ degrees of freedom. Only if the two methods show comparable precision then the paired sample Student $t-$test is appropriate for testing the inter-method bias. Therefore, it should only be used in succession to the Pitman-Morgan test. Furthermore, these tests are only valid in the case of non-replicate measurements.

## 1.2 Bradley Blackwood

Bradley and Blackwood (1989) construct the conditional expectation of D given S as linear model. They used this result to propose a test of the joint hypothesis of the mean difference and equal variances. If the intercept and slope estimates are zero, the two methods have the same mean and variance.

Bradley and Blackwood (1989) offers a formal simultaneous hypothesis test for the mean and variance of two paired data sets. Using simple linear regression of the differences of each pair against the sums, a line is fitted to the model, with estimates for intercept and slope ($\hat{\beta}_0$ and $\hat{\beta}_1$). The null hypothesis of this test is that the mean ($\mu$) and variance ($\sigma^2$) of both data sets are equal if the slope and intercept estimates are equal to zero(i.e $\sigma_1^2 = \sigma_2^2$ and $\mu_1 = \mu_2$ if and only if $\beta_0 = \beta_1 = 0$ )

This technique offers a formal simultaneous hypothesis test for the mean and variance of two paired data sets. The null hypothesis of this test is that the mean ($\mu$) and variance ($\sigma^2$) of both data sets are equal if the slope and intercept estimates are equal to zero(i.e $\sigma_1^2 = \sigma_2^2$ and $\mu_1 = \mu_2$ if and only if $\beta_0 = \beta_1 = 0$).

Both regression coefficients are derived from the respective means and standard deviations of their respective data sets. We determine if the respective means and variances are equal if both beta values are simultaneously equal to zero. The test is conducted using an F test, calculated from the results of a regression of D on M.

A test statistic is then calculated from the regression analysis of variance values (Bradley and Blackwood, 1989) and is distributed as '$F$' random variable. The degrees of freedom are $\nu_1 = 2$ and $\nu_1 = n - 2$ (where $n$ is the number of pairs). The critical value is chosen for $\alpha\%$ significance with those same degrees of freedom.

Bartko (1994) amends this approach for use in method comparison studies, using the averages of the pairs, as opposed to the sums, and their differences. This approach can facilitate simultaneous usage of test with the Bland-Altman technique.

Bartko's test statistic take the form:

$$F^* = \frac{(\Sigma d^2) - SSReg}{2MSReg}.$$

For the Grubbs data, $\Sigma d^2 = 5.09$, $SSReg = 0.60$ and $MSreg = 0.06$ Therefore the test statistic is 37.42, with a critical value of 4.10. Hence the means and variance of

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Averages | 1 | 0.04 | 0.04 | 0.74 | 0.4097 |
| Residuals | 10 | 0.60 | 0.06 | | |

Table 1.2.1: Regression ANOVA of case-wise differences and averages for Grubbs Data

the Fotobalk and Counter chronometers are assumed to be simultaneously equal.

Importantly, this approach determines whether there is both inter-method bias and precision present, or alternatively if there is neither present. It has previously been demonstrated that there is a inter-method bias present, but as this procedure does not allow for separate testing, no conclusion can be drawn on the comparative precision of both methods.

### 1.2.1 Blackwood-Bradley Model

Bradley and Blackwood (1989) have developed a regression based procedure for assessing the agreement. This approach performs a simultaneous test for the equivalence of means and variances of the respective methods. The Bradley Blackwood test is a simultaneous test for bias and precision. They propose a regression approach which fits $D$ on $A$, where $D$ is the difference and average of a pair of results.

$$D = (X_1 - X_2) \tag{1.7}$$

$$A = (X_1 + X_2)/2 \tag{1.8}$$

Using simple linear regression of the differences of each pair against the sums, a line is fitted to the model, with estimates for intercept and slope ($\hat{\beta}_0$ and $\hat{\beta}_1$).

The Bradley-Blackwood procedure fits $D$ on $A$ as follows:

$$D = \beta_0 + \beta_1 A \tag{1.9}$$

### 1.2.2 Model for Replicate Measurements

We generalize the single measurement model for the replicate measurement case, by additionally specifying replicate values. Let $y_{mir}$ be the $r-$th replicate measurement for item $i$ made by method $m$. Further to Barnhart et al. (2007) fixed effect can be expressed with a single term $\alpha_{mi}$, which incorporate the true value $\mu_i$.

$$y_{mir} = \mu_i + \alpha_m + e_{mir}$$

Combining fixed effects (Barnhart et al., 2007), we write,

$$y_{mir} = \alpha_{mi} + e_{mir}.$$

The following assumptions are required $e_{mir}$ is independent of the fixed effects with mean $\mathrm{E}(e_{mir}) = 0$. Further to Barnhart et al. (2007) between-item and within-item variances $\mathrm{Var}(\alpha_{mi}) = \sigma^2_{Bm}$ and $\mathrm{Var}(e_{mir}) = \sigma^2_{Wm}$

### 1.2.3 Statistical Model For Replicate Measurements

Let $y_{Aij}$ and $y_{Bij}$ be the $j$th repeated observations of the variables of interest $A$ and $B$ taken on the $i$th item. The number of repeated measurements for each variable may differ for each individual. Both variables are measured on each time points. Let $n_i$ be the number of observations for each variable, hence $2 \times n_i$ observations in total.

It is assumed that the pair $y_{Aij}$ and $y_{Bij}$ follow a bivariate normal distribution.

$$\begin{pmatrix} y_{Aij} \\ y_{Bij} \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma) \text{ where } \mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \tag{1.10}$$

The matrix $\Sigma$ represents the variance component matrix between response variables at a given time point $j$.

$$\Sigma = \begin{pmatrix} \sigma^2_A & \sigma_{AB} \\ \sigma_{AB} & \sigma^2_B \end{pmatrix} \tag{1.11}$$

$\sigma_A^2$ is the variance of variable $A$, $\sigma_B^2$ is the variance of variable $B$ and $\sigma_{AB}$ is the covariance of the two variable. It is assumed that $\Sigma$ does not depend on a particular time point, and is the same over all time points.

### 1.2.4 Carstensen's Model for Replicate Measurements

Carstensen et al. (2008) develop their model from a standard two-way analysis of variance model, reformulated for the case of replicate measurements, with random effects terms specified as appropriate. For the replicate case, an interaction term $c$ is added to the model, with an associated variance component. Their model describing $y_{mir}$, again the $r$th replicate measurement on the $i$th item by the $m$th method ($m = 1, 2$, $i = 1, \dots, N$, and $r = 1, \dots, n$), can be written as

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \epsilon_{mir}. \tag{1.12}$$

The fixed effects $\alpha_m$ and $\mu_i$ represent the intercept for method $m$ and the 'true value' for item $i$ respectively. The random-effect terms comprise an item-by-replicate interaction term $a_{ir} \sim \mathcal{N}(0, \varsigma^2)$, a method-by-item interaction term $c_{mi} \sim \mathcal{N}(0, \tau_m^2)$, and model error terms $\varepsilon_{mir} \sim \mathcal{N}(0, \varphi_m^2)$. All random-effect terms are assumed to be independent. For the case when replicate measurements are assumed to be exchangeable for item $i$, $a_{ir}$ can be removed.

The model expressed in (2) describes measurements by $m$ methods, where $m = \{1, 2, 3 \dots\}$. Based on the model expressed in (2), Carstensen et al. (2008) compute the limits of agreement as

$$\alpha_1 - \alpha_2 \pm 2\sqrt{\tau_1^2 + \tau_2^2 + \varphi_1^2 + \varphi_2^2}$$

Carstensen et al. (2008) notes that, for $m = 2$, separate estimates of $\tau_m^2$ can not be obtained. To overcome this, the assumption of equality, i.e. $\tau_1^2 = \tau_2^2$ is required.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \tag{1.13}$$

## 1.3 Regression Methods

Conventional regression models are estimated using the ordinary least squares (OLS) technique, and are referred to as 'Model I regression' (Cornbleet and Cochrane, 1979; Ludbrook, 1997). A key feature of Model I models is that the independent variable is assumed to be measured without error. Cornbleet and Cochrane (1979) argue for the use of methods that based on the assumption that both methods are imprecisely measured, and that yield a fitting that is consistent with both '$X$ on $Y$' and '$Y$ on $X$' formulations. These methods uses alternatives to the OLS approach to determine the slope and intercept. The fundamental flaw of simple linear regression is that it allows for measurement error in one variable only. This causes a downward biased slope estimate.

As often pointed out in several papers (Altman and Bland, 1983; Ludbrook, 1997), this assumption invalidates simple linear regression for use in method comparison studies, as both methods must be assumed to be measured with error.

Errors-in-variables models are regression models that account for measurement errors in the independent variables, as well as the dependent variable.

The use of regression models that assumes the presence of error in both variables $X$ and $Y$ have been proposed for use instead (Cornbleet and Cochrane, 1979; Ludbrook, 1997). These models are collectively known as 'Model II regression'.

Regression estimates depend on formulation of the model. A formulation with one method considered as the $X$ variable will yield different estimates for a formulation where it is the $Y$ variable. With Model I regression, the models fitted in both cases will entirely different and inconsistent. However with Model II regression, they will be consistent and complementary.

## 1.4 Error In Variable Models

Cornbleet and Cochrane (1979) comparing the three methods, citing studies by other authors, concluding that Deming regression is the most useful of these methods. They found the Bartlett method to be flawed in determining slopes.

Model II regression is suitable for method comparison studies, but it is more difficult to execute.

The Bland-Altman plot is uninformative about the comparative influence of proportional bias and fixed bias. Model II approaches, such as Deming regression, can provide independent tests for both types of bias.

However the author point out that *clinical laboratory measurements usually increase in absolute imprecision when larger values are measured.* However one of the assumptions that underline Deming and Mandel regression is constancy of the measurement errors throughout the range of values.

## 1.5 Deming Regression

The most commonly known Model II methodology is known as Deming's Regression, an approach that assumes error in both variables, and is recommended by Cornbleet and Cochrane (1979) as the preferred Model II regression for use in method comparison studies.

Informative analysis for the purposes of method comparison, Deming Regression is a regression technique taking into account uncertainty in both the independent and dependent variables.

The Deming regression method calculates a line of best fit for two sets of data. It differs from simple linear regression in that it is derived in a way that factors in for error in the x-axis, as well as the y-axis.

The sum of the square of the residuals of both variables are simultaneously minimized. This derivation results in the best fit to minimize the sum of the squares of the perpendicular distances from the data points. Normally distributed error of both variables is assumed, as well as a constant level of imprecision throughout the range of measurements.

The sum of squared distances from measured sets of values to the regression line is minimized at an angle specified by the ratio $\lambda$ of the residual variance of both variables. When $\lambda$ is one, the angle is 45 degrees. In ordinary linear regression, the distances are minimized in the vertical directions (Linnet, 1999).

In cases involving only single measurements by each method, $\lambda$ may be unknown and is therefore assumes a value of one. While this will produce biased estimates, they are less biased than ordinary linear regression.

### 1.5.1 Kummel's Estimates

The appropriate estimates were derived by Kummel (1879), but were popularized in the context of medical statistics and clinical chemistry by Deming (1943). For a given $\lambda$, Kummel (1879) derived the following estimate that would later be used for the Deming regression slope parameter.

$$\hat{\beta} = \frac{S_{yy} - \lambda S_{xx} + [(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2]^{1/2}}{2S_{xy}}, \tag{1.14}$$

with $\lambda$ as the variance ratio. The intercept estimate $\alpha$ is simply estimated in the same way as in conventional linear regression, by using the identity $\bar{Y} - \hat{\beta}\bar{X}$. As stated previously $\lambda$ is often unknown, and therefore must be assumed to equal one.

Carroll and Ruppert (1996) states that Deming regression is acceptable only when the precision ratio ($\lambda$, in their paper as $\eta$) is correctly specified, but in practice this is often not the case, with the $\lambda$ being underestimated. Several candidate models, with varying variance ratios may be fitted, and estimates of the slope and intercept are produced. However no model selection information is available to determine the best fitting model.

The measurement error is specified with measurement error variance related as $\lambda = \sigma^2_y/\sigma^2_x$, where $\sigma^2_x$ and $\sigma^2_y$ is the measurement error variance of the $x$ and $y$ variables, respectively. The variance of the ratio, $\lambda$, specifies the angle. When $\lambda$ is one, the angle is 45 degrees. This approach would be appropriate when errors in $y$ and $x$ are both caused by measurements, and the accuracy of measuring devices or procedures are known. In cases involving only single measurements by each method, $\lambda$ may be unknown and is therefore assumes a value of one. While this will bias the estimates, it is less biased than ordinary linear regression. Deming regression assumes that the variance ratio $\lambda$ is known. When $\lambda$ is defined as one, (i.e. equal error variances), the approach is equivalent to orthogonal regression.

### 1.5.2 Confidence Intervals

#### Inferences for Deming Regression

As with conventional regression methodologies, Deming regression calculates an estimate for both the slope and intercept for the fitted line, and standard errors thereof. Therefore there is sufficient information to carry out hypothesis tests on both estimates, that are informative about presence of constant and proportional bias. The test for the intercept estimate acts as a test for the presence of constant bias between both measurement methods.

The intercept and slope are calculated according to Cornbleet and Cochrane (1979). The standard errors and confidence intervals are estimated using the jackknife method (Armitage et al., 2002). Bootstrap techniques can be used to obtain confidence intervals for Deming regression estimates. Authors such as Carpenter and Bithell (2000) and Johnson (2001) provide relevant insights.

A hypothesis test for the intercept estimate can be used to test the intercept, and hence fixed bias, is equal to zero. Rejection of this hypothesis indicates that constant bias is present.

Similarly the test for the slope estimate can be used to test the hypothesis that the slope is equal to 1, and equivalently the presence of proportional bias. Rejection of this hypothesis indicates that a proportional bias exists.

Similarly the test for the slope estimate can be used to formally test proportional bias between the two methods.

Model selection and diagnostic technique are well developed for classical linear regression methods. Typically an implementation of a linear model fit will be accompanied by additional information, such as the coefficient of determination and likelihood and information criterions, and a regression ANOVA table. Such additional information has not, as yet, been implemented for Deming regression.

### 1.5.3 Zhang Data

For convenience, a new data set shall be introduced to demonstrate Deming regression. Measurements of transmitral volumetric flow (MF) by doppler echocardiography, and left ventricular stroke volume (SV) by cross sectional echocardiography in 21 patients with aortic valve disease are tabulated in Zhang et al. (1986). This data set features in the discussion of method comparison studies in Altman (1991, p.398) .

| Patient | MF ($cm^3$) | SV ($cm^3$) | Patient | MF ($cm^3$) | SV ($cm^3$) | Patient | MF ($cm^3$) | SV ($cm^3$) |
|---|---|---|---|---|---|---|---|---|
| 1 | 47 | 43 | 8 | 75 | 72 | 15 | 90 | 82 |
| 2 | 66 | 70 | 9 | 79 | 92 | 16 | 100 | 100 |
| 3 | 68 | 72 | 10 | 81 | 76 | 17 | 104 | 94 |
| 4 | 69 | 81 | 11 | 85 | 85 | 18 | 105 | 98 |
| 5 | 70 | 60 | 12 | 87 | 82 | 19 | 112 | 108 |
| 6 | 70 | 67 | 13 | 87 | 90 | 20 | 120 | 131 |
| 7 | 73 | 72 | 14 | 87 | 96 | 21 | 132 | 131 |

Table 1.5.2: Transmitral volumetric flow(MF) and left ventricular stroke volume (SV) in 21 patients. (Zhang et al 1986)

This ratio can be estimated if multiple measurements were taken with each method, but if only one measurement was taken with each method, it can be assumed to be equal to one.

Deming regression is undermined by several factors. Firstly it is computationally complex, and it requires specific software packages to perform calculations. Secondly,

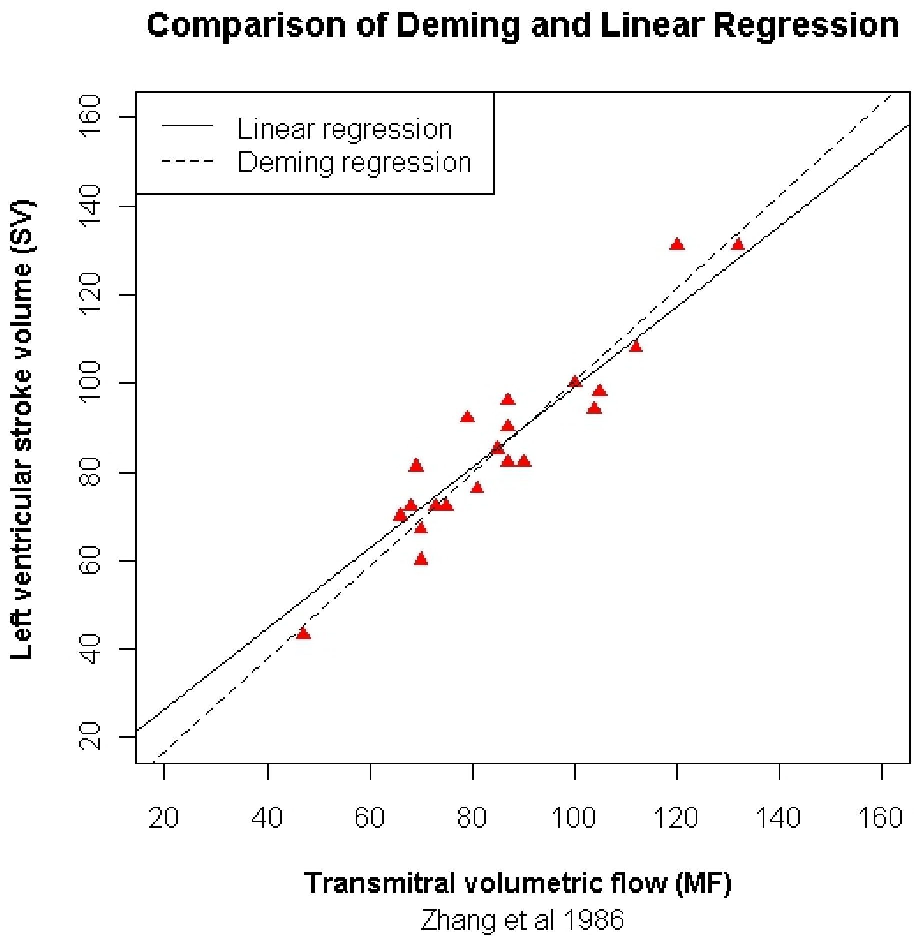


Figure 1.5.1: Deming Regression For Zhang's Data

in common with all regression methods, Deming regression is vulnerable to outliers. Lastly, Deming regression is uninformative about the comparative precision of two methods of measurement. Most importantly Carroll and Ruppert (1996) states that Deming's regression is acceptable only when the precision ratio ($\lambda$, in their paper as $\eta$) is correctly specified, but in practice this is often not the case, with the $\lambda$ being underestimated. This underestimation leads to an overcorrection for attenuation.

Several candidate models, with varying variance ratios may be fitted, and estimates of the slope and intercept are produced. However no model selection information is available to determine the best fitting model.

As noted before, Deming regression is an important and informative methodology in method comparison studies. For single measurement method comparisons, Deming regression offers a useful complement to LME models.

### 1.5.4 Structural Equation Modelling

Structural equation modelling is a statistical technique used for testing and estimating causal relationships using a combination of statistical data and qualitative causal assumptions. Carrasco (2004) describes the structural equation model is a regression approach that allows to estimate a linear regression when independent variables are measured with error. The structural equations approach avoids the biased estimation of the slope and intercept that occurs in ordinary least square regression.

Several authors, such as Lewis et al. (1991), Kelly (1985), Voelkel and Siskowski (2005) and Hopkins (2004) advocate the use of SEM methods for method comparison. In Hopkins (2004), a critique of the Bland-Altman plot he makes the following remark:

> *What's needed for a comparison of two or more measures is a generic approach more powerful even than regression to model the relationship and error structure of each measure with a latent variable representing the true value.*

Hopkins also adds that he himself is collaborating in research utilising SEM and mixed effects modelling. Kelly (1985) advised that *the Structural equations model is used to estimate the linear relationship between new and standards method. The Delta method is used to find the variance of the estimated parameters* (Kelly, 1985).

Conversely Bland and Altman (1999) also states that consider structural equation models to be inappropriate. However Altman et al. (1987) contends that it is unnecessary to perform elaborate statistical analysis, while also criticizing the SEM approach on the basis that it offers insights on inter-method bias only, and not the variability about the line of equality.

> *However, it is quite wrong to argue solely from a lack of bias that two methods can be regarded as comparable... Knowing the data are consistent with a structural equation with a slope of 1 says something about the absence of bias but nothing about the variability about $Y = X$ (the difference between the measurements), which, as has already been stated, is all that really matters.*

Dunn (2002) highlights an important issue regarding using models such as structural equatio modelling; the identifiability problem. This comes as a result of there being too many parameters to be estimated. Therefore assumptions about some parameters, or estimators used, must be made so that others can be estimated. For example, the ratio of the precision of both methods $\lambda = \frac{\sigma_1^2}{\sigma_2^2}$ must often be assumed to be equal to 1 (Linnet, 1998).

Dunn (2002) considers methodologies based on two methods with single measurements on each subject as inadequate for a serious study on the measurement characteristics of the methods, simply because there would not be enough data to allow for a meaningful analysis. There is, however, a contrary argument that in many practical settings it is very difficult to get replicate observations when the measurement method requires an invasive medical procedure.

# Bibliography


Altman, D. and J. Bland (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician) 32*(3), 307–317.

Altman, D. G. (1991). *Practical Statistics for Medical Research*. Chapman and Hall.

Altman, D. G., J. M. Bland, and G. E. Kelly (1987). Letters to the editors.

Barnhart, H., M. Haber, and L. Lin (2007). An overview of assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics 17*, 529–569.

Bartko, J. (1994). Measures of agreement: A single procedure. *Statistics in Medicine 13*, 737–745.

Bland, J. and D. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet i*, 307–310.

Bland, J. and D. Altman (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research 8*(2), 135–160.

Bradley, E. L. and L. G. Blackwood (1989). Comparing paired data: A simultaneous test for means and variances. *The American Statistician 43*(4), 234–235.

Carpenter, J. and J. Bithell (2000). Bootstrap con" dence intervals: when, which, what? a practical guide for medical statisticians.

Carrasco, J. L. (2004). Structural equation model. In *Encyclopedia of Biopharmaceutical Statistics, Second Edition*, pp. 1–7. Taylor & Francis.

Carroll, R. and D. Ruppert (1996). The use and misuse of orthogonal regression in linear errors-in-variables models. *The American Statistician 50*(1), 1–6.

Carstensen, B. (2004). Comparing and predicting between several methods of measurement. *Biostatistics 5*(3), 399–413.

Carstensen, B. (2010). Comparing methods of measurement: Extending the loa by regression. *Statistics in medicine 29*(3), 401–410.

Carstensen, B., J. Simpson, and L. C. Gurrin (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics 4*(1).

Cornbleet, P. J. and D. Cochrane (1979). Regression methods for assessing agreement between two methods of clinical measurement. *Journal of Clinical Chemistry 24*(2), 342–345.

Dunn, G. (2002). *Statistical Evaluation of Measurement Error* (Second ed.). Stanford: American Mathematical Society and Digital Press.

Grubbs, F. (1948). On estimating precision of measuring instruments and product variability. *Journal of the American Statistical Association 43*, 243–264.

Grubbs, F. (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics 15*(1), 53–66.

Hopkins, W. G. (2004). Bias in bland-altman but not regression validity analyses. *Sportscience 8*(4).

Johnson, R. W. (2001). An introduction to the bootstrap. *Teaching Statistics 23*(2), 49–54.

Kelly, G. E. (1985). Use of the structural equations model in assessing the reliability of a new measurement technique. *Applied Statistics*, 258–263.

Kinsella, A. (1986). Estimating method precision. *The Statistician 35*, 421–427.

Kummel, C. (1879). Reduction of observation equations which contain more than one observed quantity. *The Analyst 6*, 97–105.

Lewis, P., P. Jones, J. Polak, and H. Tillitson (1991). The problem of conversion in method comparison studies. *Applied Statistics 40*, 105–112.

Lewis, P., P. Jones, J. Polak, and H. Tillotson (1991). The problem of conversion in method comparison studies. *Applied Statistics*, 105–112.

Linnet, K. (1998). Performance of deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry 44*, 1024–1031.

Linnet, K. (1999). Necessary sample size for method comparison studies based on regression analysis. *Clinical Chemistry 45*(6), 882–894.

Ludbrook, J. (1997). Comparing methods of measurement. *Clinical and Experimental Pharmacology and Physiology 24*, 193–203.

Morgan, W. A. (1939). A test for the signicance of the difference between two variances in a sample from a normal bivariate population. *Biometrika 31*, 13–19.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika 31*, 9–12.

Thompson, W. (1963). Precision of simultaneous measurement procedures. *Journal of American Statistical Association 58*, 474–479.

Voelkel, J. G. and B. E. Siskowski (2005). Center for quality and applied statistics kate gleason college of engineering rochester institute of technology technical report 2005–3.

Zhang, Y., S. Nitter-Hauge, H. Ihlen, K. Rootwelt, and E. Myhre (1986). Measurement of aortic regurgitation by doppler echocardiography. *British Heart Journal 55*, 32–38.