Contents

Chapter 1

Formal tests

The Bland Altman plot is a simple tool for inspection of the data, but in itself it offers little in the way of formal testing. It is upon the practitioner opinion to judge the outcome of the methodology. A formal test, proposed by Altman and Bland (1983) on the Pearson correlation coefficient of casewise differences and means (ρ_{AD}). According to the authors, this test is equivalent to a well established tests for equality of variances, known as the 'Pitman-Morgan Test' (Pitman, 1939; Morgan, 1939).

1.1 Classical model for single measurements

Before continuing, we require a simple model to describe a measurement by method m. Carstensen (2004) presents a model to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland Altman plots. This model assumes that inter-method bias is the only difference between the two methods.

This model is based on measurements y_{mi} by method $m = 1, 2$ on item $i = 1, 2 \ldots$. We use the term *item* to denote an individual, subject or sample, to be measured, being randomly sampled from a population.

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma^2_m) \tag{1.1}$$

Here α_m is the fixed effect associated with method m, μ_i is the true value for subject i (fixed effect) and e_{mi} is a random effect term for errors.

The case-wise differences are expressed as $d_i = y_{1i} - y_{2i}$. Even though the separate variances can not be identified, their sum can be estimated by the empirical variance of the differences, using standard statistical theory. The covariance matrix for case-wise differences can be specified as

$$\Sigma = \left(\begin{array}{cc} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma_b^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{array} \right).$$

Likewise the separate α 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 μ_i. As the case-wise differences are of interest, the parameters of interest are the fixed effects for methods α_m.

1.2 Carstensen 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 rth replicate measurement on the ith item by the mth method ($m = 1, 2$, $i = 1, \ldots, N$, and $r = 1, \ldots, n$), can be written as

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \epsilon_{mir}. \tag{1.2}$$

Again, the fixed effects α_m and μ_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 $\varepsilon \sim \mathcal{N}(0, \varphi^2_m)$. 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 $m = \{1, 2, 3 \ldots\}$. Based on the model expressed in (2), Carstensen et al. (2008) compute the limits of agreement as

$$\alpha_1 - \alpha_2 \pm 2\sqrt{\tau^2_1 + \tau^2_2 + \varphi^2_1 + \varphi^2_2}$$

Carstensen et al. (2008) notes that, for $m = 2$, separate estimates of τ^2_m can not be obtained. To overcome this, the assumption of equality, i.e. $\tau^2_1 = \tau^2_2$ is required.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma^2_m), \quad c_{mi} \sim \mathcal{N}(0, \tau^2_m). \tag{1.3}$$

1.3 Morgan Pitman Testing

An early contribution to formal testing in method comparison was made by both Morgan (1939) and Pitman (1939), in separate contributions. The basis of this approach is that the distribution of the original measurements is bivariate normal.

The test of the hypothesis that the variances σ^2_1 and σ^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 $\rho_{(d,s)} = (\sigma^2_1 - \sigma^2_2)/(\sigma_D \sigma_S)$, which is zero if, and only if, $\sigma^2_1 = \sigma^2_2$. The classical Pitman-Morgan test can be adapted for the correlation value $\rho_{(a,d)}$,and is evaluated as follows;

$$\rho(a, d) = \frac{\sigma^2_1 - \sigma^2_2}{\sqrt{(\sigma^2_1 + \sigma^2_2)(4\sigma^2_S + \sigma^2_1 + \sigma^2_2)}} \tag{1.4}$$

Morgan and Pitman noted that the correlation coefficient depends upon the difference $\sigma^2_1 - \sigma^2_2$, being zero if and only if $\sigma^2_1 = \sigma^2_2$. Therefore a test of the hypothesis $H : \sigma^2_1 = \sigma^2_2$ is equivalent to a test of the hypothesis $H : \rho(D, A) = 0$. This corresponds to the well-known t test for a correlation coefficient with $n - 2$ degrees of freedom. Bartko (1994) describes the Morgan-Pitman test as identical to the test of the slope equal to zero in the regression of y_{i1} on y_{i2}, a result that can be derived using straightforward algebra.

1.4 Bland-Altman correlation test

The approach proposed by Altman and Bland (1983) is a formal test on the Pearson correlation coefficient of case-wise differences and means (ρ_{AD}). According to the authors, this test is equivalent to the 'Pitman Morgan Test'. For the Grubbs data, the correlation coefficient estimate (r_{AD}) is 0.2625, with a 95% confidence interval of (-0.366, 0.726) estimated by Fishers 'r to z' transformation (Cohen, Cohen, West, and Aiken, Cohen et al.). The null hypothesis (ρ_{AD} =0) fail to be rejected. Consequently the null hypothesis of equal variances of each method would also fail to be rejected. There has no 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) comments 'we do not see a place for methods of analysis based on hypothesis testing'. Bland and Altman (1999) also states that consider structural equation models to be inappropriate.

1.5 Paired sample *t* test

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 and is calculated as follows,

$$t^* = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \tag{1.5}$$

where $\bar{d}$ and s_d is the average of the differences of the n observations. Only if the two methods show comparable precision then the paired sample student t-test is appropriate for assessing the magnitude of the bias.

1.6 Regression Methods

Scatterplots are recommended by Altman and Bland (1983) for an initial examination of the data, facilitating an initial judgement and helping to identify potential outliers. They are not useful for a thorough examination of the data. O'Brien et al. (1990) notes that data points will tend to cluster around the line of equality, obscuring interpretation.

The Bland Altman methodology is well noted for its ease of use, and can be easily implemented with most software packages. Also it doesn't require the practitioner to have more than basic statistical training. The plot is quite informative about the variability of the differences over the range of measurements. For example, an inspection of the plot will indicate the 'fan effect'. They also can be used to detect the presence of an outlier.

Ludbrook (1997, 2002) criticizes Bland-Altman plots on the basis that they presents no information on effect of constant bias or proportional bias. These plots are only practicable when both methods measure in the same units. Hence they are totally unsuitable for conversion problems. The limits of agreement are somewhat arbitrarily constructed. They may or may not be suitable for the data in question. It has been found that the limits given are too wide to be acceptable. There is no guidance on how to deal with outliers. Bland and Altman recognize effect they would have on the limits of agreeement, but offer no guidance on how to correct for those effects.

1.6.1 Decomposition of Inter-Method Bias

Regression approaches are useful for a making a detailed examination of the biases across the range of measurements, allowing inter-method bias to be decomposed into fixed bias and proportional bias. Fixed bias describes the case where one method gives values that are consistently different to the other across the whole range.

Constant or proportional bias in method comparison studies using linear regression can be detected by an individual test on the intercept or the slope of the line regressed from the results of the two methods to be compared.

Proportional bias describes the difference in measurements getting progressively greater, or smaller, across the range of measurements. A measurement method may have either an attendant fixed bias or proportional bias, or both (?).

If the basic assumptions underlying linear regression are not met, the regression equation, and consequently the estimations of bias are undermined. Outliers are a source of error in regression estimates.

1.6.2 Inference Procedures

A 95% confidence interval for the intercept estimate can be used to test the intercept, and hence fixed bias, is equal to zero. This hypothesis is accepted if the confidence interval for the estimate contains the value 0 in its range. Should this be, it can be concluded that fixed bias is not present. Conversely, if the hypothesis is rejected, then it is concluded that the intercept is non zero, and that fixed bias is present.

Testing for proportional bias is a very similar procedure. The 95% confidence interval for the slope estimate can be used to test the hypothesis that the slope is equal to 1. This hypothesis is accepted if the confidence interval for the estimate contains the value 1 in its range. If the hypothesis is rejected, then it is concluded that the slope is significant different from 1 and that a proportional bias exists.

1.7 Bradley-Blackwood Method

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 (μ) and variance (σ^2) of both data sets are equal if the slope and intercept estimates are equal to zero(i.e $\sigma^2_1 = \sigma^2_2$ and $\mu_1 = \mu_2$ if and only if $\beta_0 = \beta_1 = 0$)

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 methodology 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 methodology. Bartko's test statistic take the form:

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

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

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

Importantly, this methodology 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.

1.8 Error-In-Variable Models

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. 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.

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 methodologies are collectively known as "Error-In-Variables Models" and "Model II regression". They differ in the method used to estimate the parameters of the regression.

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

The Bland Altman Plot is uninformative about the comparative influence of proportional bias and fixed bias. Model II approaches can provide independent tests for both types of bias.

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.

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.

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.

1.9 Deming Regression

The most commonly known Model II methodology is known as Deming's Regression, (also known an Ordinary Least Product regression). Deming regression 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. Demings method always results in one regression fit, regardless of which variable takes the place of the predictor variables.

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 fixed and proportional bias.

Deming regression method also 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 angles specified by the ratio λ of the residual variance of both variables. The measurement error (lambda or λ) is specified with measurement error variance related as

$$\lambda = \sigma^2_y / \sigma^2_x$$

where σ^2_x and σ^2_y is the measurement error variance of the x and y variables, respectively. The variance of the ratio,λ, specifies the angle. When λ 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, λ 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 ratio $\lambda = \sigma^2_\epsilon/\sigma^2_\eta$ is known. In the case where λ is equal to one, (i.e. equal error variances), the methodology is equivalent to ***orthogonal regression***.

1.9.1 Kummel's Estimates

For a given λ, Kummel (1879) derived the following estimate that would later be used for the Deming regression slope parameter. The intercept estimate α is simply estimated in the same way as in conventional linear regression, by using the identity $\bar{Y} - \hat{\beta}\bar{X}$;

$$\hat{\beta} = \frac{S_{yy} - \lambda S_{xx} + [(S_{yy} - \lambda S_{xx})^2 + 4\lambda S^2_{xy}]^{1/2}}{2S_{xy}} \tag{1.6}$$

, with λ as the variance ratio. As stated previously λ 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 (λ,in their paper as η) is correctly specified, but in practice this is often not the case, with the λ 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.

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 fixed and proportional bias.

1.10 Zhange Example

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	SV	Patient	MF	SV	Patient	MF	SV
	(cm^3)	(cm^3)		(cm^3)	(cm^3)		(cm^3)	(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 1.10.2: Transmitral volumetric flow(MF) and left ventricular stroke volume (SV) in 21 patients. (Zhang et al 1986)

1.11 Model Evaluation for Deming Regression

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.

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

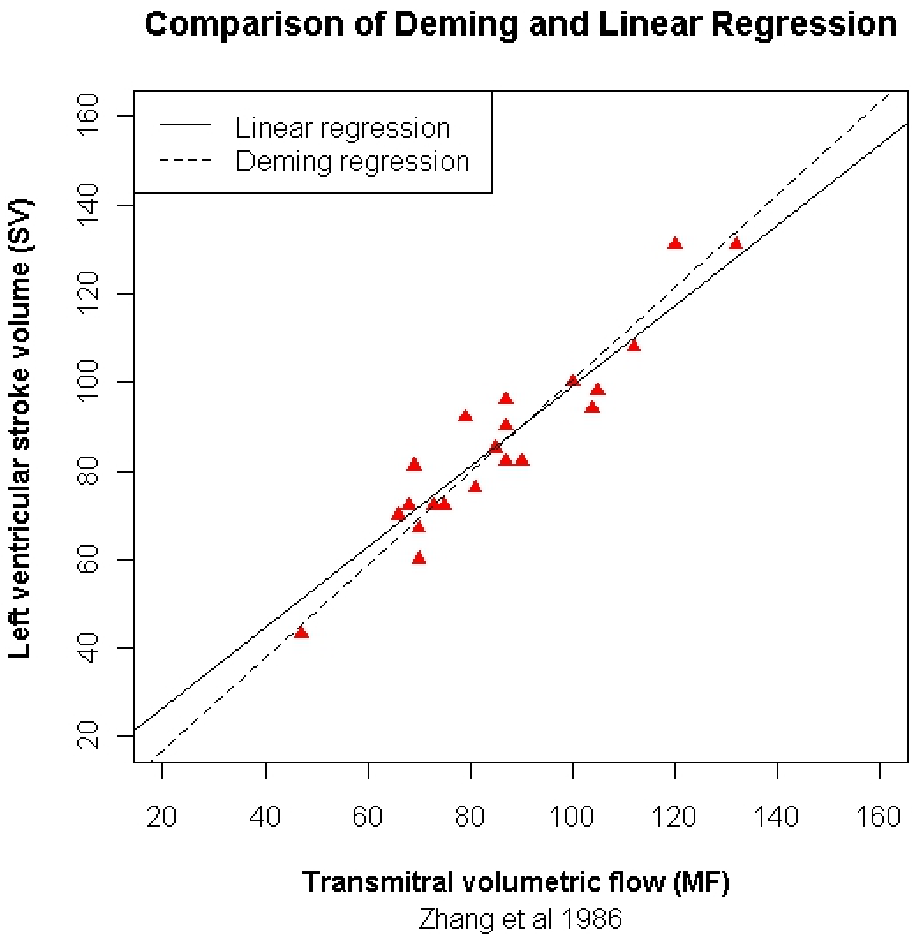

Figure 1.10.1: Deming Regression For Zhang's Data

and information criterions, and a regression ANOVA table. Such additional information has not, as yet, been implemented for Deming regression.

Deming's Regression suffers from some crucial drawbacks. Firstly it is computationally complex, and it requires specific software packages to perform calculations. Secondly, 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 (λ, in their paper as η) is correctly specified, but in practice this is often not the case, with the λ being underestimated. This underestimation leads to an overcorrection for attenuation.

1.12 Other Proposals for Formal Testing

Kinsella (1986) notes the lack of formal testing offered by this Bland-Altman plot. Furthermore, Kinsella (1986) formulates a model for single measurement observations as a linear mixed effects model, i.e. a model that additively combines 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 μ while the fixed effect due to method j is β_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 individual is assumed to give rise to a random error, represented by u_i. This random effects term is assumed to have mean zero and be normally distributed with variance σ^2. There is assumed to be an attendant error for each measurement on each individual, denoted ϵ_{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 σ^2_j. The set of observations (x_i, y_i) by methods X and Y are assumed to follow a bivariate normal distribution with expected values $E(x_i) = \mu_i$ and $E(y_i) = \tau_i$ respectively. The variance covariance of the observations $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \left[\begin{array}{cc} \sigma^2 + \sigma^2_1 & \sigma^2 \\ \sigma^2 & \sigma^2 + \sigma^2_2 \end{array} \right]$$

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 estimating the variances σ^2, σ^2_1 and σ^2_2. Grubbs (1948) offers estimates, commonly known as Grubbs estimators, for the various variance components. These estimates are maximum likelihood estimates, which shall be revisited in due course.

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

Thompson (1963) defines $\Delta_j = \sigma^2/\sigma_j^2, j = 1, 2$, to be a measure of the relative precision of the measurement methods, and demonstrates how to make statistical inferences about Δ_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 Δ_1 are

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

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 Δ_2 are found by substituting C_y for C_x in (1.2). Negative lower limits are replaced by the value 0.

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 $E(a_i) = \mu_a = (\mu_1 + \mu_2)/2$, and the variance matrix $\Sigma_{(a,d)}$ is

$$\Sigma_{(a,d)} = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{bmatrix}. \quad (1.7)$$

Chapter 2

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).

However Bland and Altman (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.*

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.

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., 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).

Cohen, J., P. Cohen, S. West, and L. Aiken. *Applied multiple regression / correlation analysis for the behavioral sciences* (Third ed.). Laurence Erlbaum Associates.

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.

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.

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: a critcal review. *Clinical and Experimental Pharmacology and Physiology 29*, 527–536.

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.

O'Brien, E., J. Petrie, W. Littler, M. de Swiet, P. L. Padfield, D. Altman, M. Bland, A. Coats, and N. Atkins (1990). The British Hypertension Society protocol for the evaluation of blood pressure measuring devices. *Journal of Hypertension 8*, 607–619.

Pinheiro, J. and D. Bates (1994). *Mixed Effects Models in S and S plus* (2nd ed.). Reading, Massachusetts: Springer.

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.