Data reconciliation and gross error diagnosis based on regression

R. Maronna[1]and J. Arcas[2]

[1] Faculty of Exact Sciencies - National University of La Plata,
and C.I.C.P.B.A.
Address: Departamento de Matemática, C.C. 172,
La Plata 1900, Argentina.
e-mail: rmaronna@mail.retina.ar

[2]Faculty of Exact Sciencies - National University of La Plata,
and CONICET
Address: CINDEFI, Calle 50 y 115, La Plata 1900, Argentina
e-mail: arcas@quimica.unlp.edu.ar

Abstract

The estimation of the conversion rates in a biochemical process, subject to the balance equations, raises the issues of detecting gross errors, estimating unobserved rates, and determining the improvement on the estimation of an observed rate contributed by the other measurements (balanceability). In this article we show that the observations constrained by the balance equations may be represented by a linear multiple regression model, with the consequence that the appropriate procedures for each issue are straightforward derivations from standard regression theory. The criteria derived from our approach are shown to be equivalent to the ones proposed by Wang and Stephanopoulos (1983) and by van der Heijden et al.(1994), which are based on special and seemingly different approaches.

A quantity familiar in the detection of regression outliers, called the *leverage* of an observation, is shown to determine both the observation's balanceability, and the probability of detecting a gross error in it.

Several examples with real and simulated data are analysed. The probabilities of detecting the existence of gross errors and of identifying their source are computed. It is shown that these probabilities may be rather low in practical cases, and an approach is proposed to remedy this difficulty.

Key words: gross errors, balanceability, multiple regression.

Running title: Data reconciliation based on regression

INTRODUCTION

The problem of estimating the conversion rates in a biochemical process, subject to the balance equations, has been considered by several authors. Basic issues are whether: (a) an observed rate is a gross error, (b) the estimation of an observed rate can be improved by using the other measurements (balanceability), and (c) an unobserved rate is estimable from the observed ones (calculability).

These issues have been addressed by several authors, who used special and seemingly different approaches for them. For issue (a), Wang and Stephanopoulos (1983) proposed an approach based on deleting one observation at a time: an observation is suspect if its deletion causes a large decrease of the error sum of squares. Van der Heijden et al. (1992 and 1994b) compared the residual vectors as functions of the different sources of error with the help of the redundancy matrix introduced by van der Heijden (1991). For issues (b) and (c), van der Heijden et al. (1994a) used the redundancy matrix and the Singular Value Decomposition for the systematic classification of the measurements according to their balanceability and calculability.

In this article we show that the observations constrained by the balance equations may be represented by a linear multiple regression model. This approach has the advantage that the appropriate procedures for each issue follow in a straightforward manner from the standard theory of linear least squares: no ad-hoc methods are necessary, and the procedures are very simple to apply.

An important consequence of the regression approach is that it naturally brings into consideration the key role of a quantity which is very familiar in outlier detection in regression. This quantity, called the leverage of an observation, is shown to determine both the observation's balanceability, and the probability of detecting a gross error in it. The regression model yields also an explicit expression for this probability.

For the detection of gross errors, we compare the criterion derived from our approach with the ones proposed in the literature, and show that all are equivalent. This criterion is also shown to yield the maximum probability of correct detection. The criteria for balanceability and calculability derived from the regression model are also shown to be equivalent to the published ones. We consider however that they are simpler and more intuitive for a user familiar with regression methods.

We consider some examples analyzed in the literature, and some simulated ones, to show how our procedures work, and also to demonstrate the effect of gross errors on the results. The probabilities of detecting the existence of a gross error, and of identifying its source, are computed through analytic and Monte Carlo methods. It turns out that these probabilities may be rather low in practical cases. Finally an approach is proposed to remedy this difficulty.

THE BASIC MODEL

Let $\mathbf{r}$ be a vector of n unknown rates $r_1, ..., r_n$, subject to $p < n$ independent constraints given by the $p \times n$ composition matrix $\mathbf{E}$:

$$\mathbf{Er} = \mathbf{0}. \tag{1}$$

Call $\mathbf{r}_{\mathrm{m}}$ and $\mathbf{r}_{\mathrm{c}}$ the vectors of (true) measured and non-measured rates, of dimensions m and c respectively, and $\mathbf{E}_{\mathrm{m}}$ and $\mathbf{E}_{\mathrm{c}}$ the respective constraint matrices, so that

$$\mathbf{r} = \left[\begin{array}{c} \mathbf{r}_{\mathrm{m}} \\ \mathbf{r}_{\mathrm{c}} \end{array} \right], \ \mathbf{E} = \left[\mathbf{E}_{\mathrm{m}} | \mathbf{E}_{\mathrm{c}} \right]. \tag{2}$$

and (1) may be written as

$$\mathbf{E}_{\mathrm{m}}\mathbf{r}_{\mathrm{m}} + \mathbf{E}_{\mathrm{c}}\mathbf{r}_{\mathrm{c}} = \mathbf{0}. \tag{3}$$

Call $\mathbf{r}_{\mathrm{ob}} = (r_{\mathrm{ob},1}, ..., r_{\mathrm{ob},m})$ the vector of observed rates: $\mathbf{r}_{\mathrm{ob}} = \mathbf{r}_{\mathrm{m}} + \mathbf{e}$ where $\mathbf{e}$ is the vector of random errors which are assumed to have zero means and known variances. Interest lies in estimating $\mathbf{r}$ —or at least those elements of it that can be estimated— subject to the balance restrictions (1), and to detect gross errors.

Finding a set of values that fits the observed values under a set of linear constraints is the subject of linear regression. The standard linear multiple regression model has the form

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \tag{4}$$

where $\mathbf{y}$ is the vector of observed responses, $\mathbf{X}$ is a fixed known matrix of predictors, β is the unknown vector of regression parameters, and $\mathbf{u}$ is a vector of independent random errors with zero means and equal variances. There are numerous texts on linear regression; e.g., (Draper and Smith, 2001; Montgomery et al., 2001; Weisberg, 1985). It will be now shown that $\mathbf{r}_{\mathrm{ob}}$ may be represented in the form (4). This form will make it possible to apply known results of linear model theory to derive criteria for estimability, balanceability, and gross error detection.

To this end, note that set of n-dimensional vectors $\mathbf{b}$ such that $\mathbf{Eb} = \mathbf{0}$ is a *linear subspace*, i.e., it is closed under addition and multiplication by a scalar. This subspace —called the *null subspace* of $\mathbf{E}$— has dimension $q = n - p$. Hence there exist q linearly independent n-dimensional vectors $\mathbf{b}_1, ..., \mathbf{b}_q$ —a *basis* of the subspace— such that any vector $\mathbf{r}$ satisfying (1) can be expressed as a linear combination of them, and therefore

$$\mathbf{r} = \sum_{j=1}^{q} \beta_j \mathbf{b}_j. \tag{5}$$

for a set of q (unknown) numbers $\beta_1, .., \beta_q$. Call $\mathbf{B}$ the $n \times q$−matrix whose columns are the $\mathbf{b}_j$, and β the q-dimensional vector with components β_j. Then

(5) can be written as $\mathbf{r} = \mathbf{B}\beta$. The matrix $\mathbf{B}$ (which is not unique) is easy to compute. In the matrix languages Matlab and Gauss, $\mathbf{B}$ is obtained through the command "null($\mathbf{E}$)". Otherwise $\mathbf{B}$ can be computed using the QR orthogonalization procedure.

If $\mathbf{B}_{\mathrm{m}}$ and $\mathbf{B}_{\mathrm{c}}$ are the $m \times q$- and $c \times q-$matrices consisting of the first m and the last c rows of $\mathbf{B}$, that is

$$\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_q] = \left[\begin{array}{c} \mathbf{B}_{\mathrm{m}} \\ \mathbf{B}_{\mathrm{c}} \end{array} \right], \tag{6}$$

then

$$\mathbf{E}_{\mathrm{m}}\mathbf{B}_{\mathrm{m}} + \mathbf{E}_{\mathrm{c}}\mathbf{B}_{\mathrm{c}} = \mathbf{0} \tag{7}$$

and

$$\mathbf{r}_{\mathrm{m}} = \mathbf{B}_{\mathrm{m}}\beta, \;\; \mathbf{r}_{\mathrm{c}} = \mathbf{B}_{c}\beta. \tag{8}$$

The observations $\mathbf{r}_{\mathrm{ob}}$ differ from $\mathbf{r}_{\mathrm{m}}$ by random errors, and hence we can write

$$\mathbf{r}_{\mathrm{ob}} = \mathbf{B}_{\mathrm{m}}\beta + \mathbf{e}, \tag{9}$$

where $\mathbf{e}$ is a random vector whose elements are the errors $e_1, .., e_m$, which has mean $\mathbf{0}$ and a covariance matrix $\mathbf{V}$ assumed known. Henceforth var ($\mathbf{x}$) and var (x) will denote the covariance matrix of the random vector $\mathbf{x}$ and the variance of the random variable x, respectively. It will be assumed that the errors are independent and hence that $\mathbf{V}$ is a diagonal matrix:

$$\mathbf{V} = \mathrm{var}(\mathbf{e}) = \mathrm{var}(\mathbf{r}_{\mathrm{ob}}) = \mathrm{diag}(\sigma_1^2, ..., \sigma_m^2) = \mathbf{\Sigma}^2, \tag{10}$$

with $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, ..., \sigma_m)$, where $\sigma_i = \mathrm{sd}(e_i)$ is the standard deviation of e_i.

It follows from (4) that $\mathbf{r}_{\mathrm{ob}}$ is the response vector of a linear regression model with predictor matrix $\mathbf{B}_{\mathrm{m}}$ and errors with possibly unequal variances. To transform it to a standard regression model, call $\mathbf{y}$ and $\mathbf{u}$ the vectors of observations and of errors normalized to unit variances, with elements

$$y_i = \frac{r_{\mathrm{ob},i}}{\sigma_i}, \;\; u_i = \frac{e_i}{\sigma_i} \;\; (i = 1, ..., m), \tag{11}$$

that is, $\mathbf{y} = \mathbf{\Sigma}^{-1}\mathbf{r}_{\mathrm{ob}}$ and $\mathbf{u} = \mathbf{\Sigma}^{-1}\mathbf{e}$; and call $\mathbf{X}$ the matrix obtained by dividing each row of $\mathbf{B}_{\mathrm{m}}$ by the respective σ_i, that is

$$\mathbf{X} = \mathbf{\Sigma}^{-1}\mathbf{B}_{\mathrm{m}}. \tag{12}$$

Then (9) is equivalent to (4). Since the u_is are independent with unit variances, we have $\mathrm{var}(\mathbf{u}) = \mathrm{var}(\mathbf{y}) = \mathbf{I}_m$ where $\mathbf{I}_m$ is the m-dimensional identity matrix. The ordinary least squares estimate $\widehat{\beta}$ of β is the solution of

$$\left\| \mathbf{y} - \mathbf{X}\widehat{\beta} \right\| = \min, \tag{13}$$

where $\|\mathbf{a}\|$ denotes the Euclidean norm of $\mathbf{a}$. A solution is given by $\widehat{\beta} = \mathbf{X}^{+}\mathbf{y}$, where $\mathbf{X}^{+}$ denotes the pseudo-inverse of $\mathbf{X}$. This estimate has the smallest variances among unbiased estimates which are linear in $\mathbf{y}$, according to the Gauss-Markov Theorem (Stapleton, 1995). If $\mathbf{u}$ is normal, then $\widehat{\beta}$ is also the Maximum Likelihood Estimate. The vectors of fitted values and of observation residuals are respectively

$$\widehat{\mathbf{r}}_{\mathrm{m}} = \mathbf{B}_{\mathrm{m}}\widehat{\beta} = \mathbf{\Sigma X}\widehat{\beta} \text{ and } \widehat{\mathbf{e}} = \mathbf{r}_{\mathrm{ob}} - \widehat{\mathbf{r}}_{\mathrm{m}}. \tag{14}$$

ESTIMATION AND CALCULABILITY

Let $d = \mathrm{rank}(\mathbf{X}) = \mathrm{rank}(\mathbf{B}_{\mathrm{m}})$; then $d \leq \min(m, q)$. If $d = q$ ("full rank"), then $\widehat{\beta}$ is unique and $\mathbf{X}^{+} = \left(\mathbf{X}^{t}\mathbf{X}\right)^{-1}\mathbf{X}^{t}$, where in general $\mathbf{X}^{t}$ denotes the transpose of $\mathbf{X}$. In this case the unmeasured rates are estimated through (8):

$$\widehat{\mathbf{r}}_{c} = \mathbf{B}_{\mathrm{c}}\widehat{\beta}. \tag{15}$$

If $d < q$, (13) has infinite solutions, but the fitted values are the same for all $\widehat{\beta}$.

In the terminology of linear regression, a parameter is *estimable* if it has a linear unbiased estimate. If $d < q$, it can be shown that the estimable elements of $\mathbf{r}_{\mathrm{c}}$ correspond to the null columns of

$$\mathbf{C} = \left[\mathbf{I}_{q} - \mathbf{X}^{t}\left(\mathbf{X}^{t}\right)^{+}\right]\mathbf{B}_{\mathrm{c}}^{t}.$$

In general, a linear combination γ of the elements of $\mathbf{r}_{\mathrm{c}}$, $\gamma = \mathbf{a}^{t}\mathbf{r}_{\mathrm{c}}$ —where $\mathbf{a}$ is a given c-dimensional vector— is estimable if and only if

$$\mathbf{Ca} = \mathbf{0}, \tag{16}$$

and its estimate is $\widehat{\gamma} = \mathbf{a}^{t}\widehat{\mathbf{r}}_{\mathrm{c}}$ (to simplify the exposition, proofs of all mathematical statements are deferred to Appendix A).

The estimable elements of $\mathbf{r}_{\mathrm{c}}$ are the *calculable* ones in the sense of van der Heijden et al. (1994a). We thus have a simple criterion for calculability.

BALANCEABILITY

The relationship between the fitted and observed values in model (4) is given by $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta} = \mathbf{Hy}$,where

$$\mathbf{H} = \mathbf{XX}^{+} \tag{17}$$

is the so called "hat matrix" (because it relates the observations $\mathbf{y}$ to the fitted "$\mathbf{y}$ hat"). Then

$$\widehat{y}_{i} = \sum_{j=1}^{m} H_{ij}y_{j}, \tag{18}$$

and hence

$$\mathrm{var}(\widehat{\mathbf{y}}) = \mathbf{H}. \tag{19}$$

The diagonal element H_{ii} of $\mathbf{H}$ is called the *leverage* of the i-th observation in regression theory, and will be denoted for simplicity as h_{i}. The name stems from the fact that an observation with a large h_i has a high weight in the determination of the least squares fit. The h_i's fulfill

$$0 \leq h_i \leq 1 \tag{20}$$

and

$$\sum_{i=1}^{n} h_i = \operatorname{rank}(\mathbf{H}). \tag{21}$$

If $h_i = 1$, then the i-th residual is null.

Call $r_{\mathrm{m},i}$ $(i = 1, ..., m)$ the elements of $\mathbf{r}_{\mathrm{m}}$. It follows from (19) that

$$\operatorname{var}(\widehat{y}_i) = \frac{\operatorname{var}(\widehat{r}_{\mathrm{m},i})}{\operatorname{var}(r_{\mathrm{ob},i})} = h_i. \tag{22}$$

According to the terminology of van der Heijden et al. (994a), $r_{\mathrm{m},i}$ is *balanceable* if it can be estimated using observations other than $r_{\mathrm{ob},i}$; in this case, the variance of the estimate is smaller than that of $r_{\mathrm{ob},i}$.

It follows from (22) and (11) that if $h_i = 1$, then $\operatorname{var}(\widehat{r}_{\mathrm{m},i}) = \operatorname{var}(r_{\mathrm{ob},i})$, and hence measurement i is not balanceable. Otherwise

$$\operatorname{var}(\widehat{r}_{\mathrm{m},i}) = h_i \operatorname{var}(r_{\mathrm{ob},i}) < \operatorname{var}(r_{\mathrm{ob},i}),$$

and hence $r_{\mathrm{m},i}$ is balanceable. Moreover, h_i measures the reduction in variance gained, and hence h_i can be taken as a measure of "unbalanceability".

Since $\mathbf{X}$ depends on $\mathbf{\Sigma}$, so do the h_is. This can be seen intuitively by noting that if for a particular i we let $\sigma_i \to 0$, then the corresponding $h_i \to 1$, which is natural since if an observation has no error, the other observations will not improve on its estimation. However, it can be shown that *exact* unbalanceability (i.e., h_i being *exactly* one) does not depend on $\mathbf{\Sigma}$.

DETECTION OF GROSS ERRORS

The residuals for model (4) are

$$\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\beta} = \mathbf{\Sigma}^{-1}\widehat{\mathbf{e}}, \tag{23}$$

with $\widehat{\mathbf{e}}$ defined in (14). Call

$$S_{\mathrm{res}} = \widehat{\mathbf{u}}^t \widehat{\mathbf{u}} = \sum_{i=1}^{m} \widehat{u}_i^2 \tag{24}$$

the residual sum of squares. If $m = q$, then $S_{\mathrm{res}} = 0$ and no further analysis can be performed. It will henceforth be assumed that $m > q$.

Assume the errors u_i to be normal. It is a standard result of multiple regression theory that S_{res} has a chi-squared distribution with $m - d$ degrees of freedom, where $d = \operatorname{rank}(\mathbf{B}_{\mathrm{m}}) = \operatorname{rank}(\mathbf{X})$. This fact may be used to carry an overall test of fit: if S_{res} is larger than —say— the 0.95 quantile of the chi-squared distribution, then the data do not support model (4), with p-value less than 0.05.

This test does not signal the specific cause of misfit. We shall now estimate the location of a gross error. The residual vector $\widehat{\mathbf{u}}$ has covariance matrix $\mathbf{I}-\mathbf{H}$, and hence $\mathrm{var}(\widehat{u}_j)=1-h_j$. Call $\widehat{e}_{\mathrm{st},i}$ the residuals standardized to unit variance :

$$\widehat{e}_{\mathrm{st},i}=\frac{\widehat{u}_i}{\sqrt{1-h_i}}=\frac{\widehat{e}_i}{\mathrm{sd}\left(\widehat{e}_i\right)}. \tag{25}$$

The $\widehat{e}_{\mathrm{st},i}$ are standard normal (but not independent!). Let i^* be the value of i that maximizes $\left|\widehat{e}_{\mathrm{st},i}\right|$, and let $\widehat{\Delta}=\widehat{u}_{i^*}$. Then i^* estimates the location of the suspect value, which can be corrected replacing y_{i^*} by $y_{i^*}-\widehat{\Delta}=\widehat{y}_{i^*}$ and hence replacing r_{ob,i^*} by $\widehat{y}_{i^*}\sigma_{i^*}$.

An alternative to the chi-squared test is to use the statistic

$$T=\max_i\left|\widehat{e}_{\mathrm{st},i}\right|. \tag{26}$$

Call G the distribution of T under the hypothesis of no gross errors. Then a new test can be defined, by declaring an observed value of T significant at level α if it is larger than the $(1-\alpha)$-percent point of G. The distribution G depends on $\mathbf{X}$, and does not have an explicit form, but accurate approximations can be found; see (Sidák, 1967).

This procedure has a theoretical justification. Call $\mathbf{x}_i$ the i-th row of $\mathbf{X}$. Consider the model of a single gross error, represented by

$$y_i=\begin{cases}\mathbf{x}_i^t\beta+u_i & \text{for} \quad i\neq i_0\\ \mathbf{x}_i^t\beta+\Delta+u_i & \text{for} \quad i=i_0\end{cases} \tag{27}$$

where i_0 and Δ are unknown. Then it is shown by Belsley et al. (1980) that i^* and $\widehat{\Delta}$ are the Maximum Likelihood Estimates of i_0 and Δ, respectively; and that T yields the Likelihood Ratio Test of the null hypothesis $\{\Delta=0\}$ against the alternative $\{\Delta\neq 0\}$. The estimator i^* maximizes the probability of correctly choosing i_0.

A natural approach to outlier detection would be to omit each observation in turn and recompute the estimates. For $i=1,...,m$ call $\widehat{\beta}_{(i)}$ the least squares estimate computed without using observation i, and v_i the respective ("leave-one-out") residual: $v_i=y_i-\mathbf{x}^t\widehat{\beta}_{(i)}$. It can be shown (Belsley et al., 1980; Chatterjee and Hadi, 1988) that $v_i=\widehat{u}_i/\left(1-h_i\right)$, and hence $v_i/\mathrm{sd}(v_i)=\widehat{e}_{\mathrm{st},i}$, so that nothing new can be obtained through this approach.

Detection probabilities

Assume model (27). Then the probability that the chi-squared test yields a p-value less than α is

$$\pi=1-F_{k,\lambda}\left[\chi_k^2(1-\alpha)\right], \tag{28}$$

where $\lambda=\Delta\sqrt{1-h_{i_0}}$, $\chi_k^2(\delta)$ is the δ-quantile of the chi-squared distribution with k degrees of freedom, and $F_{k,\lambda}$ is the distribution function of the non-central chi-squared distribution with k degrees of freedom and non-centrality parameter λ^2. Table I gives π for $\alpha=0.05$ and a range of values of k, h_i and Δ. It follows that if $h_i=0.9$, then the probability is very low even for $\Delta=4$.

In general, an observation with small σ_i will have a large h_i. In fact, the i-th row of the matrix $\mathbf{X}$ is obtained by dividing the i-th row of $\mathbf{B}_{\mathrm{m}}$ by σ_i, and this will yield large values if σ_i is small. Intuitively, an observation with a low error variance receives a high weight in weighted least squares, which makes it more dangerous if it is a gross error.

In a typical situation with $m = 5$ and $d = 2$, the sum of the h_is is 0.6 according to (21), and so we cannot hope all of them to be small.

RELATIONSHIP WITH PREVIOUS WORK

Van der Heijden (1991) and Van der Heijden et al. (1994a) define the *redundancy matrix* as $\mathbf{R} = \mathbf{E}_{\mathrm{m}} - \mathbf{E}_{\mathrm{c}}\mathbf{E}_{\mathrm{c}}^{+}\mathbf{E}_{\mathrm{m}}$. It follows from (3) that $\mathbf{r}_{\mathrm{c}} = -\mathbf{E}_{\mathrm{c}}^{+}\mathbf{E}_{\mathrm{m}}\mathbf{r}_{\mathrm{m}}$, and hence that $\mathbf{R}\mathbf{r}_{\mathrm{m}} = \mathbf{0}$. They define the weighted least squares estimate $\widetilde{\mathbf{r}}_{\mathrm{m}}$ of $\mathbf{r}_{\mathrm{m}}$ as

$$(\mathbf{r}_{\mathrm{ob}} - \widetilde{\mathbf{r}}_{\mathrm{m}})^{t} \operatorname{var}(\mathbf{r}_{\mathrm{ob}})^{-1} (\mathbf{r}_{\mathrm{ob}} - \widetilde{\mathbf{r}}_{\mathrm{m}}) = \min \quad \text{with} \quad \mathbf{R}\widetilde{\mathbf{r}}_{\mathrm{m}} = \mathbf{0}. \tag{29}$$

Let

$$\varepsilon = \mathbf{R}\mathbf{r}_{\mathrm{ob}}, \quad \mathbf{P} = \mathrm{var}(\varepsilon), \quad h_{\varepsilon} = \varepsilon^{t}\mathbf{P}^{+}\varepsilon. \tag{30}$$

The vector ε is called "residual vector" in the cited articles; note that it does not coincide with our *observation* residuals $\widehat{\mathbf{e}}$ or $\widehat{\mathbf{u}}$.

It is shown by van der Heijden et al. (1994a) that h_{ε} has a chi-squared distribution with $k = \mathrm{rank}(\mathbf{R})$ degrees of freedom. This statistic is then used for testing the significance of departures from the model. The location of gross errors is estimated with the technique of "compare vectors" (van der Heijden et al., 1994b). Call $\mathbf{c}_i$ the i-th column of $\mathbf{R}$. The technique is based on the idea that if the i-th observation contains a gross error, then $\mathbf{c}_i$ should have approximately the same direction as ε. Actually, both ε and the $\mathbf{c}_i$ can be represented in a space of dimension k (the "reduced vectors"), which allows a graphical analysis when $k = 2$. The proximity between the directions of ε and $\mathbf{c}_i$ is measured by

$$\delta_i = \frac{\left(\varepsilon^{'}\mathbf{P}^{+}\mathbf{c}_i\right)^2}{\mathbf{c}_i^{'}\mathbf{P}^{+}\mathbf{c}_i}, \tag{31}$$

and hence suspect observations correspond to large δ_is.

Alternatively, Wang and Stephanopoulos (1983) define a sum of squared residuals which coincides with h_{ε}. They estimate the location of the gross error by looking for the observation that causes the largest decrease in h_{ε}. That is, for $i = 1, .., m$ call $S_{(i)}$ the value of h_{ε} computed without using the i-th observation. Then they look for the i such that $S_{(i)}$ is minimum.

It is shown in Appendix A that

A) $\widetilde{\mathbf{r}}_{\mathrm{m}}$ coincides with $\widehat{\mathbf{r}}_{\mathrm{m}}$ in (14)

B) $\varepsilon = \mathbf{R}\widehat{\mathbf{e}}$ coincides with $\widehat{\mathbf{e}}$ defined in (14)

C) the sum h_{ε} coincides with our S_{res} in (24) and $k = m - d$

D) $\delta_i = \widehat{e}^2_{\mathrm{st},i}$

E) $S_{(i)} = S_{\mathrm{res}} - \widehat{e}^2_{\mathrm{st},i}$, and hence minimizing $S_{(i)}$ is equivalent to maximizing $|\widehat{e}_{\mathrm{st},i}|$.

It follows from (D)-(E) that our Maximum Likelihood estimator i^*, the deletion approach of Wang and Stephanopoulos (1983) and the "compare vector" approach of Van der Heijden et al. (1994b) give the same results.

EXAMPLES

Consider first the examples in pages 7-9 of (van der Heijden et al., 1994a), where $\mathbf{r} = (r_X, r_S, r_P, r_N, r_C, r_O, r_W)^t$ and

$$\mathbf{E} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1.83 & 2 & 3 & 3 & 0 & 0 & 2 \\ 0.56 & 1 & 0.5 & 0 & 2 & 2 & 1 \\ 0.17 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}. \tag{32}$$

Since *exact* balanceability does not depend on $\mathbf{\Sigma}$, we may take $\mathbf{\Sigma} = \mathbf{I}$.

In Example 1, $\mathbf{r}_{\mathrm{m}} = (r_N, r_X, r_O)^t$ and the h_i are

$$0.9719 \quad 0.02809 \quad 1.0000$$

so that r_O is not balanceable; and

$$\mathbf{C} = \begin{bmatrix} -0.697 & 0.465 & 0.232 & 0 \\ 0.204 & -0.136 & -0.068 & 0 \\ -0.340 & 0.227 & 0.113 & 0 \end{bmatrix}$$

so that r_W is calculable and r_S, r_P and r_C are not calculable. Note however that each row of $\mathbf{C}$ adds to zero, so that $\mathbf{Ca} = \mathbf{0}$ for $\mathbf{a}$ =$(1,1,1,0)$, and hence $\mathbf{a}^t\mathbf{r}_{\mathrm{c}}$= $r_S + r_P$+r_C is calculable.

In Example 2, $\mathbf{r}_{\mathrm{m}} = (r_N)$. Here $\mathbf{H} = [1]$, and hence r_N is not balanceable. The first column of $\mathbf{C}$ is null, so that r_X is calculable.

In Example 3, $\mathbf{r}_{\mathrm{m}} = (r_X, r_N, r_C, r_O)^t$, the h_i are

$$0.9719 \quad 0.02809 \quad 1.0000 \quad 1.0000$$

and hence neither r_C nor r_O are balanceable; and $\mathbf{C} = \mathbf{0}$ so that all unmeasured rates are calculable.

We now deal with the example in pages 16-17 of van der Heijden et al. (1994b), with the same matrix (32) and $\mathbf{r}_{\mathrm{m}} = (r_X, r_S, r_P, r_C, r_O)^t$ (details of the computation are given in Appendix B). Their fitted values coincide with ours, as it to be expected. The tests based on the chi-squared statistic S_{res} and on the likelihood ratio statistic T in (26) have both p-values of 0.09. The leverage values h_i and normalized residuals $\widehat{e}_{\mathrm{st},i}$ are given in table II. The absolute error standard deviations σ_i are also included for reference. The first three measurements are suspect. The first is almost unbalanceable. Note that it has both the largest h_i and the smallest σ_i.

To examine the performance of the proposed methods, we generate an artificial data set in which the gross errors are known beforehand. To this end, we first take the values fitted in the former example as the "true values" $r_{\mathrm{m},i}$ (i.e., they fulfill the model (1) exactly), and define new "observations" by $r_{\mathrm{ob},i} = r_{\mathrm{m},i} + \sigma_i z_i$, where σ_i is the standard deviation in Table II, and the z_i are independent standard normal. Table III shows the new "data" and the results. The p-values of both tests are 0.64. Now we generate a gross error by adding to observation i_0 the quantity $\Delta\sigma_{i_0}$. Table IV shows the results, for $i_0 = 1$ with $\Delta = 4$ and $i_0 = 4$ with $\Delta = 3$ and 4. When $i_0 = 1$, the gross error is not detected despite its large size. When $i_0 = 3$ and $\Delta = 3$, the result is only significant at the 0.10 level, and i_0 is incorrectly estimated; when $\Delta = 4$ the result is very significant, but O appears also as a probable error source.

Table V gives for each of the five measurements the probability $\pi\,(\mathrm{test})$ of detecting a gross error with $\Delta = 3$, and the probability $\pi\,(i^*)$ of correctly identifying i_0 in the cases that the test was significant at the 0.05 level. The first was computed using (28) , and the second was obtained by a Monte Carlo simulation, repeating 10000 times the generation of the artificial data. The simulation was also used to compute the detection probabilities of the test based on T defined in (26). Since the results were not better than those of the chi-squared test, we found no reason to recommend T.

DISCUSSION

The approach based on regression has yielded a straightforward derivation of the criteria for balanceability and calculability, and for the detection of gross errors. The proposed methods are equivalent to the former ones. Besides, this approach highlights the key role of the *leverage* h_i, in measuring both the degree of balanceability of an observation, and the probability of detecting a gross error in it.

The regression aproach makes it straightforward to calculate the probabilities of detecting the presence of a single gross error, and of locating its source. These probabilities are seen to be rather low in practical cases. For observations with high leverage, they may be extremely low. The main reason for this difficulty is that usually the number of observations is small compared to that of unknown parameters. The situation is better if there are replications, for in this case the leverage of each observation is divided by the number of replications.

However, there is a better way to reconcile the data and detect gross errors, and it is to use *all* the available information. In a biochemical experiment one seldom has an isolated set of observations. There is an external variable t, such as time, flow rate or dilution rate, and for each value of t there is a vector $\mathbf{r}_{\mathrm{ob}}(t)$. It is natural to assume that the true values $\mathbf{r}_{\mathrm{ob}}(t)$ depend smoothly on t, and hence to reconcile the data for a given t, one may use the information contained in $\mathbf{r}_{\mathrm{ob}}(t')$ for t' close to t. This approach allows us to fit the data with a much lower number of parameters, and should hence be better than dealing with each $\mathbf{r}_{\mathrm{ob}}(t)$ independently of one another. The implementation of this idea is in progress.

APPENDIX A: PROOFS OF RESULTS

For reasons of space, only the essential steps of the proofs are presented.

Proof of (16)

Let $\gamma = \mathbf{a}^t\mathbf{r}_{\mathrm{c}} = \mathbf{a}^t\mathbf{B}_{\mathrm{c}}\beta$ for some c-dimensional vector $\mathbf{a}$. If $\widehat{\beta}$ is not unique, a linear combination of β is estimable (Stapleton, 1995) if and only if it depends on β only through $\mathbf{X}\beta$ (recall that $\mathbf{X}\widehat{\beta}$ is always unique even if $\widehat{\beta}$ is not). Let $\mathbf{d} = \mathbf{B}_{\mathrm{c}}^t\mathbf{a}$. Then the estimability of γ is equivalent to $\mathbf{d} = \mathbf{X}^t\mathbf{c}$ for some m-dimensional vector $\mathbf{c}$. A solution to $\mathbf{d} = \mathbf{X}^t\mathbf{c}$ is $\mathbf{c} = (\mathbf{X}^t)^+\mathbf{d}$, and hence $\mathbf{B}_{\mathrm{c}}^t\mathbf{a} = \mathbf{X}^t\left(\mathbf{X}^t\right)^+\mathbf{B}_{\mathrm{c}}^t\mathbf{a}$, which is equivalent to (16).

The matrix H

It follows from (17) that $\mathbf{H}$ verifies $\mathbf{H} = \mathbf{H}^t = \mathbf{H}^2$, and hence

$$h_i = \sum_{j=1}^{m} H_{ij}^2 = h_i^2 + \sum_{j\neq i} H_{ij}^2,$$

which implies that $h_i \geq 0$ and $h_i(1-h_i) \geq 0$, which is equivalent to (20); and that if h_i is either 0 or 1, then $H_{ij} = 0$ for $j \neq i$.

It will now be shown that exact balanceability does not depend on $\mathbf{\Sigma}$. The *image* of $\mathbf{X}$ is the subspace $\mathrm{Im}(\mathbf{X})$ of vectors equal to $\mathbf{Xa}$ for some $\mathbf{a}$. Assume that $\mathbf{X}$ is such that $h_1 = 1$. This is equivalent to $H_{1j} = 0$ for $j \neq 1$, and hence to $\mathbf{Hv} = \mathbf{v}$ for $\mathbf{v} = (1,0,0,..,0)$. This is in turn equivalent to $\mathbf{v} \in \mathrm{Im}(\mathbf{X}) = \mathrm{Im}(\mathbf{B}_{\mathrm{m}})$; and this property does not depend on $\mathbf{\Sigma}$.

Optimality of i^*

Note that each i_0 in (27) yields a different distribution for the observations $\mathbf{r}_{\mathrm{ob}}$. Assume that i_0 is chosen at random among $\{1,..,m\}$ with equal probabilities. Estimating i_0 is then a typical problem in Discrimination (Seber, 1984). It is known that the estimator that maximizes the probability of choosing the true value of i_0 is given by choosing among the m distributions the one that attributes the highest likelihood to the observed data, and this property is fulfilled by i^* since it is the Maximum Likelihood estimator.

Proof of (28)

If the random vector $\mathbf{z}$ is k-variate normal with $\mathrm{var}(\mathbf{z}) = \mathbf{I}_k$ and expectation μ, then $\|\mathbf{z}\|^2$ has a non-central chi-squared distribution with k degrees of freedom and non-centrality parameter $\lambda^2 = \|\mu\|^2$. Under model (27), $\mathbf{y}$ has identity covariance matrix and expectation $\mathbf{X}\beta+\Delta\mathbf{v}$, where $\mathbf{v} = (v_1,..,v_m)$ with $v_{i_0} = 1$ and $v_i = 0$ for $i \neq i_0$. Hence $\widehat{\mathbf{u}}= (\mathbf{I}_m-\mathbf{H})\mathbf{y}$ has $\mathrm{var}(\widehat{\mathbf{u}}) = \mathbf{I}_m-\mathbf{H}$ and expectation $\Delta\mu$ with $\mu = (\mathbf{I}_m-\mathbf{H})\,\mathbf{v}$. Note that $\|\mu\|^2 = 1-h_{i_0}$. Since $\mathbf{I}_m-\mathbf{H}$ is an orthogonal projection matrix of rank $k = m-d$, it has k unit eigenvalues and $m-k$ null ones. Call $\mathbf{a}_i$ $(i = 1,...,k)$ the eigenvectors corresponding to the former, and let $z_i = \widehat{\mathbf{u}}^t\mathbf{a}_i$ and $\tau_i = \mu^t\mathbf{b}_i$, so that

$$\widehat{\mathbf{u}} = \sum_{i=1}^{k} z_i\mathbf{a}_i,\ \ \|\mathbf{z}\|^2 = \|\widehat{\mathbf{u}}\|^2 = S_{\mathrm{res}} \ \text{ and }\ \mu = \sum_{i=1}^{k}\tau_i\mathbf{a}_i.$$

Then $\mathbf{z} = (z_1,...,z_k)$ has $\mathrm{var}(\mathbf{z}) = \mathbf{I}_k$ and expectation $\Delta\tau$, where $\tau = (\tau_1,..,\tau_k)$ has $\|\tau\| = \|\mu\|$.

The redundancy matrix

Proof of (A): It suffices to show that the condition $\mathbf{R}\widetilde{\mathbf{r}}_{\mathrm{m}}= \mathbf{0}$ is equivalent to $\widetilde{\mathbf{r}}_{\mathrm{m}}= \mathbf{B}_{\mathrm{m}}\beta$ for some β. It follows from (7) that

$$\mathbf{R}\mathbf{B}_{\mathrm{m}} = \mathbf{0}, \tag{33}$$

and hence $\widetilde{\mathbf{r}}_{m}= \mathbf{B}_{\mathrm{m}}\beta$ implies $\mathbf{R}\widetilde{\mathbf{r}}_{m}= \mathbf{0}$. On the other hand, if $\mathbf{R}\widetilde{\mathbf{r}}_{m}= \mathbf{0}$, the vector

$$\mathbf{r} = \left[\begin{array}{c} \mathbf{r}_{\mathrm{m}} \\ \mathbf{0} \end{array} \right]$$

verifies (1), and it follows from (8) that $\widetilde{\mathbf{r}}_{\mathrm{m}}= \mathbf{B}_{\mathrm{m}}\beta$.

Proof of (B): The result follows from $\mathbf{R}\widehat{\mathbf{e}} = \mathbf{R}\mathbf{r}_{\mathrm{ob}} - \mathbf{R}\widehat{\mathbf{r}}_{\mathrm{m}}$ and $\mathbf{R}\widehat{\mathbf{r}}_{\mathrm{m}} = \mathbf{R}\widetilde{\mathbf{r}}_{\mathrm{m}} = \mathbf{0}$ by (A) and (29).

Proof of (C): The *null subspace* of a matrix $\mathbf{A}$ is the set Null($\mathbf{A}$) of the vectors $\mathbf{a}$ such that $\mathbf{A}\mathbf{a} = \mathbf{0}$. The equivalence proved in (A) can be restated as

$$\mathrm{Null}(\mathbf{R}) = \mathrm{Im}(\mathbf{B}_{\mathrm{m}}). \tag{34}$$

The dimensions of Null($\mathbf{R}$) and of Im($\mathbf{B}_{\mathrm{m}}$) are $m - \mathrm{rank}\left(\mathbf{R}\right)$ and $\mathrm{rank}\left(\mathbf{B}_{\mathrm{m}}\right)$, respectively, and hence $m - k = d$.

Let z_i, $\mathbf{z}$ and $\mathbf{a}_i$ be the variables and vectors defined in the proof of (28). Then $\mathbf{z}^t \mathrm{var}\left(\mathbf{z}\right)^{-1} \mathbf{z} = \left\|\mathbf{z}\right\|^2 = \left\|\widehat{\mathbf{u}}\right\|^2 = S_{\mathrm{res}}$.

On the other hand, we may write $\varepsilon = \left(\mathbf{R}\boldsymbol{\Sigma}\right)\widehat{\mathbf{u}}$. Since $\widehat{\mathbf{u}} \in \mathrm{Im}\left(\mathbf{I}_m - \mathbf{H}\right)$, ε belongs to the subspace $V = \mathrm{Im}\left(\mathbf{R}\boldsymbol{\Sigma}\left(\mathbf{I}_m - \mathbf{H}\right)\right)$. Since (34) implies that Null($\mathbf{R}\boldsymbol{\Sigma}$) =Im($\mathbf{X}$), which is orthogonal to $\mathrm{Im}\left(\mathbf{I}_m - \mathbf{H}\right)$, we have that V has dimension k. Let $\mathbf{v}_1, ..., \mathbf{v}_k$ be an orthogonal basis of V, so that ε may be written as $\varepsilon = \sum_{i=1}^{k} w_i \mathbf{v}_i$. Let $\mathbf{w} = \left(w_1, ..., w_k\right)$. Then $\mathbf{w} = \mathbf{A}\mathbf{z}$ for some $k \times k$-matrix $\mathbf{A}$ of rank k, and hence

$$\varepsilon^t \mathrm{var}\left(\varepsilon\right)^{+} \varepsilon = \mathbf{w}^t \mathrm{var}\left(\mathbf{w}\right)^{-1} \mathbf{w} = \mathbf{z}^t \mathrm{var}\left(\mathbf{z}\right)^{-1} \mathbf{z}.$$

Proof of (D): It will be shown that the denominator of (31) equals $(1 - h_i)/\sigma_i^2$. Let $\mathbf{Q} = \left(\mathbf{R}\boldsymbol{\Sigma}\right)^t \left[(\mathbf{R}\boldsymbol{\Sigma})(\mathbf{R}\boldsymbol{\Sigma})^t\right]^{+} (\mathbf{R}\boldsymbol{\Sigma})$. Then the denominator is the i-th diagonal element of $\boldsymbol{\Sigma}^{-1}\mathbf{Q}\boldsymbol{\Sigma}$, and is hence equal to Q_{ii}/σ_i^2. We shall show that $\mathbf{Q} = \mathbf{I}_m - \mathbf{H}$.

A result from Linear Algebra states that if $\mathbf{c}$ is orthogonal to Null($\mathbf{A}$) —i.e., $\mathbf{c}^t\mathbf{a} = 0$ for all $\mathbf{a} \in$Null($\mathbf{A}$), then $\mathbf{c} \in$Im($\mathbf{A}^t$).

Since Null($\mathbf{R}\boldsymbol{\Sigma}$) =Im($\mathbf{X}$), the image of $\left(\mathbf{R}\boldsymbol{\Sigma}\right)^t$ coincides with the orthogonal complement of Im($\mathbf{X}$), which is $\mathrm{Im}\left(\mathbf{I}_m - \mathbf{H}\right)$. Finally, the definition of $\mathbf{Q}$ implies that $\mathbf{Q} = \mathbf{Q}^t = \mathbf{Q}^2$, and hence $\mathbf{Q}$ is the orthogonal projection on the image of $\left(\mathbf{R}\boldsymbol{\Sigma}\right)^t$.

A similar argument shows that the denominator of (31) equals u_i^2/σ_i^2.

Proof of (E): The result is proved in (Belsley et al., 1980). By the way, since $S_{(i)} \geq 0$ for all i, it follows that the statistic (26) fulfills $T^2 \leq S_{\mathrm{res}}$.

APPENDIX B: DETAILS OF COMPUTATION

The purpose of this section is to give the reader the elements to reproduce our results corresponding to the example in pages 16-17 of van der Heijden et al. (1994b). Table VI gives the matrix $\mathbf{E} = [\mathbf{E}_{\mathrm{m}}|\mathbf{E}_{\mathrm{c}}]$ in (1)-(2) Table VII gives the matrix $\mathbf{B}$ in (6) resulting from the Matlab command "Null(E)". Finally table VIII gives the vectors $\mathbf{r}_{\mathrm{ob}}$ and $\mathbf{y}$, the matrix $\mathbf{X}$ and the parameter vector β defined in (11), (12) and (13).

Recall that another software may yield a different $\mathbf{B}$ and hence different $\mathbf{y}$, $\mathbf{X}$ and β; but $\mathbf{H}$ and therefore the fitted values and residuals are the same.

References

Belsley, DA, Kuh, E and Welsch, RE. 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley and Sons

Chatterjee, S and Hadi, AS. 1988. Sensitivity Analysis in Linear Regression. New York: John Wiley and Sons.

Draper, NR and Smith, H. 2001. Applied Regression Analysis, 3rd Ed.. New York: John Wiley and Sons.

Montgomery, DC, Peck, EA and Vining, GG. 2001. Introduction to Linear Regression Analysis, 3rd Edition, New York: John Wiley and Sons.

Seber, GAF. 1984. Multivariate Observations. New York: John Wiley and Sons.

Sidák, Z 1967. Rectangular regions for the means of multivariate normal distributions. Jr. Amer. Statist. Assoc. 62**:**626-633.

Stapleton, JH. 1995. Linear Statistical Models. New York: John Wiley and Sons.

van der Heijden, RTJM. 1991. State estimation and error diagnosis for biotechnological processes. PhD thesis. Delft, the Netherlands.

van der Heijden, RTJM, Romein, B, Heijnen, JJ, Hellinga, C and Luyben, KChAM. 1992. Error detection and diagnosis. In: Modeling and Control of Biochemical Processes. IFAC Symposia Series:135-140.

van der Heijden, RTJM, Heijnen, JJ, Hellinga, C, Romein, B and Luyben, KChAM. 1994a. Linear constraint relations in biochemical reaction systems: I. Classification of the calculabitlity and the balanceability of conversion rates. Biotechnol. Bioeng. 43: 3-10.

van der Heijden, RTJM, Romein, B, Heijnen, JJ, Hellinga, C and Luyben, KChAM. 1994b. Linear constraint relations in biochemical reaction systems: II. Diagnosis and estimation of gross errors. Biotechnol. Bioeng. 43: 11-20.

Wang, NS and Stephanopoulos, G. 1983. Application of macroscopic balances to the identification of gross measurement errors. Biotechnol. Bioeng. 25: 2177-2208.

Weisberg, S. 1985. Applied Linear Regression, 2nd. Ed.. New York: John Wiley and Sons.

Table I: Rejection probabilities of the chi-squared test

Δ	h_i	deg. fr.		
		2	4	6
3	0.2	0.67	0.55	0.48
	0.4	0.54	0.43	0.36
	0.6	0.38	0.29	0.24
	0.8	0.21	0.16	0.13
	0.9	0.12	0.10	0.09
4	0.2	0.90	0.83	0.77
	0.4	0.80	0.70	0.62
	0.6	0.61	0.50	0.43
	0.8	0.34	0.26	0.22
	0.9	0.19	0.14	0.12

Table II: Example

	σ_i	h_i	$\widehat{e}_{\mathrm{st},i}$
X	1.07	0.958	-2.11
S	4.17	0.388	-2.05
P	2.20	0.628	-2.161
C	2.62	0.250	0.114
O	1.46	0.776	1.311
p		0.09	

Table III: Artificial data without gross errors

	$r_{\mathrm{ob},i}$	$\widehat{e}_{\mathrm{st},i}$
X	13.909	-0.89
S	-24.642	-0.91
P	0.156	-0.59
C	5.340	-0.56
O	-5.456	-0.04
p	0.65	

Table IV: Artificial data with gross errors (altered values in boldface)

$i_0 = 1,\ \Delta = 4$		$i_0 = 4,\ \Delta = 3$		$i_0 = 4,\ \Delta = 4$	
$r_{\mathrm{ob},i}$	$\widehat{e}_{\mathrm{st},i}$	$r_{\mathrm{ob},i}$	$\widehat{e}_{\mathrm{st},i}$	$r_{\mathrm{ob},i}$	$\widehat{e}_{\mathrm{st},i}$
18.181	-0.07	13.909	-0.24	13.909	-0.03
-24.642	-0.09	-24.642	-0.07	-24.642	0.21
0.156	0.12	0.166	-1.29	0.156	-1.52
5.340	-0.36	**13.190**	2.04	**15.811**	2.90
-5.456	-0.31	-5.456	2.13	-5.456	2.85
$p = 0.94$		0.09		0.01	

Table V: Detection and identification probabilities

	$\pi\,(\text{test})$	$\pi\,(i^*)$
X	0.08	0.15
S	0.55	0.49
P	0.35	0.52
C	0.64	0.75
O	0.23	0.48

Table VI: Balance matrix of example

		$\mathbf{E}_\mathrm{m}$			$\mathbf{E}_\mathrm{c}$	
1	1	1	1	0	0	0
1.83	2	3	0	0	3	2
0.56	1	0.5	2	2	0	1
0.17	0	0	0	0	1	0

Table VII: Matrix **B** given by Matlab

$\mathbf{B}_\mathrm{m}$	-0.2714	0.7745	-0.1469
	-0.3114	-0.5235	-0.5977
	0.6555	-0.0346	0.2247
	-0.0727	-0.2163	0.5198
	0.3868	0.2378	-0.4148
$\mathbf{B}_\mathrm{c}$	0.0461	-0.1317	0.0250
	-0.4927	0.0643	0.3575

Table VIII: Elements of regression model

$\mathbf{r}_\mathrm{ob}$	$\mathbf{y}$			$\mathbf{X}$	
21.37	20.00		-0.2540	0.7248	-0.1375
-69.45	-16.67		-0.07473	-0.1256	-0.1434
14.7	6.667		0.2973	-0.01569	0.1019
23.57	9.009		-0.02779	-0.08267	0.1987
-12.51	-8.547		0.2643	0.1625	-0.2834
		β	7.284	42.82	63.69